The Evolution of Statistical Hypothesis Testing

Bayesian Statistical Solutions to the Replication Crisis in the Biomedical Sciences

Riko Kelter

Dissertation
to obtain the degree of

Doctor rerum naturalium

submitted by
Riko Kelter
from
Siegen

submitted to the School of Science and Technology
of the University of Siegen
Siegen 2021

gedruckt auf alterungsbeständigem holz- und säurefreiem Papier

SUPERVISORS

Prof. Dr. rer. nat. Gregor Nickel
Department of Mathematics
University of Siegen
Siegen, Germany

Prof. Dr. Julio Michael Stern
Institute of Mathematics and Statistics
University of São Paulo
São Paulo, Brazil

Date of the oral examination: 22.10.2021

For Victoria and my family

The only theories which are formally complete, and of adequate scope for treating statistical evidence and its interpretation in scientific research contexts, are Bayesian.

Alan Birnbaum

The anomalous concept of statistical evidence

ABSTRACT

Statistical hypothesis testing is a central method for the judgement of empirical findings in the medical, social and natural sciences. In recent years, the ongoing problems with null hypothesis significance testing and p-values have shown that the underlying paradigm for quantifying statistical evidence about a research hypothesis is highly problematic, and the situation has been termed a replication crisis. In this thesis, the evolution of statistical hypothesis testing is reconstructed, and it is shown that various of the recently observed problems with the reproducibility of research can be attributed to the underlying statistical theory of widely used inferential statistical methods. In the first part, the development of an inconsistent hybrid approach to statistical hypothesis testing which emerged out of Fisher's theory of significance tests, p-values and the Neyman-Pearson theory is analyzed. In part two, the evolution of Bayesian approaches to hypothesis testing is detailed with a focus on the Bayes factor. Part three discusses the development of modern Markov-Chain-Monte-Carlo algorithms and their impact on Bayesian hypothesis testing. Part four then provides an axiomatic analysis of the concept of statistical evidence in the context of statistical hypothesis testing and it is shown that various substantial problems which were observed in the replication crisis can be attributed to purely axiomatic inconsistencies and conflicts with the likelihood principle. Based on the axiomatic analysis, it is shown that robust Bayesian methods, in particular robust Bayesian hypothesis tests provide a solution to some substantial problems with the reproducibility of research. Bayesian statistical solutions to the replication crisis are provided in the fifth part with a focus on widely used Bayesian statistical models in the biomedical sciences. New results demonstrate that the implicit error control of Bayesian hypothesis tests is comparable to frequentist tests based on p-values, and that a variety of Bayesian evidence measures attains reasonable type I error control and power in practice. Also, a shift towards the Hodges-Lehmann paradigm which advocates testing small interval instead of point null hypotheses is explored, and new theoretical results show that such a shift may be an appealing additional step towards increasing the reproducibility of science which has not received enough attention in the discussion about the validity of statistical hypotheses and the reproducibility of scientific research.

ZUSAMMENFASSUNG

Statistische Hypothesentests bilden eine zentrale Methode für die Bewertung empirischer Studien in der Medizin, den Sozialwissenschaften und den Naturwissenschaften. Die zunehmenden Probleme mit der Reproduzierbarkeit wissenschaftlicher Resultate haben gezeigt, dass die zu Grunde liegende mathematische Theorie zur Quantifizierung statistischer Evidenz im Kontext statistischer Hypothesentests hochgradig problematisch ist, und die Situation wird weitläufig als wissenschaftliche Reproduzierbarkeitskrise bezeichnet. In dieser Arbeit wird die Evolution statistischer Hypothesentests rekonstruiert und gezeigt, dass diverse im Rahmen der Reproduzierbarkeitskrise beobachtete Probleme ursächlich auf die den Verfahren zu Grunde liegende statistische Theorie zurückgeführt werden können. Im ersten Teil wird die Entwicklung eines inkonsistenten hybriden Ansatzes statistischer Hypothesentests beschrieben, welcher sich aus Fisher's Theorie der Signifikanztests mittels p-Werten und der Neyman-Pearson-Theorie entwickelte. In Teil zwei wird die Entwicklung Bayes'scher Ansätze zum Hypothesentesten und insbesondere die Entwicklung des Bayes-Faktors analysiert. Teil drei zeigt die Entwicklung moderner Markov-Chain-Monte-Carlo-Algorithmen und deren Bedeutung für die Anwendbarkeit Bayes'scher Hypothesentests auf. Im vierten Teil bildet eine axiomatische Analyse des Konzepts statistischer Evidenz im Kontext von Hypothesentests den Ausgangspunkt und es wird gezeigt, dass eine Vielzahl der Probleme der Reproduzierbarkeitskrise ursächlich auf die axiomatischen Grundlagen und Konflikte mit dem Likelihood-Prinzip zurückgeführt werden können. Die Ergebnisse der axiomatischen Analyse zeigen, dass robuste Bayes'sche Analysen und Hypothesentests eine Lösung für eine Vielzahl der Probleme mit der Reproduzierbarkeit wissenschaftlicher Resultate darstellen. Bayes'sche statistische Lösungen für die Reproduzierbarkeitskrise werden im fünften Teil der Arbeit mit einem Schwerpunkt auf statistische Modelle aus der Biostatistik und medizinischen Biometrie vorgestellt und es wird demonstriert, dass robuste Bayes'sche Hypothesentests für den Großteil dieser statistischen Modelle verfügbar sind. Neue Resultate zeigen, dass Bayes'sche Hypothesentests über eine vergleichbare Fehlerkontrolle verfügen wie frequentistische auf p-Werten basierende Tests, und dass mehrere Bayes'sche Evidenzmaße praxistaugliche Typ-I-Fehlerraten sowie Trennschärfen erzielen. Zusätzlich wird ein Paradigmenwechsel zum Hodges-Lehmann-Paradigma analysiert, nach welchem das Testen von kleinen Intervallhypothesen an Stelle von präzisen Punkthypothesen für eine Vielzahl von Anwendungskontexten – insbesondere in der Medizin – realistischer ist. Neue theoretische Resultate zeigen, dass ein solcher Paradigmenwechsel eine attraktive zusätzliche Lösung der Reproduzierbarkeitsprobleme darstellt, welcher bisher zu wenig Aufmerksamkeit erhalten hat.

PREFACE

This manuscript has been written in the last three years during my time at the Department of Mathematics at University of Siegen. The on-going problems with the reproducibility of empirical research results in the biomedical and cognitive sciences still pose a serious problem to scientists, academic institutions and the funding public. While the reasons of these problems are still widely debated in the statistical literature, most research takes a static perspective on the situation. In this thesis, a dynamic perspective is taken which focusses on how statistical hypothesis testing has evolved over the last century. By embracing such a perspective a much richer picture is provided which illustrates how and why the replication problems in the biomedical sciences are rooted deeply inside statistical theory itself. Based on this perspective, this thesis contributes to the literature by presenting several Bayesian statistical and biometrical solutions to improve the reliability of research in the biomedical and cognitive sciences with an emphasis on Bayesian hypothesis tests and Bayesian evidence measures.

However, this thesis would not have been written without the support of several people. First of all, I would like to thank my supervisor Gregor Nickel, who supported me from the first ideas until the final typesetting of the manuscript. I had the immense privilege to work in the most comfortable scientific environment imaginable, having the freedom to pursue my interests while simultaneously being supported by my supervisor. Without our many discussions this thesis would be a different. Also, I would like to thank Julio Stern from the Institute of Mathematics and Statistics at the University of São Paulo for inspiring exchanges on Bayesian evidence measures and Bayesian hypothesis tests, for co-supervision of this thesis and for pointing me towards the Full Bayesian Significance Test which now builds the basis of a variety of further research.

Furthermore, I am grateful to Philip Dawid for sharing some of his wisdom about statistical principles with me. I also thank Eric-Jan Wagenmakers, John Kruschke and Daniel Lakens for exchanges on Bayesian evidence measures and equivalence testing from the perspective of psychologists. In the last two years, my academic home has been the AG42 at University of Siegen. I would like to thank Susanne Spies, Shafie Shokrani, Daniel König and Heiko Hoffmann for their support throughout the last three years. Also, I am grateful to Edgar Kaufmann and Alexander Schnurr for their support.

Doing research under these desirable conditions was made possible primarily by the Graduate Center of the University of Siegen, in particular the HYT PhD program. I am grateful to the HYT for being awarded with a doctoral scholarship which made this research possible after all. Without this financial support this manuscript would not have been written.

After all, this thesis would not exist without the support of my family, in particular, without the support of my parents. I owe you everything. Finally, this manuscript would not have been written without you, Victoria. Thank you for your never-ending support throughout this marvellous adventure. It's done.

Partial results of the presented work have been published in:

- Kelter, R. (2020). bayest: An R Package for effect-size targeted Bayesian two-sample t-tests. *Journal of Open Research Software* 8:14.
https://doi.org/10.5334/jors.290

- Kelter, R. (2020). Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology* 20(88).
https://doi.org/10.1186/s12874-020-00968-2

- Kelter, R. (2020). Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Research Notes*, 13(1).
https://doi.org/10.1186/s13104-020-05291-z

- Kelter, R. (2020). Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology* 1(142).
https://doi.org/10.1186/s12874-020-00980-6

- Kelter, R. (2020). Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *WIREs Computational Statistics*, 7.
https://doi.org/10.1002/wics.1523

- Kelter, R. (2020). Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. *Measurement: Interdisciplinary Research and Perspectives*, 18(2), 101–119.
https://doi.org/10.1080/15366367.2019.1689761

- Kelter, R. (2019). bayest – Effect size targeted Bayesian two-sample t-tests via Markov Chain Monte Carlo in Gaussian mixture models. *Comprehensive R Archive Network* (*CRAN*).
https://cran.r-project.org/web/packages/bayest/index.html

- Kelter, R. (2020). New Perspectives on Statistical Data Analysis: Challenges and Possibilities of Digitalization for Hypothesis Testing in Quantitative Research. In J. Radtke, M. Klesel, & B. Niehaves (Eds.), *Proceedings on digitalization at the Institute for Advanced Study of the University of Siegen*, p. 100–108, Institute for Advanced Study of the University of Siegen.
http://dx.doi.org/10.25819/ubsi/1894

Contents

Chapter 1

Introduction

> There is increasing concern that most current published research findings are false.
>
> John P. Ioannidis
> *Why most published research findings are false*

A few years ago, in March 2016, the American Statistical Association (ASA) published a warning about a common statistical method (Wasserstein and Lazar, 2016). The issue was published in *The American Statistician*, with several leading statisticians involved in the publication, suggesting the statistical method was leading to wrong conclusions, wasting taxpayers money and damaging science and the reputation of scientific work in general. The method named of course was the p-value.

Today, substantial parts of scientific research are heavily influenced and even based on statistical methods. From clinical trials, epidemiology and psychological studies over experiments in the domains of physics and engineering to economics, p-values have been used since their earliest days to back claims for the discovery of *significant* effects in often noisy data. Statisticians know about these problems since a long time, so for them, the situation is nothing to be surprised about. Already decades ago scientists warned about the incorrect use or interpretation of statistical methods - especially p-values - and today the same old problem confronts science again. Statistical inference as a mathematical discipline is shaped by the turbulent history of change and transformation in the last century and its evolution was influenced by feuds and debates about the right approach for a given statistical problem. The debate often centered around the existence of two competing core models for statistical inference: the frequentist and the Bayesian approach.

1.1 The Replication Crisis in the biomedical Sciences

Null hypothesis significance testing (NHST) is the leading but also controversial scientific method for establishing new results. The use of hypothesis tests in scientific research has grown steadily over the last decades (Halpin and Stam, 2006; Hubbard and Ryan, 2000; Hubbard, 2004) and in recent years, more and more problems have been identified in NHST. These problems undermine the reliability of research and can be attributed to the overuse of hypothesis tests in research (Kelter, 2020b,d; Nuzzo, 2014)

The implication of these problems is the irreproducibility of statistically significant findings. Therefore, the situation has been termed a reproducibility crisis of science (Baker and Penny, 2016; Wagenmakers and Pashler, 2012; Ioannidis, 2005b; Colquhoun, 2017) and started a debate about the appropriateness of hypothesis tests for complex research scenarios like clinical trials, psychological interventions or sociological experiments (Gigerenzer, 2004). Various authors have critisized NHST, in particular, p-values (Ioannidis, 2005b). p-values are probably the most widespread statistical tool to separate between significant and non-significant research findings, and they are an essential part of NHST. Technically, the p-value is the probability of observing a result equal to or more extreme than the one obtained, under assumption of the null hypothesis $H_0 : \theta = \theta_0$ (Held and Sabanés Bové, 2014). Here, θ is the parameter of interest, and θ_0 the value specified by the null hypothesis. Although the p-value has a long tradition in statistical science, the recently rediscovered problems have made the p-value fall into disrepute (Haaf et al., 2019; Halsey, 2019; Greenland, 2019; Kelter, 2021a).

In 2016 the American Statistical Association (ASA) released a statement on the proper use and interpretation of p-values (Wasserstein and Lazar, 2016). Although the statement included six principles how to deal with p-values, few has changed since (Hubbard, 2019; Matthews et al., 2017). This can be attributed to a lack of alternative methodologies and precise guidance for researchers how to replace p-values. The situation did not improve even after a second statement was published three years afterwards (Wasserstein et al., 2019). What is however clear by now, is that the problems of NHST and p-values are not overblown (Pashler and Harris, 2012) and that the false-discovery rate of scientific findings and frequent misinterpretations of p-values slow scientific progress (McElreath and Smaldino, 2015; Colquhoun, 2014). Even worse, by now, there is few consensus how to solve these issues (Wasserstein et al., 2019). Also, the scientific areas affected by these problems range from neuroscience (Button et al., 2013) and the cognitive sciences (Haaf et al., 2019; Kelter, 2021c; Ly et al., 2020) over political science (Gigerenzer, 2004; Gelman et al., 2019) to medical research (Kelter, 2020b; Ioannidis, 2005b,a, 2016) which illustrates the dimension of the problem.

Proposed solutions range from stricter thresholds for statistical significance (Benjamin et al., 2018) over various methodological modifications (Benjamin and Berger, 2019; Brownstein et al., 2019; Hurlbert et al., 2019) to proposals like adopting an entirely different paradigm like the Bayesian approach.

Already before the ASA statement in 2016, the scientific discussion heated up with the beginning of the so-called replication crisis. In 2005, John P.A. Ioannidis published his landmark paper *Why Most Published Research Findings Are False* (Ioannidis, 2005b). In it, Ioannidis modelled a framework for calculating false-positive findings in research to explain the noticeably high rate of failed replication attempts of some highly prestigious research. This research was often assessed by statistical significance in the form of a p-value.

Ioannidis' argument went as follows: First, according to Ioannidis (2005b) "both true and false hypotheses can be made about the presence of relationships." By denoting R as the ratio of true relationships #TR to false relationships #FR among all tested in the research field (a kind of unknown but existing a priori ratio), the pre-study probability (PSP) of a relationship being true is derived as

$$\text{PSP} = \frac{\#\text{TR}}{\#\text{TR} + \#\text{FR}} = \frac{\frac{\#\text{TR}}{\#\text{FR}}}{\frac{\#\text{TR} + \#\text{FR}}{\#\text{FR}}} = \frac{\frac{\#\text{TR}}{\#\text{FR}}}{\frac{\#\text{TR}}{\#\text{FR}} + \frac{\#\text{FR}}{\#\text{FR}}} = \frac{R}{R+1} \tag{1.1}$$

The probability of claiming a relationship when none exists is denoted as the type I error rate α, and one assumes further that c relationships in total are being probed in the research field. A type II error is defined as a non-significant finding when indeed there exists a true relationship and the type II error rate is denoted as β. The power to detect an effect when one is present is then given as 1-β. After a relationship has been claimed as statistically significant, the post-study probability that this relationship indeed is true and exists is denoted by Ioannidis as the *positive predictive value* (PPV). The PPV equals the complement of the *false positive report probability*, also denoted as the *false discovery rate* (FDR).

$$\begin{aligned} \texttt{PPV} &= \texttt{P(True Relationship|Significant Finding)} && (1.2) \\ \texttt{FDR} &= \texttt{P(False Relationship|Significant Finding)} && (1.3) \end{aligned}$$

Ioannidis (2005b) then calculated the cells of a 2x2 table with probabilities for all four combinations of significant findings and true relationships (see Table 1.1).

Significant	True Relationship		
Finding	Yes	No	Total
Yes	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1))$
No	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

Table 1.1: Significant findings and true relationships according to (Ioannidis, 2005a)

By these probabilities one can obtain the PPV via Bayes' theorem:

$$\begin{aligned} \texttt{PPV} &:= \texttt{P(True Relationship|Significant Finding)} \\ &= \frac{c(1-\beta)R/(R+1)}{c(R+\alpha-\beta R)/(R+1))} = \frac{(1-\beta)R}{(R+\alpha-\beta R)} = \frac{(1-\beta)R}{(R+\alpha-\beta R)} \end{aligned} \tag{1.4}$$

Solving this equation for PPV $>$ 0.5, that is, a significant finding is post-study more likely to be true than false, results in the condition

$$\begin{aligned} \texttt{PPV} \overset{(1.4)}{:=} & \frac{(1-\beta)R}{(R+\alpha-\beta R)} > 0.5 \\ \Leftrightarrow & (1-\beta)R > 0.5(R+\alpha-\beta R) \\ \Leftrightarrow & (1-\beta)R > 0.5R(1-\beta)+0.5\alpha \\ \Leftrightarrow & 0.5R(1-\beta) > 0.5\alpha \\ \Leftrightarrow & R(1-\beta) > \alpha \end{aligned} \tag{1.5}$$

As most researchers use the (arbitrary) threshold $\alpha = 0.05$, Ioannidis (2005b) concluded that a post-study significant finding is more likely true than false if $R(1-\beta) > 0.05$ due to Equation (1.5). Ioannidis (2005b) then introduced *bias* as another confounding variable. Therefore, he denoted u as the "proportion of probed analysis that would not have been "research findings", but nevertheless end up presented and reported as such, because of bias." (Ioannidis, 2005b, p. 2), which is not to be conflated with findings to be non-significant by chance. He defines bias as procedures entailing manipulation of data analysis or the reporting of findings, respectively selected reporting

Significant	True Relationship		
Finding	Yes	No	Total
Yes	$(c[1-\beta]R+uc\beta R)/(R+1)$	$(c\alpha+uc(1-\alpha))/(R+1)$	$c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1))$
No	$(1-u)c\beta R/(R+1)$	$(1-u)c(1-\alpha)/(R+1)$	$c(1-u)(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

Table 1.2: Significant findings and true relationships in the presence of bias according to (Ioannidis, 2005a)

of findings. Reasonably assuming independence of u from the fact that a true relationship exists or not Ioannidis (2005b) then set up the following probabilities shown in Table 1.2.

Taking for example the entry for both true relationship and significant finding set to *Yes*, the probability has two components:

1. The first, being the product of the c probes made in total with the pre-study probability $R/(R+1)$ of a true relationship, multiplied by $[1-\beta]$. Here, $[1-\beta]$ is the proportion of findings where no type II error has happened, meaning no true relationship has incorrectly been classified as non-significant. The first component thus results in $c[1-\beta]R/(R+1)$.

2. The second, being the product of the c probes made in total with the pre-study probability $R/(R+1)$ of a true relationship, this time multiplied by β and u. Again, the fraction of true relationships are the product of $cR/(R+1)$. The multiplication with β results in $c\beta R/(R+1)$, a quantity which describes the fraction of type II errors made, which is the fraction of true existing relationships which have incorrectly been classified as non-significant. By finally multiplying with the bias u, the resulting portion describes the portion of type II errors made, which have become significant again due to bias. The second component thus results in $uc\beta R/(R+1)$.

Adding both components results in the probability of the given cell. As the total portion of true relationships in the underlying data does not change, the cell with a true relationship and no significant finding can be calculated by subtracting the above quantity from the total quantity for true relationships $cR/(R+1)$, resulting in $(1-u)c\beta R/(R+1)$. Ioannidis (2005b) considered now the cell for no true relationship and a significant finding. Again, there are two components which build the resulting term:

1. The first, being $\frac{c\alpha}{R+1}$. This component is simply the portion of non-true relationships being probed where a type I error is made, resulting in a significant finding which in reality has no underlying relationship. The term is derived by multiplying the total of relationships being probed in the field, c, with the fraction of non-true relationships in the underlying data, $c(1-\frac{R}{R+1})$. The term is then multiplied with α.

2. The second, being $uc(1-\alpha)/(R+1)$. This term is derived from the proportion of relationships being probed in the field with no true relationship in the underlying data, $\frac{c}{R+1}$. Multiplying with $1-\alpha$ results in the fraction of cases where no true relationship exists and no significant finding is stated, $\frac{(1-\alpha)c}{R+1}$. By introducing bias (like selective reporting or faulty data analysis) through multiplying

this quantity with u, these correctly non-significant classified cases are then incorrectly converted to significant research findings although no true relationship exists. These calculations result in the second term $uc(1-\alpha)/(R+1)$.

Adding both terms results then in $(c\alpha + uc(1-\alpha))/(R+1)$. Given these derivations, it is straightforward to carry out the calculations for all other entries in Table 1.2. By using the probabilities respecting bias Ioannidis (2005b) then calculated the posterior predictive value (PPV) for a true relationship after a significant finding:

$$\begin{aligned} \texttt{PPV} &= P(\text{true relationship}|\text{significant finding}) && (1.6)\\ &= \frac{(c[1-\beta]R + uc\beta R)/(R+1)}{c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1))} && (1.7)\\ &= \frac{([1-\beta]R+u\beta R)}{(R+\alpha-\beta R+u-u\alpha+u\beta R)} && (1.8)\end{aligned}$$

By regarding α, β and R as constant and calculating the derivative with respect to u it follows that the PPV as a function of u is decreasing when the condition $\alpha \geq 1-\beta$ holds. When α is set to 0.05, this equals the condition $0.05 \geq 1-\beta$ or $\beta \geq 0.95$. The PPV

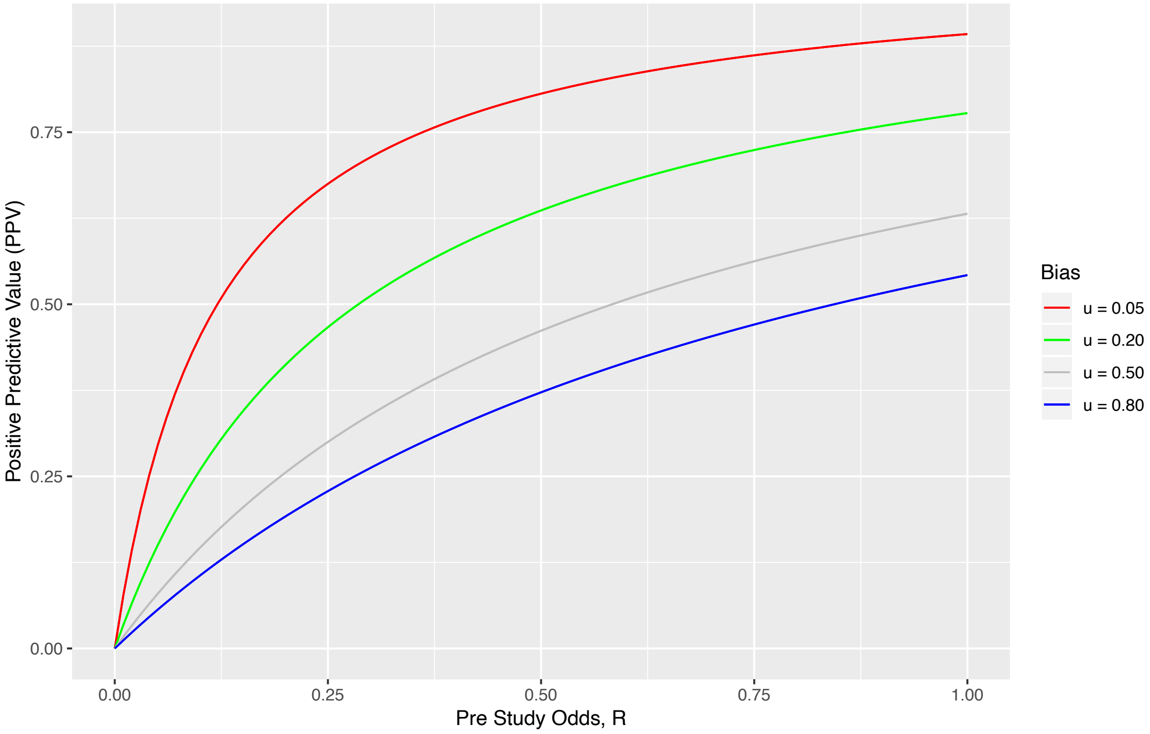

Figure 1.1: PPV calculations for different pre-study odds with test power of .80 ($power = 1-\beta$), (Ioannidis, 2005a)

for different pre-study odds can then be calculated. Figure 1.1 shows the PPV values for different pre-study odds R (rescaled from $R/(R+1)$) with a given test power[1] $1-\beta$ of 0.80. As can be seen from Figure 1.1, the PPV is strongly influenced by bias u and the pre study odds R. Also, even when assuming only very small bias $u = 0.05$, when the

[1] The test power is the probability of correctly stating a significant finding when there exists a true relationship. Therefore, it is equal to $1-\beta$, the proportion of cases where no type II error (incorrectly stating no significant finding when indeed there is a true relationship) is made.

pre study odds R of true relationships among all probed is small, say 12.5%, the PPV is only about 50%. Thus, one could flip a coin to decide whether a research finding is true or false instead of testing for statistical significance.

The last argument of Ioannidis (2005b) involved the testing by several independent research teams and its influence on the PPV. He argued, that when testing by several independent teams is conducted the post-study probability of no significant finding in all n teams when a true relationship exists equals $cR\beta^n/(R+1)$, derived by the matching entry of Table 1.1. This term equals the probability that all n teams make a type II error, and that happens exactly with probability β^n multiplied with the product of the total relationships being probed, c, and the pre-study odds $R/(R+1)$ for a relationship to be true. Accordingly, the cell for no significant finding when no true relationship exists is calculated as the complement and all other cells can be calculated straightforward in the same manner by the given marginal probabilities of the table, see Table 1.3. Calculation of the PPV for different number of studies n shows that for larger n the PPV is smaller. Thus, when a large number of research groups is working on a single topic, the probability of reliable results does, paradoxically, not increase, but decrease. The probability that a research finding is true does not increase when several independent teams work together. Instead, the PPV is even smaller.[2] Finally, Ioannidis (2005b) stated multiple corollaries of the described derivations, the three most important being:

Significant	True Relationship		
Finding	Yes	No	Total
Yes	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1))$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

Table 1.3: Significant findings and True Relationships in the Presence of Multiple Studies, reproduced from (Ioannidis, 2005a)

1. “The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.” (Ioannidis, 2005b, p. 3). The argument for this corollary involves that all PPV functions stated above decrease as functions of the power $1-\beta$. The sample size determines the amount of sampling error of a test result. As decreasing sample size N increases the standard error $\frac{\hat{\vartheta}}{\sqrt{N}}$ of an unbiased estimator $\hat{\vartheta}$, smaller sample sizes lead to more sampling error in test results and therefore less power of accurately stating a significant finding when there is a true relationship, that is $1-\beta$. Conversely, increasing sample size easily boosts the statistical power of a hypothesis test (Casella and Berger, 2002).

2. “The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.” (Ioannidis, 2005b, p. 3). This statement refers to the results of bias-respecting PPV functions above. As bias increases, the PPV decreases as can easily be seen in Figure 1.1.

[2] Statistically, this is obvious in the context of hypothesis testing also from the fact that the power of hypothesis tests is in the frequentist approach a function which depends on the studies sample size n. A single study with sample size n has much higher power to detect a true effect than 10 small studies with sample sizes $n/10$.

3. "The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true" (Ioannidis, 2005b, p. 3). This statement refers to the results derived by calculating the PPV in the presence of multiple studies, compare Table 1.3.

While Ioannidis (2005b) derivations raised attention and started a discussion in the scientific community, one big problem remained, which was the lack of alternatives to proceed with:

> "Better powered evidence, e.g. large studies or low-bias meta-analyses, may help, (...). However, large studies may still have biases (...). Moreover, large-scale evidence is impossible to obtain for all of the millions and trillions of research questions posed in current research. Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high so that a significant research finding will lead to a post-test probability that would be considered quite definitive."
> Ioannidis (2005b, p. 3)

As the pre-study probability for most research questions is unknown and can only be estimated by conducting meta-studies with noticeable effort, following these rules would strongly increase the amount of required work to do reliable research. Maybe the most important point stated in the last section of his paper was the following:

> "Diminishing bias through enhanced research standards and curtailing of prejudices may also help. However, this may require a change in scientific mentality that might be difficult to achieve."
> Ioannidis (2005b, p. 3)

Importantly, Ioannidis was among the first who argued against the contemporary *scientific mentality* when it comes to statistical analyses of scientific data. This mentality encompasses bias in research and the use of inappropriate statistical methods for answering a given research question.

In the same year, Ioannidis (2005a) also conducted a meta-study following his theoretical arguments and successfully reported problems concerning the reproducibility in highly cited clinical research, showing that in 49 medical studies conducted between 1990 and 2003 the results of 16% were contradicted by future studies, and another 16% found stronger effects than their successors (Ioannidis, 2005a). This result also confirmed the much earlier results from Glick (1992), who conducted a similar meta-analysis about the period from 1977 to 1990.

While Ioannidis' work raised concern in some scientific domains, among them especially clinical research in medicine and psychology, researchers soon went on to business as usual. In the early 2010s then, the replication crisis reemerged: Simmons et al. (2011) showed in simulations that small changes in data-analysis decisions could increase the false-positive rate of a single study to 60%, showing clearly that false-positive conclusions are drawn regularly from research via the current statistical methodology. In 2012, John et al. (2012) joined the debate and published a study which reported the results of over 2000 interviews held with psychologists about their research practices. These research practices were then classified into questionable research practices (QRPs) like selective reporting or partial publication of the data used for analysis, optional stopping, p-value-rounding or the manipulation of outliers. The results were discussed widely (Fiedler and Schwarz, 2016). Indeed, John et al. (2012) were able to prove that a majority of participants admitted to using at least one of the questionable QRPs, indicating that there is a definite problem about the validity regarding research in the cognitive sciences.

Pashler and Harris (2012) argued that all this may be a self-cleansing process of science itself, necessary to separate research that can withstand the test of time from non-relevant results. Ioannidis (2012) questioned this and the following years showed that he was right: Begley and Ellis (2012) showed that only 11% of cancer studies could be replicated at all, pouring fuel into the already heated discussion. One year later, Johnson (2013) claimed revised standards for statistical evidence as many others (Begley, 2013) by proposing increased thresholds for significant findings:

> “An examination of these connections suggest that recent concerns over the lack of reproducibility of scientific studies can be attributed largely to the conduct of significance tests at unjustifiably high levels of significance. To correct this problem, evidence thresholds required for the declaration of a significant finding should be increased to 25-50:1, and 100-200:1 for the declaration of a highly significant finding. In terms of classical hypothesis tests, these evidence standards mandate the conduct of tests at the 0.005 or 0.001 level of significance.”
> Johnson (2013, p. 1)

In February 2014, Nuzzo (2014) published an article about the limitations of the often-used p-value, directly addressing the connections between the replication crisis and the underlying statistical methods. Nuzzo (2014, p. 150,151) argued, that while p-values “always had their critics (...) the p value was never meant to be used the way it’s used today.”. Nuzzo (2014) made an important point in her paper which was missing in prior search for causes: Next to the scientific mentality criticized by Ioannidis (2005b), she added the historical perspective. Drafting the invention of the p-value by the british statistician Ronald Fisher in the 1920s (but leaving out most details as well as the successive work of Neyman and Pearson on hypothesis testing), Nuzzo drew the attention to the historical causes of the situation. Reciting statistician Steven Goodman in her article she stressed:

> “The basic framework of statistics has been virtually unchanged since Fisher, Neyman and Pearson introduced it.”
> Nuzzo (2014, p. 152)

Nuzzo’s paper added the historical perspective to the debate and pointed out an additional important fact, again reciting Steven Goodman:

> “Change your statistical philosophy and all of a sudden different things become important.”
> Nuzzo (2014, p. 151)

Similar to Nuzzo (2014), Colquhoun (2014) made the statistical methods and their evolution responsible for the crisis, elaborating the problems of hypothesis testing further by simulation experiments. Like Ioannidis (2005b), Colquhoun (2014) started with the definition of the positive predictive value (PPV) and the false discovery rate (FDR): While Ioannidis (2005b) in his analysis investigated primarily the PPV, Colquhoun (2014) was interested in the FDR as this rate accurately describes the portion of type I errors made. To calculate the FDR he first made the following assumptions:

(i) The test power $1 - \beta$ is assumed to be 0.80, as this is the test power most often specified in power analyses before conducting studies or clinical trials (Altman, 2000). The type I error probability α of a test[3], is assumed to be 0.05.

(ii) The fraction of relationships which are true out of all relationships tested is assumed to be constant. (which is equal to Ioannidis' pre-study odds R). This quantity cannot previously be known. For ease of notation, this quantity will also be called the pre-study odds here.

Colquhoun (2014) then proceeded with two examples, the first being a theoretical derivation of the FDR for 1000 tests with given pre-study odds of 10%, meaning 10% of tests conducted will have a true relationship to be discovered or not and 90% will be cases where there is no true relationship. By considering the 900 tests where there is no true relationship, according to classical test theory 5% (45) will result in a false positive, a type I error of stating a significant finding when there is no true relationship. Similar, in the 100 tests where there is a true relationship, because of the test power of 80%, exactly 80 tests will result in a significant finding, missing 20 true relationships which result in non-significant findings by making a type II error in those 20 cases. Therefore, the portion of false discoveries results in

$$\begin{aligned} FDR =& P(\text{no true relationship}|\text{significant result}) \\ =& \frac{P(\text{no true relationship} \cap \text{significant result})}{P(\text{significant result})} \\ =& \frac{\frac{45}{1000}}{\frac{80}{1000} + \frac{45}{1000}} = \frac{45}{80 + 45} = 0.36 \end{aligned} \tag{1.9}$$

Therefore, Colquhoun (2014) concluded that in the long run the FDR will be about 36% and not only 5% as assured by the type I error rate $\alpha = 0.05$. He noted:

> "It shows that there is a problem, but does not provide all the answers. Once we go a bit further, we get into regions where statisticians disagree with each other."
> Colquhoun (2014, p. 5)

His second argument involves a simulation of 100.000 Student's t-tests for differences between two group means. Therefore, for every test "two groups of simulated 'observations' are generated as random variables with specified means and standard deviations. The variables are simulated as normally distributed, so the assumptions of the t-tests are exactly fulfilled." (Colquhoun, 2014, p. 5). Sample sizes of $n = 16$ are chosen to match the achieved test power of 80%, and a difference of one standard deviation between means is assumed. Therefore, samples from the first and second group are simulated from a normal distributed random variable with mean and variance equal to one. By simulating 100.000 pairs of samples with 16 observations in each group and running the test, his simulations show that indeed 78% of the p-values fulfill $p \leq 0.05$, being near the specified test power of 80%.[4] However, in this setting the pre study odds were 1, that is, every relationship probed was true. Assuming now pre-study odds of

[3] See Definition C.73.

[4] The 2% deviation are due to sampling variability in the simulation.

10%, Colquhoun (2014) concluded that the FDR results in:

$$\begin{aligned}FDR =& P(\text{no true relationship}|\text{significant result})\\ =& \frac{P(\text{no true relationship} \cap \text{significant result})}{P(\text{significant result})}\\ =& \frac{0.9 \cdot \alpha}{0.9 \cdot \alpha + 0.1 \cdot 0.78} = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.78} = \frac{0.045}{0.123} \approx 0.37 \end{aligned} \tag{1.10}$$

Therefore, like in the theoretical derivation, the simulations also back the claim of a long run FDR of at least 36%. Still, when changing the pre-study odds to 50% the FDR reduces to ≈ 0.06, which is close to the assured 5% by $p = 0.05$. Still, in practice it is unrealistic to assume that one out of two relationships probed is true:

> "(...) there is no reason to think that half the tests we do in the long run will have genuine effects."
> Colquhoun (2014, p. 7)

Furthermore, Colquhoun (2014) questioned that the assumption of 80% power is realistic, mentioning the results of

- Cohen (1962), who found average power $1 - \beta$ in 70 investigated research studies in psychology of 0.18 for small, 0.48 for medium and 0.83 for large effects.
- Button et al. (2013), who found that an optimistic estimation of the median power in neuroscience research lies between 8% and 31%.

Preferring the approach of minimum FDRs as advertised by Sellke et al. (2001a), Colquhoun (2014) also makes the evolution of statistical inference itself responsible for the replication crisis, citing Matthews (1998):

> "The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug."
> Matthews (1998) in Colquhoun (2014, p. 12)

One year later, in 2015, Begley and Ioannidis (2015) summarized the current state of the crisis:

> "At the heart of this irreproducibility lie some common, fundamental flaws in the currently adopted research practices."
> Begley and Ioannidis (2015, p. 116)

After addressing multiple institutionally caused problems in science, Begley and Ioannidis (2015) discussed what constitutes the reproducibility crisis and conclude that "empirical assessments (...) showed an array of problems, including (...) to use legitimate controls, to validate reagents, and use appropriate statistical tests." (Begley and Ioannidis, 2015, p. 117). Next to this problem, the sheer number of publications increased to the staggering number of 15 million scientists publishing at least one article that was indexed in Scopus in the years from 1996 until 2011, makes it nearly impossible to stay up to date in the own scientific field. Begley and Ioannidis (2015) present a meta-analysis conducted in multiple scientific domains, showing that while the number of publications increases, the evidence of irreproducibility across most domains grows further and further, too. As an example, Begley and Ioannidis (2015) cited studies conducted in neuroscience, pharmacology, bioinformatics, chemistry and computational biology, all contributing to the claim that reproducibility is problematic and one of the causes lies in the statistical analyses carried out by researchers (Kenakin et al., 2014; McGonigle and Ruggeri, 2014). Other researchers like Winquist et al. (2014) and Marino (2014) went even further:

> "(...) the emphasis on statistical methods in pharmacology has become dominated by inferential methods often chosen more by the availability of user-friendly software than by any understanding of the data set or the critical assumptions of the statistical tests. Such frank misuse of statistical methodology and the quest to reach the mystical $\alpha < 0.05$ criteria has hampered research via the publication of incorrect analysis driven by rudimentary statistical training."
> Marino (2014, p. 1)

Similar findings were made in the fields of computational biology (Sandve et al., 2013), bioinformatics (Sugden et al., 2013) as well as in chemistry (Davis and Erlanson, 2013). Begley and Ioannidis (2015) thus concluded that while the number of researchers being aware of the problems is steadily increasing, few solutions are in sight. Large-scale, cooperative efforts to better the situation like the *Reproducibility Project*[5] in the domain of psychology are rare. Multiple initiatives were started, among them also the *Reproducibility Project: Cancer Biology*[6], which investigates the reproducibility of 50 highly influential cancer studies. In 2017, Baker and Dolgin (2017) investigated the first results from the project, resulting in a mixed message about the reproducibility. All in all, there seems to be no consensus by now. In summary, Begley and Ioannidis (2015) argued that there are essentially five major points which should be addressed when investigating causes and possible solutions to the replication crisis (Begley and Ioannidis, 2015, p. 1):

1. The generation of new data and scientific publications is observed at a by-now unprecedented rate.

2. There is convincing evidence that the majority of new discoveries will not stand the test of time.

3. The causes of the reproducibility crisis may be seen in the failure to adhere to good scientific practice and the *publish-or-perish* mentality across sciences.

4. The problem is a complex multistakeholder problem.

5. There is no solely responsible party, and no single solution will suffice.

The current situation offers the following picture concerning the above five points:

- The first point has been discussed intensively by Siebert et al. (2015) and Parolo et al. (2015), both of them pointing out that the overflow of data and publications has an impact on the quality of scientific writing and is one cause for the situation and the scientific incentive system is entangled with this cause.

[5] https://osf.io/ezcuj/

[6] https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology

- The second point can be regarded as proven by the many metastudies conducted since the beginning of the reproducibility crisis, although no appealing solution is at hand. Even the causes for the situation are not clear by now.

- Concerning the third and fourth point, social and institutional aspects can be stated as important factors, but open a completely new discussion. The problem of the underlying incentive-system in science has its definitive place among the causes of the crisis. However, institutional aspects and grant review processes differ between scientific domains and countries and a unified solution is not in sight. Still, two-stage analyses and usage of the Open Science Framework[7] as recommended by Nuzzo (2014) could help.

In 2016, Baker and Penny (2016) published an article about the results of a survey conducted among 1576 international scientists about reproducibility as well as the causes of the reproducibility crisis. Baker and Penny (2016) showed in their analysis that low statistical power and poor statistical analysis are among the three top-rated factors for the irreproducibility of research. Also, respondents were asked to rate different approaches to improving the reproducibility: Nearly 90% ticked 'better statistics' (Baker and Penny, 2016, p 454). In the same year the ASA statement mentioned above was published, which offered guidance for the purpose and interpretation of p-values for researchers (Wasserstein and Lazar, 2016), backing the arguments of Nuzzo (2014) and Colquhoun (2014):

> "What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research."
> Wasserstein and Lazar (2016, p. 2)

The six main principles offered in the statement mostly addressed misconceptions about p-values:

> 1. P-values can indicate how incompatible the data are with a specified statistical model.
> 2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
> 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
> 4. Proper inference requires full reporting and transparency.
> 5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
> 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.
>
> Wasserstein and Lazar (2016, p. 2)

While showing which problems may be the cause for the reproducibility crisis the only advice given to researchers, namely to supplement or replace p-values with other approaches, was given by advising methods

[7] The Open Science Framework: https://osf.io/

> “that emphasise estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modelling and false discovery rates.”
> Wasserstein and Lazar (2016, p. 2)

The discussion followed but one year later Matthews et al. (2017) summarized the implications and what changed (or not) in his article *'The ASA's p-value statement, one year on'*. There, he concluded:

> “Yet a year on, it is not clear that the ASA's statement has had any substantive effect at all. A quick check of the latest issues of leading journals like The Lancet or Proceedings of the National Academy of Sciences shows it's business as usual - even in papers submitted after the ASA's statement.”
> Matthews et al. (2017, p. 38)

Without few exceptions (one of them the complete banishing of p-values from the scientific journal *Basic and Applied Social Psychology* (Matthews et al., 2017, p. 39) (which itself is a consequence of the replication crisis starting earlier in the 2010s and not of the ASA statement), according to Matthews et al. (2017), the ASA statement has not changed the habits of researchers at all. Matthews et al. (2017) underlined that the ASA statement despite its criticism gave no clear vision of which statistical methods should be used in science instead of p-values. According to him, the “workaday researcher's question goes unanswered: how do I turn my data into insight?” (Matthews et al., 2017, p. 40). He argued that for decades the standard methodology of hypothesis testing was shaped by the monograph *'Statistical methods for research workers'* by british statistician Ronald Fisher. He also questioned Fisher's *“one paragraph dismissal of Bayesian methods in the introduction to the book.”* (Matthews et al., 2017, p. 40), which among the methods proposed as alternatives to p-values by the ASA.

Somewhat simplifying the situation, Matthews et al. (2017) suggested that nearly a century after the original publication the take-away message of Fisher's work has “seeped into software, lecture courses and countless “statistics for scientists” texts: inference is simply a game - and anyone can play it” (Matthews et al., 2017, p. 40). Matthews work shows that an important aspect is the evolution of statistical inference as a mathematical concept itself for understanding the replication crisis.

However, it is not clear that Ronald Fisher's approach seeped into software, lecture courses and statistic texts for scientists. The situation is more complex: Statisticians today separate at least between three different kind of inference models: *Frequentist Inference*, *Bayesian Inference* and *Fisherian Inference* (Efron, 1998). In most discussions about the replication crisis, Fisher-Inference is set equal to the frequentist approach of statistical inference and in particular to null hypothesis significance testing. However, this clearly is an incomplete picture of the connections between various modes of inference as shown in Part I.

1.2 Research Question and Contribution

1.2.1 Research question

While substantial progress has already been made in the last years in identifying the causes of the replication crisis, one major problem is that the evolution of the underlying statistical theory and its role in the recent replication crisis has not been analyzed in detail. The majority of solutions which have been proposed are either motivated by personal conviction or only give a vague reference that the historical developments played a key role in the problems observed today (Nuzzo, 2014; Ziliak, 2019). The main research question can thus be formulated as follows:

> **How can the problems with statistical hypothesis testing which have been observed during the scientific replication crisis be solved by changing statistical practice based on a detailed reconstruction of the evolution of statistical hypothesis testing?**

While it is easy to blame the historical developments partially for today's situation, tracing the evolution of hypothesis testing to the points where things started to derail is difficult and important to explain the causes of the replication crisis. A detailed analysis of the evolution of statistical hypothesis testing is thus a preliminary requirement to develop possible statistical solutions to the replication crisis in a second step.

1.2.2 Contribution

The contribution of this thesis is to investigate the evolution of hypothesis testing and develop statistical solutions to the scientific replication crisis with an emphasis on statistical models in the biomedical sciences. Taking the results of Ioannidis (2005b), Nuzzo (2014), Colquhoun (2014, 2017), Benjamin et al. (2018), Baker and Penny (2016) and Matthews et al. (2017) as a motivation, in this thesis the evolution of statistical hypothesis testing is reconstructed and it is shown that various of the recently observed problems with the reproducibility of research can be attributed to the 1) underlying statistical theory, 2) computational obstacles, 3) axiomatic problems, in particular, inconsistencies with the foundations of statistical inference and 4) the application context, that is, the inappropriate use of a statistical method that was never designed to be used in such a context. Statistical solutions to some major problems of the scientific replication crisis are provided based on the results of this reconstruction.

There are four core aspects which play a substantial role in the evolution of statistical hypothesis testing and which need to be taken into account in the reconstruction of the evolution of statistical hypothesis testing, and before statistical solutions are proposed in a second step:

1. **The evolution of mathematical theory**
   Without the existence of an adequate mathematical theory to test a hypothesis in a statistical manner, no research hypothesis in a study or experiment can be analyzed at all. The analysis of the underlying mathematical theory of the competing approaches to statistical hypothesis testing is thus a necessary first step.

2. **The availability of computational resources**
   The availability of computational resources comprise the possibility to compute theoretically sound algorithms backed by a sufficient mathematical theory. While the existence of a rigid mathematical theory is necessary in the first place, the availability of computing power may limit the everyday application of methods used by researchers in the second place, or as Efron (1998, p. 112) stated: "Equipment is destiny in science, and statistics is no exception to that rule. Second, statisticians are being asked to solve bigger, harder, more complicated problems, under

such names as pattern recognition, DNA screening, neural networks, imaging and machine learning." The analysis of the computational obstacles that prevented a more widespread use are, in particular, relevant in the reconstruction of the Bayesian approach to statistical hypothesis testing.

3. **Philosophical and axiomatic aspects**: While mathematical correctness is a first requirement for any statistical analysis, philosophical issues arise in practice. These are connected to the problem of induction (Popper, 1959; Mayo, 2018) and have to be considered. The underlying scientific theory and justification of statistical approaches to hypothesis testing from a philosophy of science perspective is important. Furthermore, the axiomatic justification of these different modes of inference is an important and highly controversial topic in statistical science (Birnbaum, 1962; Berger and Wolpert, 1988), and any solution presented to the replication crisis must obey certain axiomatic implications which follow from the foundations of statistical inference and theoretical statistics. Thus, proposed statistical solutions need to incorporate the results of the reconstruction of both philosophical and axiomatic aspects.

4. **The application context**: Although mathematical theory, computational resources and philosophical and axiomatic aspects play a crucial role in the evolution of hypothesis testing, often the context of application is of substantial importance. As shown in Part I, the two main theories for frequentist hypothesis testing differ strongly in the intended application context, which offers first insights why the current status quo of statistical hypothesis testing may be seen as highly problematic. Also, as shown in Part II and Part IV, statistical hypothesis testing can be loosely coupled with a supporting theory of science or be designed as such a theory from the start. Contextual and practical aspects also play a major role in the development of computational resources as shown in Part III and for the development of statistical solutions as presented in Part V.

1.2.3 Chapter outline

The first two parts of the thesis are concerned with the first of the above four points and they reconstruct the evolution of mathematical theory for statistical hypothesis testing.

Part I: The Evolution of Frequentist Significance- and Hypothesis Testing

In Part I, it is shown that current dominating statistical methodology is the product of an inconsistent hybrid approach to statistical hypothesis testing which emerged out of Fisher's theory of significance tests, p-values and the Neyman-Pearson theory. In Chapter 2 and Chapter 3, the development of Fisher's theory of significance tests and his introduction of p-values is detailed. Chapter 4 presents the evolution of the Neyman-Pearson theory for statistical hypothesis testing and Chapter 5 contrasts both approaches, their application context, the statistical differences and the consequences of the predominant use of an inconsistent hybrid approach which emerged out of both theories for the replication crisis today. The reconstruction shows why the current status quo of NHST is highly problematic and clarifies that the dominant approach to statistical hypothesis testing today was never intended by the creators of the underlying statistical theories in such an application context.

Part II: The Evolution of Bayesian Hypothesis Testing

In Part II, the evolution of Bayesian approaches to hypothesis testing is detailed with a focus on the Bayes factor as a possible alternative to the current status quo. Chapter 6 outlines the basics of Bayesian statistics and contrasts these with the frequentist approach, and Chapter 7 details the evolution of the Bayes factor as an alternative to frequentist hypothesis tests based on p-values. It is shown that although the Bayes factor approach did not succeed in the decades that followed, the primary reasons were mostly computational aspects which prevented a more widespread use of Bayesian methods across science. Also, the differences between the frequentist and Bayesian approach are analyzed and it is shown that the more appropriate approach for hypothesis testing in scientific contexts is the Bayesian one.

Part III: The Evolution of Markov-Chain-Monte-Carlo

The following Part III then discusses the development of modern Markov-Chain-Monte-Carlo algorithms and their impact on Bayesian hypothesis testing and thus is concerned with the second point above by reconstructing the evolution of statistical hypothesis testing with a focus on computational aspects. Chapter 8 provides the basics of Markov-Chain-Monte-Carlo (MCMC), and Chapter 9 details the Markov-Chain-Monte-Carlo revolution, which introduced a Bayesian renaissance from a statistical perspective. It is shown that the development of these methods has simplified Bayesian hypothesis testing substantially and that the largest hurdle in employing Bayesian hypothesis tests has been removed through the advent of modern Markov-Chain-Monte-Carlo algorithms. Also, it is outlined why the burden of manual calibration of these algorithms has been taken from researchers through the introduction of modern Hamiltonian-Monte-Carlo algorithms which present a new generation of MCMC algorithms.

Part IV: On the axiomatic Foundations of Statistical Inference

Part IV then is concerned with the third point above and reconstructs philosophical and axiomatic aspects in the evolution of statistical hypothesis testing. Chapter 10 provides a justification why Bayesian inference can be accepted as a probabilistic version of enumerative induction, and discusses arguments from philosophers of science against enumerative induction and thus against Bayesian inference as a grounded scientific theory. It is shown why Bayes' theorem can be interpreted as a statistical implementation of probabilistic enumerative induction, which justifies Bayesian hypothesis tests for use in scientific contexts. Chapter 11 then provides a detailed axiomatic analysis of the concept of statistical evidence in the context of hypothesis testing. It is shown that the majority of observed problems in the replication crisis are due to purely axiomatic reasons and conflicts with the likelihood principle. Based on the axiomatic analysis, Chapter 11 then shows that robust Bayesian methods, in particular robust Bayesian hypothesis tests provide a solution to some substantial problems with the reproducibility of research.

Part V: Bayesian statistical solutions to the replication crisis

Bayesian statistical solutions to the replication crisis are provided in Part V with a focus on Bayesian biostatistical and biometrical models which are widely used in medical research and the cognitive sciences. In this part, the results from the reconstruction of the evolution of statistical hypothesis testing are incorporated. While Part I and Part II show in their analysis of the underlying mathematical theory that the application context as listed in the fourth point above can be seen as a primary reason why the problems in the replication crisis are observed, Part III and Part IV provide the justification for the developed statistical solutions in this Part V from a philosophy of science and axiomatic perspective. In Chapter 12 it is shown that robust Bayesian hypothesis tests based on the Bayes factor are available for the majority of statistical models used in biomedical research. Chapter 13 demonstrates that even complex models like parametric survival models in medical statistics become tractable by use of the Hamiltonian-Monte-Carlo algorithms as discussed in Part III. Chapter 14 provides new results which show that the implicit error control of Bayesian hypothesis tests is comparable to frequentist tests based on p-values, and that a variety of Bayesian evidence measures attains reasonable type I error control and power in practice in parametric two-sample tests. Chapter 15 then proposes a shift towards the Hodges-Lehmann paradigm which advocates testing small interval instead of point null hypotheses, and new theoretical results show that such a shift may be an appealing additional step towards increasing the reproducibility of science which has not received enough attention in the discussion about the validity of statistical hypothesis testing. In Chapter 16 the dissertation is concluded by revisiting the replication crisis in light of the proposed statistical solutions and provides a discussion for future research.

Part I

The Evolution of Frequentist Significance - and Hypothesis Testing

Chapter 2

The Protagonists

> The record of a month's roulette playing at Monte Carlo can afford us material for discussing the foundations of knowledge.
>
> Karl Pearson

This section gives a short overview over the persons who played a major role in the evolution of frequentist significance and hypothesis testing and who invented the classical approach of testing hypotheses in a statistical manner. This serves to frame the classical frequentist theory into its application context later, in particular, in Part IV.

Karl Pearson

Karl Pearson (1857-1936) can be considered as the initiator of modern statistics. Pearson studied mathematics in Cambridge and was interested strongly in philosophy and theology (Porter, 2006). Although the British statistician Ronald Fisher and polish mathematician Jerzy Neyman basically created the discipline of modern statistics from scratch, their work would have been impossible without the previous work of Karl Pearson. Pearson mainly invented two things, which heavily influenced the work of Fisher and many others:

1. He introduced in 1895 his system of Pearson curves, which today is not used anymore but at that time was the quasi-standard for defining different probability densities (Porter, 2006).

2. In 1902, he proposed the method of moments for estimating parameters of the Pearson curves under given data and thus invented one of the first parameter estimation techniques (Morant, 1939; Stigler, 2008).[1]

Principally, the method of moments was Fisher's first reason to use maximum likelihood in his 1912 paper (Fisher, 1912) although he did not explicitly call it maximum likelihood then. Also, Fisher was a great admirer of Pearson's achievements, as can be

[1] A modern introduction into the method of moment estimation is given in Rüschendorf (2014, Chapter 5).

seen in his early writing style, often referring to Pearson and his achievements in the newly founded research area of mathematical statistics (Fisher, 1922a).

Next to the two concepts mentioned above, Pearson's biggest achievement was the development of the χ^2-test for the goodness of fit. Until then, the strategy to determine whether a given series of observations could be assumed to belong to a normal distribution was to see if the fit by eye was acceptable. The χ^2-test introduced by Pearson offered a quantifiable way that was more reliable and a first statistical test, which established itself quickly among mathematicians. Next to these achievements, Pearson wrote the influential book *'The Grammar of Science'* and founded the first statistics department in London. Furthermore, his work on the correlation coefficient, eugenics and psychology is worth mentioning, for details see Porter (2006).

Ronald Aylmer Fisher

Fisher was born in London, on February 17, 1890 as a surviving twin. George Fisher, his father, was an auctioneer for fine art. What is known from the biography written by his daughter Joan Fisher Box, Fisher went to Harrow School in 1904 and was good in mathematics from the very beginning (Box, 1978). He entered Gonville and Caius College, Cambridge in 1909 and graduated three years later. His statistical career started with his 1912 paper. A lot of his early work was influenced by the correspondence to William Sealy Gosset, known better under the pseudonym Student. However, in the early work of Fisher, likelihood was already used but ironically never mentioned. Fisher spent a postgraduate year in Cavendish Laboratory, Cambridge under supervision of F.J.M. Stratton, studying the theory of errors. He stayed at Cambridge another year, studying the theory of errors, and in 1914 he volunteered for military service. He was rejected due to his poor eyesight and therefore spent the next five years as a high school teacher for mathematics and physics. Also in 1914, Fisher worked on the exact derivation of the distribution of the correlation coefficient r and sent his derivation to Karl Pearson for publication in the journal *Biometrika* (Fisher, 1915). Pearson accepted Fisher's paper but wrote one two years later together with H.E. Soper, where parts of Fisher's paper were criticised for using inverse probability – better known as Bayesian statistics today – in his derivation (Soper et al., 1917). A feud started between Fisher and Pearson which resulted in two rejected papers of Fisher in the following years. In the meantime, Fisher married Ruth Eileen Guiness in 1917 and was offered a position as leading statistician at the Galton Laboratory in 1919. However, he rejected due to concerns regarding his ability to publish anything in this position as Karl Pearson was involved in the Galton Laboratory, too. Instead, he accepted a position as statistician at Rothamsted Experimental Station. Rothamsted was an agricultural experimental station in Harpenden, 25 miles north from London (Box, 1978, Chapter 4). After his first year, Fisher's contract was extended to a permanent assignment, and he stayed at Rothamsted until 1933. His work mainly consisted of analyzing agricultural records of the last years, and improving the methods used in agricultural work. In these years, Fisher's theory of significance testing was widely popularized as a framework for statistical hypothesis testing, partially because of his influential textbook *Statistical methods for research workers* (Fisher, 1925a).

In the same year Fisher quit at Rothamsted and Karl Pearson retired from University College London as head of the applied statistics department. The department was split and his son Egon Pearson followed him as head of the department of statistics, and

Fisher was assigned head of the department of eugenics so that he never worked as a professor for statistics.

In 1939, with the beginning of WWII, Fisher's department at University College was evacuated and until 1943, Fisher did not find a new position. Then, he was appointed to the Arthus Balfour Chair of Genetics at Cambridge University. Until his retirement in 1957 he remained at Cambridge and spent the remaining years of his life in Australia at Adelaide until he died in 1962.

Fishers definitive biography has been written by his daughter Joan Fisher Box (Box, 1978). Fisher's life and especially work has been reviewed also by Savage (1976), Rao (1992), Efron (1998), Healy (2003) and Stigler (2006). While all of these provide detailed insights to specific aspects of Fisher's work, none of them has investigated the evolution of hypothesis testing to which Fisher is directly connected with regard to the scientific replication crisis. Fisher's work, as noted by various authors (compare Chapter 1), of course plays a major role in understanding the causes of the replication crisis.

Jerzy Neyman

Jerzy Neyman (1894-1981) was a polish statistician. The first half of his life he spent in Europe, until he emigrated to America in 1938 where he spent the rest of his life. He was born in Russia and in 1912 entered the University of Kharkov, majoring in both physics and mathematics. According to Reid (1982), the lectures of Lebesgue motivated Neyman to focus on mathematics instead of physics, and he read 'Lessons on the integration and the research of the primitive functions' by Henri Lebesgue. However, how strong the influence of this early connections to measure theory were for his later career remains unknown. It is known that he also visited lectures of Sergei Natanowitsch Bernstein which dealt with probability theory (Reid, 1982).

In 1921, Neyman went to Warsaw to work at the Agricultural Institute in Bydgoszcz. Later, he switched to the national Meteorological Institute and finally got a position at the University of Warsaw as an assistant. He obtained his doctorate in 1924 also at Bydgoszcz for a dissertation titled *On the Applications of the Theory of Probability to Agricultural Experiments*. After getting a fellowship he spent a couple of years in London and Paris to work with Karl Pearson and Émile Borel. In London, he got in touch with Karl Pearson's son Egon Pearson who also tried to make a career as a statistician.

In 1926, Egon Pearson sent him a letter with ideas about which statistical problems to collaborate on, and a collaboration started which yielded the Neyman-Pearson of hypothesis testing. After the year in Paris, Neyman returned to Warsaw and had a tough time to work for a minimum living. In 1934, the situation improved, and Egon Pearson was able to offer Neyman a position as new head of the department of statistics at University College, London. After a year, the position was extended into a permanent one and Neyman could concentrate on his research. Still, 2 years later, in 1937, he was offered a professorship at the University of California at Berkeley, and he went to the United States. There, he remained professor until his retirement and later death in 1981. His biggest achievement can be seen in the Neyman-Pearson theory of hypothesis testing, which he developed together with Egon Pearson.

The definitive reference to Jerzy Neyman is his biography by Constance Reid (Reid, 1982). Other references include Kendall et al. (1982), who provide insights to Neyman's academic contributions in their obituary, and Salsburg (2001). Also, another valuable source is Pearson (1966).

William Sealy Gosset ('Student')

Next to Karl Pearson, a person who had a crucial impact on the development of the statistical hypothesis testing was William Sealy Gosset (1876-1937). He studied chemistry and mathematics at the New College in Oxford and in 1899 took a position at the Arthur Guinness Son and Co. brewery. At work he was soon confronted with a lot of statistical problems for which no solution was available. Therefore, he educated himself with textbooks and corresponded with Fisher which resulted in a very productive cooperation. Gossets most popular achievement was his 1908 publication, in which he initiated under his pseudonym Student (he was not allowed to publish while working for Guinness) a completely new statistical approach (Student, 1908b). Gosset's idea was basically to test the value of a population mean, and by 1908, it had been customary to choose the well-known Student's t statistic, and compare it to a normal distribution. Of course, the Student's t statistic is not normally distributed and therefore, only for large samples the approximation holds.[2] Gosset wanted to find a formula for small samples, where the approximation did not hold. He obtained his results by restricting the form of the distribution of the observations to a normal distribution and then derived his popular small-sample formula. While he was not able to give a definitive proof, Fisher (1915) obtained the first proof a few years later and both of them worked together on the correlation coefficient for small samples. Gosset had already tried to derive it in another paper earlier (Student, 1908a). Fisher (1921b) found a proof, and from then on several small sample statistics were derived by them which were of significant help in practical analysis because often the sample size was limited so asymptotic arguments could not be used to determine the distribution of a statistic of interest. The collaboration culminated in the publication of Fisher's 1925 book *'Statistical Methods for Research Workers'* (Fisher, 1925b). Gosset can therefore be considered as an important catalysator for Fisher's early work. Gosset worked at Guinness until he died at the age of 61.[3]

Egon S. Pearson

The last person of interest is the son of Karl Pearson, Egon S. Pearson, who was born in 1895 and died in 1980. Egon Pearson went to Cambridge to study mathematics and obtained his degree in 1919. He continued with graduate studies in astronomy and in 1921 finally joined his father's Department of Applied Statistics at University College, London as a teaching assistant. Five years later, in 1926 he began to teach on his own and finally in 1933, when Karl Pearson retired, he was appointed chair of the Department of Statistics. Egon Pearson took particular interest in Fisher's 1925 book *'Statistical Methods*

[2] Expressed differently, the t_n density converges in probability to a standard normal density $\varphi_{0,1}$ for $n \to \infty$. For small n, the distribution of the t_n density was unknown at that time.

[3] In 1939, Fisher in an obituary for 'Student' noted:

> "The untimely death of W. S. Gosset, at the age of 61, in October 1937, has taken one of the most original minds in contemporary science. Without being a professed mathematician, he first published, in 1908, a fundamentally new approach to the classical problem of the theory of errors, the consequences of which are still only gradually coming to be appreciated in the many fields of work to which it is applicable."
> Fisher (1939, p. 1)

for Research Workers' as he later wrote:

> "I was in a state of puzzlement, and realized that, if I was to continue an academic career as a mathematical statistician, I must construct for myself what might be termed a statistical philosophy, which would have to combine what I accepted from K. P.'s large-sample tradition with the newer ideas of Fisher."
> Pearson et al. (1990, p. 77)

He exchanged letters with Gosset and like Gosset already did in the collaboration wish Fisher, he functioned as a catalysator for the Neyman-Pearson theory of hypothesis testing. Gosset and Pearson exchanged opinions about testing for the mean of a sample and Gosset wrote:

> "In your large samples with a known normal distribution you are able to find the chance that the mean of a random sample will lie at any given distance from the mean of the population. (Personally I am inclined to think your cases are best considered as mine taken to the limit n large.) That doesn't in itself necessarily prove that the sample is not drawn randomly from the population even if the chance is very small, say ·00001: what it does is to show that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability, say ·05 (such as that it belongs to a different population or that the sample wasn't random or whatever will do the trick) you will be very much more inclined to consider that the original hypothesis is not true."
> W.S. Gosset, in Pearson (1939, Letter I, dated May 1926, p. 243)

This had a huge impact on the problems Egon Pearson decided to work on, as Pearson himself later remembered:

> "Gosset's reply had a tremendous influence on the direction of my subsequent work, for the first paragraph contains the germ of that idea which has formed the basis of all later joint researches of Neyman and myself."
> Pearson (1939, p. 242)

In total, 'Student' therefore not only initiated Fisher's interest in small sample statistics, but also gave Egon Pearson an interesting unsolved problem, which was highly important in the statistical community then and started the Neyman-Pearson theory. In the same year, 1926, Egon Pearson started the collaboration with Jerzy Neyman.

Chapter 3

Fisher's Theory of Significance Testing

> We may discuss the probability of occurrence of quantities which can be observed or deduced from observations, in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of the probability of hypotheses or hypothetical quantities.
>
> Ronald Aylmer Fisher
> *Studies in crop variation*

Ronald Fisher can be described as the founder of modern statistics. Until the work of Fisher, statistics was a scientific discipline which had no precise terminology and also lacked a clear mathematical foundation to answer the practical problems which were often faced in agriculture, medicine or physics. Although Bayesian inference was well known since centuries, it was refused by most statisticians in the early 20th century. The reasons are twofold: First, modern measure theory was not invented at that time which troubled the application of Bayesian data analysis. Second, the claim of subjectivity was often made against the use of Bayesian inference. It remained unclear how to elicit the prior distribution of a parameter in practice, and as statistical science was only an emerging subdiscipline of mathematics, researchers searched for objective procedures which were optimal in some sense (Howie, 2002).

Fisher's work can be divided roughly into two distinct parts: In the first part of his career, he mainly worked on deriving his maximum likelihood theory, including a complete theory of estimation and the necessary statistical vocabulary. This period is marked by a frequent exchange with W.S. Gosset and the derivation of multiple distributions of practically relevant test statistics.[1] At this time, Gosset worked at the Guiness

[1] A test statistic, from a modern perspective, is a decision rule $d : (\mathcal{X}, \mathcal{A} \rightarrow (\Delta, \mathcal{A}_\Delta)$ of the sample space $(\mathcal{X}, \mathcal{A})$ into a measure space $(\Delta, \mathcal{A}_\Delta)$, where in practice, often $(\Delta, \mathcal{A}_\Delta) := (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Examples are the sample mean $d(x) := \bar{x}_n := \frac{1}{n}\sum_{i=1}^n x_i$ or the sample variance $d(x) := s_n^2 := \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$. For details see Appendix C. The exact distribution of these "test statistics" was unknown at that time (e.g. when the observed data $X \sim \mathcal{N}(\mu, \sigma^2)$ one can show that $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$), but researchers and statisticians wanted to judge if an observed sample mean or variance indicates strong deviations from the assumed distribution for the data X.

brewery and was challenged with a variety of statistical tasks for which no solutions were available. The exchange with Fisher led to a variety of solutions which Gosset could apply in his everyday work. For Gosset, the distribution of a specific parameter (like alcohol percentage of the beer) was a quantity of practical interest. For example, the empirical mean as a statistic for the mean parameter of a given distribution was often computed in practice. To decide if a hypothesis about the parameter was reasonable or not, the distribution of this statistic (under infinite repetition and assumption of the hypothesis) was required. Based on this distribution, researchers could then decide if the observed test statistic value was significantly deviating from the value specified in the hypothesis or not. To follow this general procedure, the corresponding distributions of practically relevant test statistics had to be derived. Examples of this first part of Fisher's work include the derivation of the distribution of the correlation coefficient and the derivation of the t-statistic. Over a decade, Fisher and Gosset derived one solution after the other to the problems Gosset faced in his everyday work at Guinness. However, Fisher himself was more interested in developing a general statistical theory, while Gosset was inspired more strongly from his everyday work. Together with Gosset, Fisher was involved in the creation of statistical hypothesis testing, so investigation of his foundational papers is essential to understand the replication problems arising in science today. Fisher's methodology concerning the testing of hypotheses can be attributed to one of the earliest approaches that established itself in the scientific community, as well as the cornerstones for the later development of the Neyman-Pearson theory (compare Chapter 4).

The second part of Fisher's career is concerned with a different goal: The focus of Fisher's work shifted from estimation theory and the derivation of distributions to hypothesis testing, or in his words: significance testing. In this second part of his work, Fisher was head statistician at Rothamsted experimental station, and challenged with the analysis of agricultural data. Next to significance testing, he was also concerned with minute experimental design and statistical methods to analyze experiments, which can be attributed to his work at Rothamsted station. For example, hypotheses like "Is the crop yield larger for a specific sort of potatoes than for another sort of potatoes?" had to be tested, and therefore the focus of Fisher's work shifted.

In summary, while the first part of Fisher's career was concerned both with propagating his maximum likelihood theory as a universal theory for parameter estimation in statistical models and deriving a variety of distributions for statistics, the second part shifted towards more applied work which was inspired by the agricultural context at Rothamsted. There, the importance of hypothesis testing and connecting statistical theory to scientific theory dominated the work of Fisher.

The next section details the achievements of Fisher's early papers to get an overview of these early developments.

3.1 Fisher's Foundation of Estimation Theory

Whereas Fisher's career started as early as 1912 with the Metron publication (Fisher, 1912), terms like likelihood, sufficiency or even hypothesis in a statistical interpretation were not invented then. While Fisher in his early papers like (Fisher, 1915), (Fisher, 1918) and (Fisher, 1920) was already interested in the derivation of specific distributions or characteristics of these, the main vocabulary of modern statistics was laid out in Fisher's 1922 article *'On the mathematical foundations of theoretical statistics'* (Fisher, 1922b), which is also discussed by Geisser (1980). To see how the basic principles of modern statistics and hypothesis testing were invented, attention is directed first at the 1922 paper of Fisher.

3.1.1 The Cornerstone: Mathematical Foundations of Theoretical Statistics

Fisher started his paper with 15 definitions of the necessary vocabulary to distinguish several mathematical objects. Specifically, in the list for the first time, the terms 'statistic', 'scaling', 'sufficiency', 'location', 'efficiency' and 'consistency' were introduced. Also, the term 'maximum likelihood' (ML) was mentioned for the first time in the paper (Fisher, 1922b, p. 323). Furthermore, one of the most crucial distinctions was made by Fisher for the first time:

> "... it is customary to apply the same name, mean, standard deviation, correlation coefficient, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation."
> Fisher (1922b, p. 311)

The distinction made by Fisher is important, in that he first separated between the population parameters, which cannot be precisely known from the sample and have to be estimated, and the *statistic* which is based only on the sample. Such a statistic[2] can be an *estimator* (see Appendix C) for the true population parameter. This distinction was not precisely made before, and Fisher presented a more structured way to work on statistical topics in general. While some statisticians protested against the use of these new terms, Fisher's terminology succeeded. According to Bennett (1990), the mathematician Arne Fisher (1887-1944) was especially offended by the term statistic:

> "I am more inclined to quarrel with you over the introduction by you in statistical method of some outlanding and barbarous technical terms. They stand out like quills upon the porcupine, ready to impale the skeptical critic. Where, for instance, did you get that atrocity, a statistic?"
> Arne Fisher in Bennett (1990, p. 311-313)

Fisher replied to the latter:

> "I use special words as the best way of expressing special meanings. Thiele and Pearson were quite content to use the same words for what they were estimating and for their estimates of it. Hence the chaos in which they left the problem of estimation."
> Ronald Fisher in Bennett (1990, p. 311-313)

Thus, previously statisticians like Karl Pearson did not separate these terms carefully from each other by the time Fisher published his paper and therefore made the whole discipline look like an unstructured approach, leaving much space for ambiguity. After the introduction of the fifteen terms, Fisher set the stage for the general aims of statistical methods:

[2] As noted above, in modern words, a statistic corresponds to a (randomised) decision rule, which is an estimator for a parameter of the statistical model, see Appendix C.

> "... to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: (...) the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or (...) in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data."
> Fisher (1922b, p. 311)

Fisher stated his vague definition of probability directly afterwards:

> "This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion."
> Fisher (1922b, p. 311)

Fisher's probability concept was for the first time explicitly stated in the 1922 paper. Later on, it played a major role in the feud between Fisher and Neyman-Pearson about the way hypotheses should be tested. After concentrating on the reduction of data as the statistician's first goal, the third chapter of Fisher's paper then introduced the three types of problems arising as obstacles for this aim:

> 1. "Problems of Specification - These arise in the choice of the mathematical form of the population.
> 2. Problems of Estimation - These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.
> 3. Problems of Distribution - These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."
>   Fisher (1922b, p. 313)

The first problem is the well-known model choice problem: As one cannot precisely know which statistical model for the population is true, there is uncertainty in the model specification. For example, one could specify a model for the population which follows a normal distribution, or a Cauchy distribution. Maybe both considered options are wrong, which complicates the specification of the form of the population.

The second problem refers to parameter estimation for a fixed statistical model. When a model is selected for the population, its parameters remain unknown. In the example of the normal distribution under consideration above, the mean and standard deviation need to be specified. These population parameters need to be estimated via sample statistics like the empirical mean or the empirical standard deviation. However, there are multiple available options, and it remains unclear how to derive an estimator, in general, for a given population parameter.

The third problem aims at hypothesis testing: Even when an estimator is selected for a given parameter in a fixed statistical model, it remains unknown which values can be expected in practice. However, to decide if a research hypothesis like $H_0 : \mu = 0$

about a population parameter is reasonable or not, one needs to know which values of the statistic are expected under H_0. Mathematically, this requires the distribution of the sample statistic (like the empirical mean) under the assumption of the hypothesis H_0.

To illustrate, in the above example, the distribution of the empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is given as $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, because under the assumption of the statistical model $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, ...n$ where $\mu = 0$, one obtains

$$\mathbb{E}[\bar{X}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n} \cdot n\mathbb{E}[X_i] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

because of the linearity of the expectation and

$$\mathbb{V}[\bar{X}] = \mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\mathbb{V}[\sum_{i=1}^{n} X_i] = \frac{1}{n^2} \cdot n\mathbb{V}[X_i] = \frac{1}{n} \cdot \sigma^2$$

because X_i are independent and identically distributed (i.i.d.). As a consequence, observing a sample statistic value $\bar{X} = 5$ under $H_0 : \mu = 0$ is quite implausible, because this value is located in the tails of the $\mathcal{N}(\mu, \sigma^2/n)$ distribution when for example $\sigma^2 = 1$ and $n = 10$.

Thus, Fisher argued that if the first problem is solved and the model is chosen, the mathematical form is identified correctly. From today's perspective, model selection still plays a major role in statistics and remains an unsolved problem which is addressed in practice with computational methods. When the second problem is solved and an estimator is selected, one can calculate estimates from the sample for parameters of interest. If the third problem is solved, the exact form of the distribution can be determined by the sample, and then *'the theoretical aspect of any particular body of data has been completely elucidated.'* (Fisher, 1922b, p. 314).

Fisher was concerned with the second and third problem over his career. In the fourth chapter he then gave the three essential properties a good estimator should possess: consistency, efficiency and sufficiency.

> "The common-sense criterion employed in the problems of estimation may be stated thus: – That when applied in the whole population the derived statistic should be equal to the parameter. This may be called the Criterion of Consistency."
> Fisher (1922b, p. 316)

Using the absolute mean error $\sigma_1 := \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum|x_i - \bar{x}|$ and the mean squared error $\sigma_2 := \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$ as estimates for the standard deviation of a normally distributed i.i.d. population, that is, X_i i.i.d. and distributed as $\mathcal{N}(\mu, \sigma^2)$, Fisher showed that both of these estimates are consistent. Consistency was a crucial property of an estimator according to Fisher. It was even more important than unbiasedness for him, as Bennett (1990, p. 196) notes. After demonstrating the property of consistency, Fisher went on with efficiency:

> "Consideration of the above example will suggest a second criterion, namely: – That in large samples, when the distribution of the statistics tend to normality, that statistic is to be chosen which has the least probable error. This may be called the Criterion of Efficiency."
> Fisher (1922b, p. 316)

As the Cramér-Rao Lower Bound was not discovered then (compare Theorem C.44), Fisher defined efficiency by calculating the ratio between the "probable error of the statistic calculated (...), and that of the most efficient statistic which could be used. The square ratio of these two quantities then measures the efficiency." (Fisher, 1922b, p. 316). While there was no general lower bound available in 1922, the idea was the same which is still used today, see Definition C.48. Using as examples the absolute mean error $\sigma_1 := \frac{1}{n}\sqrt{\frac{\pi}{2}}\sum|x_i - \bar{x}|$ and the mean squared error $\sigma_2 := \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$ as estimates for the standard deviation of a normally distributed i.i.d. population, Fisher argued that the squared error is preferable because of its lower large sample variance. Fisher then argued that different methods of calculation might tend to the same results for large samples, but differ for smaller samples, which motivated the criterion of sufficiency:

> "... the statistic chosen should summarise the whole of the relevant information supplied by the sample. This may be called the Criterion of Sufficiency."
> Fisher (1920, p. 316)

Fisher added as an explanation:

> "In mathematical language we may interpret this statement by saying that if θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of the distribution of pairs of values of θ_1 and θ_2, for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2 does not involve θ."
> Fisher (1920, p. 316/317)

Comparison with Definition C.50 shows that this concept has been extended into its measure-theoretic version since the introduction by Fisher in 1922, but the idea has stayed the same. In fact, the modern measure-theoretic definition contains the original definition of Fisher, see Schervish (1995). Fisher had already discovered sufficiency in (Fisher, 1920), but sufficiency was not defined explicitly there.

One of the key elements of the 1922 paper was the first appearance of the term maximum likelihood.[3] While Fisher already used likelihood in his very first paper in 1912, it was only used as an alternative approach to Karl Pearsons system of error curves and far away from a general estimation theory (Fisher, 1912). In his 1922 paper, Fisher separated that consistency, efficiency and sufficiency were properties of an estimator, but not methods to find any estimator.

> "... the criterion of sufficiency (...) is not of direct assistance in the solution of problems of estimation. For it is necessary first to know the statistic concerned and its surface of distribution, with an infinite number of other statistics, before its sufficiency can be tested. For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum

[3] A concise measure-theoretic introduction to maximum likelihood estimation is given in Rüschendorf (2014, Chapter 5). A brief introduction is provided in Appendix C, and an accessible introduction that assumes no familiarity with measure theory is given in Myung (2003).

> Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect."
> Fisher (1922b, p. 323)

Fisher was critical about his work, inviting readers to "form their own opinion as to the possibility of the method of maximum likelihood leading in any case to an insufficient statistic." (Fisher, 1922b, p. 323). Thus, the theory of maximum likelihood was introduced as a method to find estimators having the desirable property of sufficiency. In general, maximum likelihood estimates need not be sufficient, but when a sufficient statistic exist, the maximum likelihood solutions are a function of the sufficient statistic (Rüschendorf, 2014, Chapter 5).[4] While Fisher gave a proof for his intuition, he was not satisfied with its mathematical rigour. First, he explained the method of maximum likelihood as follows: He supposed a random variable X has density $f(x|\theta_1, \theta_2, ...\theta_r)$, where $\theta_1, \theta_2, ...$ are unknown parameters. Fisher formulated the chance of a single observation to fall into the range dx as

$$\mathbb{P}(x \leq X \leq x + dx) = f(x|\theta_1, \theta_2, ...\theta_r) \cdot dx \tag{3.1}$$

Generalizing his idea to a sample of n observations, he proceeded by stating that the chance of n_1 falling into dx_1, n_2 falling into dx_2, and so on is

$$\frac{n!}{\prod_{i=1}^{p} n_i!} \prod_{i=1}^{p} \{f(x_p|\theta_1, \theta_2, ...)dx_p\}^{n_p} \tag{3.2}$$

Fisher then concluded:

> "The method of maximum likelihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum, and since in this expression the parameters are only involved in the function f, we have to make
>
> $$S(logf)$$
>
> a maximum for variations of $\theta_1, \theta_2, \theta_3$ (...). In this form, the method applies to the fitting of populations involving any number of variates, and equally to discontinuous as continuous distributions."
> Fisher (1922b, p. 323/324)

Here, S stands for sum, because Fisher directly maximized the log-likelihood (compare Definition C.38). This simplifies computations, because the logarithm of products is turned into a sum of logarithms. After introducing his method, Fisher went on to show that maximum likelihood estimates are invariant under one-to-one-transformations, while Bayesian inference does not have this property.

To show that Bayesian inference, in general, does not produce estimates which are invariant to one-to-one transformations, Fisher presented a counterexample. He considered a binomial distribution where the number of successes x out of n trials is known,

[4] The existence of a sufficient statistic is related in turn to the statistical model $\mathcal{P}$. If $\mathcal{P}$ is dominated, there exists a sufficient statistic (even minimal sufficient) (Rüschendorf, 2014, Theorem 4.2.9), and the model being dominated is related to the separability of the underlying metric space $(\mathcal{P}, d_r)$ with regard to the total variation norm d_r, compare Rüschendorf (2014, Theorem 3.1.17).

and the probability p is unknown. At this time, it was usual to stick to Bayes' theorem to solve this problem. One expressed complete ignorance about a parameter by assigning a uniform prior $f(p) = 1$ for all $p \in [0, 1]$ to p. Bayes' theorem then yields

$$f(p|x) \propto p^x(1-p)^{n-x} \cdot 1 \propto p^x(1-p)^{n-x} \tag{3.3}$$

Fisher then strongly critizised this postulate:

> "The postulate would, if true, be of great importance in bringing an immense variety of questions within the domain of probability. It is, however, evidently extremely arbitrary. Apart from evolving a vitally important piece of knowledge, that of the exact form of the distribution of the values of p, out of an assumption of complete ignorance, it is not even a unique solution."
> Fisher (1922b, p. 325)

He considered the transformation $\sin(\theta) = 2p - 1$. Complete ignorance about p is the same as complete ignorance about θ then, and therefore $f(\theta) = \frac{1}{\pi}, -\pi/2 < \theta < \pi/2$ is a uniform prior on $[-\pi/2, \pi/2]$. Then the posterior can be written as

$$f(\theta|x) \propto \theta^x(1-\theta)^{n-x} f(\theta) \propto p^x(1-p)^{n-x} \tag{3.4}$$

because the probability to observe $X = n$ successes given θ is identical to the probability to observe $X = n$ successes given p under the uniform prior. By a change of variables for $g : p \longmapsto \theta, p \longmapsto \sin^{-1}(2p-1)$, and

$$\left|\frac{dg^{-1}(\theta)}{d\theta}\right| = \left|\frac{dp}{d\theta}\right| = \left|\frac{d\sin(\theta)+1}{2d\theta}\right| = \frac{1}{2}\cos(\theta) \tag{3.5}$$

from $|\frac{dp}{d\theta}|^{-1} = \frac{2}{\cos(\theta)}$ one obtains

$$f(p|x) = f(\theta|x)\left|\frac{dp}{d\theta}\right|^{-1} \propto p^x(1-p)^{n-x}\frac{1}{\cos(\theta)} \tag{3.6}$$

As

$$p^{\frac{1}{2}}(1-p)^{\frac{1}{2}} = [p(1-p)]^{\frac{1}{2}} = \left[\frac{1+\sin(\theta)}{2} \cdot (1 - \frac{1+\sin(\theta)}{2})\right]^{\frac{1}{2}} \tag{3.7}$$

$$= \left[\frac{1-\sin^2(\theta)}{2}\right]^{\frac{1}{2}} = \left[\frac{\cos^2(\theta)}{2}\right]^{\frac{1}{2}} \propto \cos(\theta) \tag{3.8}$$

because of $\sin^2(\theta) + \cos^2(\theta) = 1$, Equation (3.6) is proportional to

$$p^x(1-p)^{n-x}\frac{1}{\cos(\theta)} \propto p^x(1-p)^{n-x}\frac{1}{p^{1/2}(1-p)^{1/2}} = p^{x-1/2}(1-p)^{n-x-1/2} \tag{3.9}$$

By obtaining this last equation, Fisher demonstrated that this contradicted Equation (3.4). Therefore, his example showed a problem with inverse probability (in modern language, with Bayesian inference): The missing invariance under parameter-transformations. For example, the modes of both posterior distributions will, in general, differ and lead to differing estimates for the parameter. Fisher then proceeded to show that his maximum likelihood method was invariant under one-to-one-transformations. He used the same example as above, the probability p the unknown parameter of the binomial distribution with n trials and x successes. The likelihood $L(p|x)$ of p is given as

$$L(p|x) \propto p^x(1-p)^{n-x} \tag{3.10}$$

Using the one-to-one transformation $\sin(\theta) = 2p - 1$, the likelihood of θ is given by

$$\begin{aligned} L(\theta|x) &= \left(\frac{1+\sin(\theta)}{2}\right)^x \left(\frac{1-\sin(\theta)}{2})\right)^{n-x} \qquad (3.11) \\ &\propto (1+\sin(\theta))^x(1-\sin(\theta))^{n-x} \qquad (3.12) \end{aligned}$$

Differentiating the right-hand side of Equation (3.10) Fisher obtained the maximum likelihood estimate

$$\hat{p} = \frac{x}{n} \tag{3.13}$$

Proceeding equally for Equation (3.11), he obtained

$$\hat{\theta} = \sin^{-1}\left(\frac{2x}{n} - 1\right) \tag{3.14}$$

Finally, substituting $\theta = \sin^{-1}\left(\frac{2x}{n} - 1\right)$ into $\sin(\theta) = 2p - 1$, he obtained the same maximum likelihood estimator $p = x/n$. With this example Fisher tried to demonstrate that the results produced by his theory of maximum likelihood were more objective than the results produced via the Bayesian approach. In general, maximum likelihood parameters are indeed invariant under one-to-one transformations, compare Held and Sabanés Bové (2014).[5]

Next, Fisher turned to another important problem of estimation: The variance of an estimator, which should be as small as possible.[6] He showed that the variance of estimators obtained via his theory of maximum likelihood were in some sense optimal. Fisher used an asymptotic argument and assumed

$$\theta_1 \sim \mathcal{N}(\theta, \sigma^2) \tag{3.15}$$

Then, he denoted the density function of θ_1 as

$$\varphi_{\theta,\sigma^2}(\theta_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta_1-\theta)^2}{2\sigma^2}\right) \tag{3.16}$$

The likelihood of θ of course is proportional to $\exp(-\frac{(\theta_1-\theta)^2}{2\sigma^2})$ and has its maximum at

$$\theta = \theta_1 \tag{3.17}$$

[5] It took 40 years until Zehna (1966) showed that this desirable property of the method of maximum likelihood also holds in cases when the parameter transformation is not one-to-one. However, notice that Bayesian parameter estimates are invariant under one-to-one transformations, too, when Jeffrey's prior is used, compare Chapter 6.

[6] It is clear from Appendix C, Definition C.48, that a good estimator should possess a small variance.

Fisher noted, that

$$\frac{\partial}{\partial\theta}\log(\varphi_{\theta,\sigma^2}(\theta_1)) = \frac{\theta_1-\theta}{\sigma^2} \tag{3.18}$$

$$\frac{\partial^2}{\partial\theta^2}\log(\varphi_{\theta,\sigma^2}(\theta_1)) = -\frac{1}{\sigma^2} \tag{3.19}$$

Fisher then reasoned that $\varphi(\theta_1)$ is the density of all samples for which the statistic has value θ_1. He denoted ϕ as the density of such a single sample and therefore the density of all samples can be written as $\varphi = \sum\phi$ (Fisher, 1922b, p. 328). If f is the density of an observation in a given sample, then

$$\log\phi = C + \sum f \tag{3.20}$$

where $C \in \mathbb{R}$ is a constant which does not depend on the parameters and the sum is over all observations of the given sample. Using a Taylor series expansion for f around $\theta = \theta_1$, he wrote

$$\log(f(\theta)) = \log(f(\theta_1)) + (\theta-\theta_1)\frac{\partial}{\partial\theta}\log(f(\theta_1)) + \frac{1}{2}(\theta-\theta_1)^2\frac{\partial^2}{\partial\theta^2}\log(f(\theta_1)) + ... \tag{3.21}$$

Rewriting the above as

$$\log(f) = \log(f_1) + a(\theta-\theta_1) + \frac{b}{2}(\theta-\theta_1)^2 + ... \tag{3.22}$$

where

$$f_1 = f(\theta_1) \tag{3.23}$$

$$a = \frac{\partial}{\partial\theta}\log f(\theta_1) \tag{3.24}$$

$$b = \frac{\partial^2}{\partial\theta^2}\log(f(\theta_1)) \tag{3.25}$$

Fisher (1922b, p. 328) obtained

$$\log\phi = C + \sum\log(f_1) + (\theta-\theta_1)\sum a + \frac{1}{2}(\theta-\theta_1)^2\sum b + ... \tag{3.26}$$

$$= C + (\theta-\theta_1)\sum a + \frac{1}{2}(\theta-\theta_1)^2\sum b + ... \tag{3.27}$$

where the term $\sum\log(f_1)$ vanished because θ_1 is the MLE according to Equation (3.17). By the CLT, it follows that

$$\frac{\sum b - n\mathbb{E}(b)}{\sqrt{n\mathbb{V}(b)}} \sim \mathcal{N}(0,1) \tag{3.28}$$

This is equivalent to

$$\sum b - n\mathbb{E}(b) \sim O(n^{1/2}) \tag{3.29}$$

In the same way it follows that $\theta - \theta_1 \sim O(n^{-1/2})$. Fisher therefore concluded, that "the only terms in log ϕ, which are not reduced without limits, as n is increased" are (here Equation (3.26))

$$\log \phi = C + \frac{1}{2} n(\theta - \theta_1)^2 \mathbb{E}(b) \tag{3.30}$$

so that finally

$$\phi \propto \exp \left\{ \frac{1}{2} n(\theta - \theta_1)^2 \mathbb{E}(b) \right\} \tag{3.31}$$

Fisher then argued, that the proportionality constant given in Equation (3.31) was applicable to all samples with the value θ_1, and therefore also for φ. Finally, he obtained from Equation (3.30)

$$\log \varphi_{\theta,\sigma^2}(\theta_1) = C' + \frac{1}{2} n(\theta - \theta_1)^2 \mathbb{E}(b) \tag{3.32}$$

with $C' = C + (\theta - \theta_1) \sum a$. This led to the equation

$$\frac{\partial^2}{\partial \theta^2} \log \varphi_{\theta,\sigma^2}(\theta_1) = n\mathbb{E}(b) \tag{3.33}$$

Thus, in result Fisher this way obtained the variance of the MLE as

$$\begin{aligned} \mathbb{V}(\theta_1) &= \sigma^2 && \text{(3.34)} \\ &= -\frac{1}{n\mathbb{E}(b)} && \text{by Equation (3.33) and Equation (3.19)} \quad \text{(3.35)} \\ &= -\frac{1}{n\mathbb{E}(\frac{\partial^2}{\partial \theta^2} \log f(\theta_1))} && \text{by Equation (3.25)} \quad \text{(3.36)} \end{aligned}$$

Fisher in summary showed that the MLE attains the by then not discovered Cramér-Rao Lower Bound for the variance of an estimator, and thus is in this respect optimal (compare Theorem C.44). In section 7 of the 1922 paper, he then tried to show that MLEs are always sufficient. As indicated already above, he was not satisfied by the rigour of his proof, and indeed he was wrong. MLEs are not always sufficient, but when a sufficient statistic exists, the MLE is a function of the sufficient statistic.

3.1.2 Significance testing and p-values

The last section showed that the basic concepts of statistical science were introduced by Fisher in 1922. However, these developments were less concerned with any form of hypothesis testing, but aimed at providing a coherent framework for parameter estimation in form of maximum likelihood theory. Also, properties like sufficiency, optimality or efficiency were all properties of estimators, and their purpose was to find good estimators. Thus, Fisher's work was primarily concerned with the second problem of statistics, the problem of estimation.

Fisher's first significance test already occured in a 1921 paper called *On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample* (Fisher, 1921b). Although the treatment was far from a full outline of a theory of hypothesis testing, it can be interpreted as the first use of significance tests by Fisher. In it, the last section deals with the example of a correlation coefficient derived from a sample of twins. The degree of correlation to be expected between both twins depends on multiple factors and is assumed to be $p = 0.18$. After that, Fisher calculates the correlation coefficient for a given sample of 39 twins, and notices:

> "The value found from 39 pairs of twins each measured in 6 traits was $-0.016 \pm$ [0.048] ..."
> Fisher (1921b, p. 23)

Fisher then calculated the distribution of the correlation coefficient under the hypothesis $p = 0.18$.[7] With reference to the observed difference of the sample correlation coefficient from the assumed value of $p = 0.18$, Fisher concluded after the calculations:

> "Its difference from the point
>
> $$p = 0.18 \tag{3.37}$$
>
> is now much more significantly apparent; for using the original estimate, this difference is only 4.1 times its probable error, for which $P = 0.0051$ (...). The evidence in favour of a single type of origin for this group of twins is thus stronger than I had previously imagined."
> Fisher (1921b, p. 23)

The first p-values were calculated as a byproduct of the whole paper, the calculations are not very detailed, and the procedure seems to be clear to Fisher, not worth any further explanations.[8] He did not even define the symbol P used to calculate p-values. Most of the ideas probably already took form in the early papers he published with Student. However, no exact p-values were calculated there. Fisher then distinguished his procedure from the Bayesian philosophy:

> "My treatment of the problem differs radically from that of Bayes. Bayes (1763) attempted to find, by observing a sample, the actual *probability* that the population value lay in any given range. In the present instance the complete solution of this problem would be to find the probability integral of the distribution of p. Such a problem is indeterminate without knowing the statistical mechanism under which different values of p come into existence."
> Fisher (1921b, p. 24)

Note that Fisher's primary objection to the solution of Bayes is the challenge to elicit the prior distribution for p. In the last paragraph of the paper, Fisher summarized that his approach makes it possible to test statistical hypotheses:

> "We may discuss the probability of occurrence of quantities which can be observed or deduced from observations, in relation to any hypotheses which may be suggested to explain these observations. We can know nothing of

[7] This was a solved problem at that time because of the earlier derivations in his 1915 paper *Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population* (Fisher, 1915).

[8] On page 214 of the same paper, a p-value is presented in a column of a table which includes various other estimates. Formally, this is the first p-value Fisher reported, but the first explanations for a p-value were given by him for the example on page 23 of the paper as cited above.

> the probability of hypotheses or hypothetical quantities. On the other hand we may ascertain the likelihood of hypotheses and hypothetical quantities by calculation from observations (...)."
> Fisher (1921b, p. 24)

The first phrasing of *significance testing* then appears in the 1922 paper *'The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients.'* (Fisher, 1922c). In it, Fisher derived the distribution of regression coefficients.[9] After conducting his derivations and finding the t-distribution with appropriate degrees of freedom as the distribution of the regression coefficients under the null hypothesis, he concluded:

> "Tables of the Probability Integral of the above Type VII distribution have been prepared by "Student" [8], for values of $n - p$ from 0 to 30. These tables are in a suitable form for testing the significance of an observed regression coefficient. For larger samples the curve will be sufficiently normal for most purposes (...)."
> Fisher (1922c, p. 610)

Also, Fisher already reasoned that these derivations also hold for the two-sample t-test. Together, both papers provide the first appearance of p-values and significance testing, which would become the main focus of Fisher's work later in his career. While in 1924 he already frequently used the term *tests of significance*, the use of p-values appeared already earlier, and first in the 1921 paper described above. Also, in his 1921 paper *'Studies on crop variation'* Fisher (1921a), he listed a table (Table II) with p-values and discussed the results. However, these first hints at hypothesis testing were still far from a concise, clearly articulated theory.

3.1.3 Fisher's refining of the statistical theory

Three years later, in 1925, Fisher published his paper *'Theory of statistical estimation'* (Fisher, 1925c). In this paper, Fisher refined multiple aspects of his 1922 paper and also introduced ancillary statistics (compare Definition C.60), one of the most controversial heritages of Fisher. At the beginning of the paper, Fisher stated probability as being a frequency ratio obtained from a hypothetical infinite population:

> "It has been pointed out to me that some of the statistical ideas employed in the following investigation never received a strictly logical definition (...). The idea of a frequency curve, for example, evidently implies an infinite hypothetical population distributed in a definitive manner; (...) The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability. (...) Also, the word infinite is to be taken in its proper mathematical sense as denoting the limiting conditions approached by increasing a finite number indefinitely."
> Fisher (1925c, p. 700)

Fisher's concept of probability later was one of the reasons of a bitter feud between Jerzy Neyman, Egon Pearson and Fisher himself about the nature of hypothesis testing. Furthermore, as can be seen from the above, Fisher's probability concept was far from

[9] As is well known, these are t_{n-p} distributed, where p and n are the number of parameters used in the regression and the sample size.

a modern measure-theoretic definition as Kolmogorov's axioms were not available at that time.

The structure of the 1925 paper *'Theory of statistical estimation'* can be delineated as follows: The first two sections were quite brief and dealt with the same ideas about estimation and consistency already introduced in the 1922 paper. Section three dealt mainly with efficiency and in section four, Fisher dealt with the ratio of an efficient and non-efficient statistic. Later in the paper, he introduced the method of scoring and gave a proof for the asymptotic efficiency of the MLE[10]. The last section is the most important as ancillarity is introduced (Fisher, 1925c).

Before the introduction of ancillarity in the last section, the introduction of the information number in section 6 of the 1925 paper was a milestone. Today, the information number is better known as the *Fisher-Information* (compare Definition C.46). Fisher discussed it under the name of "intrinsic accuracy of error curves" (Fisher, 1925c, p. 709).

> "The variance of efficient statistics from a distribution of any form affords us a measure of an important property of the distribution itself. (...) We may thus obtain a measure of the intrinsic accuracy of an error curve, and so compare together curves of entirely different form. If the variance of an efficient estimate derived from a large sample of n is A/n, then the intrinsic accuracy of the distribution is defined as $1/A$."
> Fisher (1925c, p. 709)

In modern terms, the intrinsic accuracy can be interpreted as the large sample precision of an estimator T, where the precision is equal to $n \cdot \mathbb{V}^{-1}(T) = \frac{1}{n\mathbb{V}(T)}$ for n large. Using his previously derived asymptotic variance of MLEs in Equation (3.36) it can be seen that the intrinsic accuracy equals

$$\frac{1}{n\mathbb{V}(T)} = -\mathbb{E}\left(\frac{\partial^2 \log y}{\partial \theta^2}\right) \tag{3.38}$$

Arriving at this point, Fisher gave then a formal definition of what he understood under the information of one observation:

> "What we have spoken of as intrinsic accuracy of an error curve may equally be conceived as the [expected] amount of information of a single observation belonging to such a distribution."
> Fisher (1925c, p. 709)

If the Fisher-Information is small this implies that the corresponding estimator variance is large, so a single observation yields only a small amount of information. If on the contrary, the Fisher-Information is large, the variance of the corresponding estimator needs to be small, and a single observation yields a greater amount of information.[11] Thus, when the Fisher information is large, the data provide substantial amount of information about the parameter to be estimated. By its introduction in the 1925 paper, Fisher did not know that his information number would later become the key quantity in the Cramér-Rao Lower Bound (Theorem C.44).

[10] For a modern proof see Rüschendorf (2014, p. 166-167).

[11] For details, see Schervish (1995, p. 111).

Section 7 of the 1925 paper (Fisher, 1925c, p. 710) then started with a proof of the asymptotic efficiency of the MLE before introducing ancillarity. The asymptotic efficiency of the MLE is a key property that can be regarded as a reason for the widespread appeal to maximum likelihood theory.

The last section of Fisher's 1925 paper then finally introduced ancillary statistics. Fisher first noted that there might exist no sufficient statistic in some cases:

> "(...) there exists no sufficient statistic, and some loss of information will necessarily ensue upon the substitution of a single estimate for the original data upon which it was based."
> Fisher (1925c, p. 718)

Fisher introduced the concept of ancillary statistics to solve this problem:

> "Since the original data cannot be replaced by a single statistic, without loss of accuracy, it is of interest to see what can be done by calculating, in addition to our estimate, an ancillary statistic which shall be available in combination with our estimate in future calculations.
> If our two statistics specify the values of $\partial L / \partial \theta$ and $\partial^2 L / \partial^2 \theta$ for some central value of θ, such as $\hat{\theta}$, then the variance of $\partial L / \partial \theta$ over the sets of samples for which both statistics are constant, will be that of
>
> $$\frac{1}{2}(\theta - \hat{\theta})^2 \frac{\partial^2 L}{\partial \theta^3} \tag{3.39}$$
>
> which will ordinarily be of order n^{-1} at least. With the aid of such an ancillary statistic, the loss of accuracy tends to zero for large samples."
> Fisher (1925c, p. 724)

Fisher's reasoning here was as follows: If there exists no sufficient statistic for θ at all, then definitely some information will be lost. Still, this loss can be quantified by using a Taylor series expansion. For any statistic T, such an expansion of $l'(\theta)$ around T yields

$$l'(\theta) = l'(T) + (\theta - T)l''(T) + \frac{1}{2}(\theta - T)^2 l'''(T) + ... \tag{3.40}$$

Fisher then assumed T to be the MLE of θ, denoted by $\hat{\theta}$ and inferred (due to $l'(\hat{\theta}) = 0$), that

$$\mathbb{V} = [l'(\theta) | l'(\hat{\theta}), l''(\hat{\theta})] = \mathbb{V}\left[\frac{1}{2}(\theta - \hat{\theta})^2 l'''(T)\right] \tag{3.41}$$

because the first and second terms of the right-hand side of Equation (3.40) are zero and constant. Equation (3.41) resembles the equation Fisher gave in the last quotation, and indeed, is of $O(1/n)$. This way, asymptotically, using the information provided by the ancillary statistic $l''(\hat{\theta})$ reduces the loss of information incurred by using an insufficient statistic to zero. Fisher added:

> "The function of the ancillary statistic is analogous to providing a true, in place of an approximate, weight for the value of the estimate."
> Fisher (1925c, p. 724)

In addition to the MLE $\hat{\theta}$, the ancillary statistic $l''(\hat{\theta})$ provided a measure of the curvature of the likelihood function, and the use of both MLE and the ancillary statistic $l''(\hat{\theta})$, therefore, result in a statistic with asymptotic sufficiency, even when the MLE itself is not sufficient.

3.1.4 Ancillarity

One of the problems with ancillarity from the first appearance in Fisher's 1925 paper was the lack of a precise definition. Even in 1962, Savage et al. (1962a, p. 19) argued that the "concept of ancillary statistic, introduced by Fisher, has been difficult to grasp and to define precisely". While the reasoning above may seem clear from today's perspective, it barely was back in 1925 and even years later. Another problem was that Fisher introduced ancillarity only for his maximum likelihood theory and not in a general way. This most probably led to confusion about how to use ancillary statistics separate from likelihood theory at all. Also, while in theory, the usage of $l''(\theta)$ may sound convincing, in multivariate problems, the calculation of the conditional distributions quickly becomes complicated. In 1934 and 1935, Fisher therefore tackled exactly these problems and tried to clarify his intentions to a wider audience (Fisher, 1934c, 1935). Only then, ten years later, ancillarity was accepted more widely. Fisher's 1934 paper, titled *'Two new properties of Mathematical Likelihood'*, explained how to recover the information lost when the location parameter θ of a Laplace distribution was estimated via the method of maximum likelihood. Fisher used as ancillary statistics the configuration of the sample composed of the i-th order statistics and concluded at the end of the paper:

> "The process of taking account of the distribution of our estimate in samples of the particular configuration observed has therefore recovered the whole of the information (...)."
> Fisher (1934c, p. 303)

In his 1935 paper, *'The logic of inductive inference'* (Fisher, 1935), Fisher discussed ancillarity again:

> "It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the parameter, but, instead tell us how good an estimate we have made of it."
> Fisher (1935, p. 48)

This closely resembles the definition still used today, compare Definition C.60. A statistic T is ancillary for a parameter θ of interest if its distribution does not depend on θ. Fisher used the 2 × 2 table to illustrate ancillarity:

> "The use of ancillary statistics may be illustrated in the well-worn topic of the 2 × 2-table. Let us consider a classification as Lange supplies in his study on criminal twins. Out of 13 cases judged to be monozygotic, the twin brother of a known criminal is in 10 cases also a criminal; and in the remaining 3 cases he has not been convicted. Supposing the data to be accurate, homogenous, and unselected, we need to know with what frequency so large a disproportion would have arisen if the causes leading to conviction had been the same in the two classes of twins. We have to judge this from the 2 × 2 table of frequencies."
> Fisher (1935, p. 48)

Fisher then gave the following 2 × 2-table (see Table 3.1): Fisher's idea consisted of taking the information in the margins of the table, which themselves supply no information about the ratios of the twins inside the table cells. He proceeded:

	Convicted	Not Convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Table 3.1: From (Fisher, 1935), page 48

> "If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary;"
> Fisher (1935, p. 48)

After that, Fisher noted that there were in total 13 possible combinations of margins for the 2 × 2 tables, identified by iterating from 0 to 12 in the dizygotic convict number in the lower-left cell of the table body. He then used the binomial distribution to model that $(x+1)$ are not convicted and $(12-x)$ are convicted out of the 13 monozygotic twins[12]. He assumed $p_1 = p_2 =: p$, where these are the probabilities for a conviction for both groups. In modern terms, this matches the assumption of a null hypothesis (compare Definition C.65, where in Fisher's calculation, $\Theta := \{(p_1, p_2) : p_1, p_2 \in (0,1]\}$ and $\Theta_0 := \{p_1 = p_2\}$). So he first wrote

$$\frac{13!}{(12-x)!(x+1)!} p^{12-x}(1-p)^{x+1} \tag{3.42}$$

for the probability that $12-x$ are convicted while $x+1$ are not convicted out of the 13 monozygotic twins as well as

$$\frac{17!}{x!(17-x)!} p^{x}(1-p)^{17-x} \tag{3.43}$$

for the probability that $17-x$ are not convicted while x are convicted out of the 17 dizygotic twins (Fisher, 1935, p. 49). The probability of both events was written by Fisher as the product

$$\frac{13!17!}{(12-x)!(1+x)!x!(17-x)!} p^{12}(1-p)^{18} \tag{3.44}$$

Fisher (1935, p. 49) argued that this quantity is proportional (as a function of x) to

$$\frac{1}{(12-x)!(1+x)!x!(17-x)!} \tag{3.45}$$

and concluded:

> "...and on summing the series obtained by varying x, the absolute probabilities are found to be
>
> $$\frac{13!17!12!18!}{30!} \frac{1}{(12-x)!(1+x)!x!(17-x)!} \tag{3.46}$$
>
> "
>
> Fisher (1935, p. 49)

[12] Note that p_1 is assumed to be greater than 0 and therefore at least one twin needs to be not convicted which is the case for $x = 0$.

which matches the hypergeometric distribution for $x = 0, 1, ...$ The absolute probability can thus be written as:

$$\frac{\binom{13}{12-x}\binom{17}{x}}{\sum_{x=0}^{12}\binom{13}{12-x}\binom{17}{x}} = \frac{\binom{13}{12-x}\binom{17}{x}}{\binom{30}{12}}, x = 0, 1, 2, ... \tag{3.47}$$

Fisher then noted:

> "...the significance of the observed departure from proportionality is therefore exactly tested by observing a discrepancy from proportionality as great or greater than that observed, will arise, subject to the conditions specified by the ancillary information."
> Fisher (1935, p. 50)

Fisher then concluded, that as there are 10 monozygotic twins in the table who were convicted, the probability of a result as large as 10 or larger, is

$$\frac{\binom{13}{10}\binom{17}{2} + \binom{13}{11}\binom{17}{1} + \binom{13}{12}\binom{17}{0}}{\binom{30}{12}\binom{17}{0}} = 0.000465 \tag{3.48}$$

which shows that the null hypothesis $p_1 = p_2$ would be rejected at the 5% significance level in modern notation. In contrast to his earlier uses of p-values, this example shows how settled the concept had become a decade later. The derivations are much clearer and better commented as in the first calculations of p-values by Fisher. The above example would later be named a one-sided p-value (compare Definition C.83). The ancillary statistics of the table margin counts make it possible that the conditional distribution of x (conditioned on the table margins) is independent of the parameter of interest, p. While Fisher in 1925 barely did anticipate the paramount importance of such quantities for the later development and use of hypothesis tests[13], he essentially calculated the first *pivot* (under the null hypothesis) here, compare Definition C.88. The above test over time became popular as the *exact Fisher test*, published first in 1934 in the fifth edition of *'Statistical Methods for Research Workers'* (Fisher, 1934b, p. 99), and is still widely used in medical research today (Held and Sabanés Bové, 2014). So in summary, ancillarity directly led to the first precise formulation of p-values. Furthermore, the exact Fisher test and the concept of pivots was outlined for the first time in connection with ancillarity.[14]

3.2 Fisher's Significance Testing Framework

The above sections detailed how Fisher's estimation theory was outlined, and how the central concepts like sufficiency, consistency, maximum likelihood and ancillary statistics were invented. The shift from deriving distributions under specific assumptions to significance testing and the calculation of p-values was described in the last section

[13] The principal idea of the p-value seeped into likelihood ratio tests, score tests and Wald tests, compare Held and Sabanés Bové (2014). Also, the p-value often is used in connection with Neyman-Pearson tests, a situation which created the hybrid inconsistent approach of hypothesis testing described in Chapter 5.

[14] Note that the exact Fisher test – although a procedure following conditional inference as outlined below – has drawbacks: It rejects a true null hypothesis too often because of the hypergeometric distribution's discreteness. Therefore, in practice it has a smaller size than 0.05, even if $\alpha = 0.05$, compare Definition C.71.

and seems natural, as these calculations were only possible because of the previous derivations of the distributions of specific statistics like those of regression coefficients or the correlation coefficient. The previous work allowed Fisher to transition from his earlier allusions to hypothesis testing as in his 1921 paper or the ancillary statistics twin example to a thorough treatment of hypothesis tests. As already mentioned, Fisher's work can be separated into two quite distinct parts. In the first part of his scientific career, he was mainly concerned with problems of estimation and deriving a theory to find, evaluate and compare estimators as shown in the previous sections. This first period lasted between the beginning of Fisher's career in 1912 until the mid of the 1920's when the first edition of *'Statistical Methods for Research Workers'* was published. In these years, Fisher built successively upon his maximum-likelihood-theory, from its first introduction, over the addition of more concepts like sufficiency in his 1922 paper (Fisher, 1922a) until the refining of his theory in his 1925 paper *'Theory of Statistical Estimation'* (Fisher, 1925c). It is not easy to draw a clear line between both periods of Fisher's work, because, after the 1922 paper on the mathematical foundations of theoretical statistics, Fisher's maximum likelihood method was published in full account. Hypothesis testing started with the 1921 and 1922 papers described above. Therefore the transition is continuous.[15] Additionally, Fisher did not formally introduce his theory of significance tests: For him, the logic behind these hypothesis tests seemed clear probably because of the years of earlier work together with Gosset, and as a consequence, the logic of his significance test was motivated in large parts on intuitive grounds.

3.2.1 Fisher's Shift from Estimation Theory to Significance Testing: *'Statistical Methods for Research Workers'*

From 1922 on, Fisher's primary focus shifted gradually from improving his estimation theory and deriving exact distributions towards the theory of significance tests. Three years later, when *'Statistical Methods for Research Workers'* (SMRWI) was published, this paradigm is clearly articulated in the introduction, where Fisher stated the scope of the book:

> "...the prime object of this book is to put into the hands of research workers... the means of applying statistical tests accurately to numerical data accumulated in their own laboratories."
> Fisher (1925a, p. 17)

Fisher's book, therefore, was targeted directly at researchers and advocated the use of his significance tests. He also shaped the way scientific work was conducted by presenting handy tables of the distributions necessary to conduct the significance tests:

> "The tables of distributions supplied at the ends of several chapters form a part essential to the use of the book."
> Fisher (1925a, Introduction, Section 5)

[15] It should be noted, that the first part of Fisher's work marked a transition in statistical science. The older inverse probability approach (which equals a modern Bayesian approach) was demised and rejected more and more because of Fisher's work: "the framework for likelihood, if not the term itself, was present in 1912, and thus that the break with inverse probability, if not clean, was at least clear." (Howie, 2002, p. 68).

In most cases, Fisher assumed normality of the sample, and therefore the choice and derivation of the appropriate test statistic were obvious for him. In contrast, when working on estimation theory, the choice of a good estimator was not particularly clear and differed from case to case. Thus, Fisher's framework in *'Statistical Methods for Research Workers'* consisted of choosing the right table which listed tabulated values of the distribution of the appropriate test statistic and declaring significance of the results for sample values larger than a specific threshold. Statistically, the only difficult problem was to derive the correct distributions – which of course had already been obtained in earlier years for most standard statistical models and were presented in the book. Once this step was done and the values were tabulated, everything else could be done in a nearly algorithmic fashion by plugging in the observed experimental data.

The rest of the first part of *'Statistical Methods for Research Workers'* treated diagrams and distributions, the χ^2-test of goodness of fit, homogeneity and independence.

The second part of SMRWI – starting with Chapter 5 –introduced exact small sample tests then, specific tests for differences of two means and tests for regression coefficients via the use of the t-distribution. In the chapter, it was shown that the t-test, first introduced for the differences of two means, also applies for the testing of regression coefficients in linear regression settings. Regarding the level of content, it is worthwhile to remember that these concepts had been developed by Fisher a few years ago, as described above (Fisher, 1922c). The content of the book was therefore not only targeted at practitioners but also represented state-of-the-art statistical theory of the time.

Chapter 6 then focussed on correlation coefficients, giving a treatment of the correlation coefficient p in a bivariate normal distribution including its estimation and judging the significance of an observed correlation coefficient r. Chapter 7 extended the previous ideas to intraclass correlations and tackling tests arising in the analysis of variance, providing the general ideas of this procedure as well as examples and illustrations for the use of the appropriate table in the appendix. The last Chapter 8 then showed applications of the analysis of variance, mainly resulting out of Fisher's work at Rothamsted: In one case the analysis is carried out for agricultural plots assigned to different treatments randomly, versus randomized blocks, versus Latin squares. SMRWI also can be seen as the place where Fisher first advocated the (completely arbitrary) threshold for his significance tests to reject a hypothesis:

> "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy."

This recommendation should become close to a natural law in social and medical sciences over time, as already detailed in Chapter 1. Thus, the guidelines given by Fisher clearly attributed to the problems which are observed in the reproducibility crisis witnessed today: They can be seen as the mental attitude or established research habits which were listed as one of the major problems in the replication crisis by Ioannidis (2005b).

3.2.2 The impact and reception of SMRW

The impact of SMRW was enormous, measured objectively. In eight chapters, Fisher managed to construct a solid statistical foundation for researchers interested in statistical analysis of their data. McGrayne (2011, p. 47) entitled SMRW a "cookbook of ingenious statistical procedures for nonstatisticians", which "turned frequency into the de facto statistical method". Also, according to (McGrayne, 2011, p. 47), "no one today can discuss statistics – what he called "mathematics applied to observational data" – without using some of Fisher's vocabulary". Still, Fisher's writing was complex, and as proofs were intentionally not included, the ideas were even harder to grasp for readers interested in why the analyses and tests worked. Lehmann (2011) analysed some official reviews published in statistical journals in the period after the publication of SMRWI, which all show that most readers found it hard to understand and criticised the lack of proofs. Nevertheless, the first edition was sold out in a short period, and in 1928 Fisher published the second edition, also including no proofs but adding a chapter titled *'The Principles of Statistical Estimation'*, in which he tried to explain the earlier developed concepts of sufficiency and consistency. Lehmann (2011) noted that while the reviews for the second edition were slightly more favourable than for the first, most people still had their issues with the style of presentation of the topics. This remaining criticism can be attributed to the demanding level of content which was included in the book.

Some even took Fisher's maximum likelihood method as "nothing more than an application of inverse probability with uniform priors" (Howie, 2002, p. 75). The only favourable review was that of Student, at that time a good friend of Fisher, and as Howie (2002) notes, "the initial reviews, from the biometricians who still dominated the statistical community, were uniformly negative." (Howie, 2002, p. 74) Still, this did not inhibit the interest of applied researchers, especially "social scientists, who saw Fisher's book as a route to objectivity and thus legitimacy." (Howie, 2002, p. 74)

Regarding the impact, SMRWI achieved to tie together two distinct requirements of statistics at that time. The small sample tests of Student, as well as the χ^2-test of Pearson, had been part of Fisher's work in the years before publication. They were presented as modern methods for practical data analysis in SMRWI, although the underlying theory remained mysterious for the reader. Next to giving a compendium of available tests, it also presented these in realistic situations, a rarity, which attracted lots of non-mathematicians to the book. The first edition of 1050 copies was sold out after three years, and Fisher published the second edition two years later. Most of the other editions also followed in two-year cycles. The size of the editions steadily increased with the popularity of the book, reaching its maximum of 7500 copies of the eleventh edition published in 1950. The last edition, the fourteenth, was published in 1970 posthumously and incorporated changes based on notes Fisher had made before his death in 1962 for the next edition. Retrospective, the book was praised by lots of institutions for its impact as noted by Lehmann (2011, p. 25/26).

The most problematic issue with SMRW was Fisher's writing style. It was simply too complicated for mass appeal (Howie, 2002, p. 76) and Fisher's daughter Joan Fisher Box wrote in the biography of Fisher, that

> "It was George W. Snedecor ... who was to act as a midwife in delivering the new statistics in the United States."
> Box (1978, p. 313)

Snedecor (1937) published a book called *'Statistical Methods'* in 1937 which covered roughly the same content as Fisher's SMRWI. The χ^2-test, two-sample and one-sample tests, regression, correlation coefficients, as well as the analysis of variance, were included in it. In contrary to Fisher's book, it was presented in a much easier way, allowing a broader audience to consume the content and understand it. Snedecors *'Statistical Methods'* was sold over 200000 times in multiple editions and can be seen as the accessible translation of Fisher's SMRW for the masses. Thus, Fisher's ideas were populated strongly in by Snedecor's textbook.

3.2.3 The conventional frequency theory of probability and Fisher's Conditional Inference

While Fisher from the beginning resented the Bayesian philosophy due to its subjectiveness of priors, he also had his problems with the traditional frequency interpretation of probability. Fisher's interpretation of probability was also crucial in the later debate about the correct hypothesis testing framework to be used by scientists between him and Jerzy Neyman and Egon Pearson. Much later, Fisher (1956b) criticised the conventional theory of frequency probability. One of Fisher's legacies remains in the concept of so-called conditional inference, which involved the concept of *relevant subsets*:

> "(...) information supplied by a mathematical statement such as: "If a aces are thrown in n trials, the probability that the difference in absolute value between $1/6$ and a/n shall exceed any positive value ε, however small, shall tend to zero as the number n is increased indefinitely", will seem not merely remote, but also incomplete and lacking in definiteness in its application to the particular throw in which he is interested. Indeed, by itself it says nothing about that throw. It is obvious, moreover, that many subsets of future throws, which may include his own, can be shown to give probabilities, in this sense, either greater or less than $1/6$. Before the limiting ratio of the whole set can be accepted as applicable to a particular throw, a second condition must be satisfied, namely that before the die is cast no such subset can be recognized. This is a necessary and sufficient condition for the applicability of the limiting ratio of the entire aggregate of possible future throws as the probability of anyone particular throw. On this condition we may think of a particular throw, or of a succession of throws, as a random sample from the aggregate, which is in this sense subjectively homogeneous and without recognizable stratification."
> Fisher (1956b, p. 32-33)

What Fisher (1956b) tried to express in his much later published book *'Statistical Methods and Scientific Induction'*, was that a concept of probability like the classical frequency-based concept has to possess two properties. First, the relative frequencies need to converge to a limiting value like in the dice-example given by him. Second, subsequences of a given sequence need to converge to the same value. Additionally, according to Fisher, a given sequence must possess no *relevant subset*. A general subset can be interpreted as a subset of the sample space. Often, this relevant subset is provided by ancillary statistics. One of the most prominent examples showing what conditional inference in Fisher's sense meant is not given by Fisher (1956b), but by Cox (1958), which is today known as one of the classic examples in favour of Fisher's conditional inference. Cox (1958) wrote in his paper *'Some problems connected with statistical inference'* about the following situation:

> "Suppose that we are interested in the mean θ of a normal population and that, by an objective randomization device, we draw either (i) with probability 1/2, one observation, x, from a normal population of mean θ and variance σ_1^2, or (ii) with probability 1/2, one observation x, from a normal population of mean θ and variance σ_2^2, where σ_1^2, σ_2^2 are known, $\sigma_1^2 >> \sigma_2^2$, and where we know in any particular instance which population has been sampled.
> The sample space formed by indefinite repetition of the experiment is clearly defined and consists of two real lines $\sum_1, \sum_2$, each having probability 1/2, and conditionally on $\sum_i$, there is a normal distribution of mean θ and variance σ_i^2.
> Now suppose that we ask, accepting for the moment the conventional formulation, for a test of the null hypothesis $\theta = 0$, with size say 0.05, and with maximum power against the alternative $\theta' \approx \sigma_1 >> \sigma_2$."
> Cox (1958, p. 360)

Cox proceeded by investigating two tests for the given hypothesis $H_0 : \theta = 0$ against $H_1 : \theta > 0$. In the first case, Cox conditionalized on the population samples, and in the second case, he did not. He supposed a fair coin had been tossed to choose the population to sample from, and the outcome of the tossed coin was used as the ancillary statistic. This way, Cox (1958) obtained the rejection region $X > c$ of the conditionalized test as

$$\mathbb{P}\{X > c|\sum_i\} = 0.05 \tag{3.49}$$

$$\Leftrightarrow 1 - \Phi\left(\frac{c}{\sigma_i}\right) = 0.05 \tag{3.50}$$

which yields $c = \sigma_i \Phi^{-1}(.95) = 1.645\sigma_i$ for c and where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution. The rejection region (see Definition C.69) of the conditionalized test can therefore be written as $1.645\sigma_1$, if the first population was chosen by the coin, and $1.645\sigma_2$, if not.
Cox (1958) then proceeded by investigating the unconditionalized test. The level 0.05 was also chosen, and the rejection region $X > c$ therefore

$$\mathbb{P}\{X > c\} = 0.05 \tag{3.51}$$

$$\Leftrightarrow \mathbb{P}\{X > c|\textstyle\sum_1\}\mathbb{P}\{\sum_1\} + \mathbb{P}\{X > c|\sum_2\}\mathbb{P}\{\sum_2\} = 0.05 \tag{3.52}$$

$$\Leftrightarrow \frac{1}{2}\left[\mathbb{P}\{X > c|\textstyle\sum_1\} + \mathbb{P}\{X > c|\sum_2\}\right] = 0.05 \tag{3.53}$$

$$\Leftrightarrow \frac{1}{2}\left[\mathbb{P}\{Z > \frac{c}{\sigma_1}\} + \mathbb{P}\{Z > \frac{c}{\sigma_2}\}\right] = 0.05 \tag{3.54}$$

$$\Leftrightarrow \frac{1}{2}\left[1 - \Phi(\frac{c}{\sigma_1}) + 1 - \Phi(\frac{c}{\sigma_2})\right] = 0.05 \tag{3.55}$$

$$\Leftrightarrow \left[\Phi(\frac{c}{\sigma_1}) + \Phi(\frac{c}{\sigma_2})\right] = 1.9 \tag{3.56}$$

where $Z \sim \mathcal{N}(0,1)$. For fixed σ_1 and σ_2, it is possible to obtain solutions for c then. Cox (1958) chose as an example $\sigma_1 = 100, \sigma_2 = 1$, so that $c \approx 128.2$. Cox (1958) then noted, that the problem occurring with the unconditionalized test is, that the average α level of 0.05 is a mixture of

$$\alpha_1 = \mathbb{P}\{Z > \frac{c}{\sigma_1}\} = 1 - \Phi(\frac{128.2}{100}) = 1 - \Phi(1.282) = 0.100 \tag{3.57}$$

$$\alpha_2 = \mathbb{P}\{Z > \frac{c}{\sigma_2}\} = 1 - \Phi(\frac{128.2}{1}) = 1 - \Phi(128.2) \approx 0.000 \tag{3.58}$$

The quintessence therefore is, that the unconditionalized test maintains its long run error rate of 0.05 by averaging both these two error rates α_1 and α_2. For any *particular* test at hand, this average error rate - or in modern notation, the test level, compare Definition C.71 - is simply not attained. If the sample indeed is chosen from the first population, then the true type I error rate is 0.100. This is too high and therefore unacceptable. If on the other hand the sample is from the second population, the true Type I error rate is exactly 0.000, so no error occurs at all. Cox (1958) therefore added:

> "(...) if the object (...) is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say "what we can learn from the data that we have", the unconditional test is surely no good."
> (Cox, 1958, p. 360)

The dilemma of unconditional inference, therefore, is, that for practical purposes, it is not helpful according to Cox (1958). Returning to Fisher, this was also his argument against unconditional inference (Fisher, 1956b, p. 32-33), and the relevant subset in the above example to obtain a correct probability statement in terms of statistical inference corresponds to the test actually performed. Fisher's relevant subsets thus were a vague formulation of what Cox called conditional inference, and according to Fisher statistical inference needed to be performed conditional on the relevant subset, that is, conditional on the observed data and performed experiment or study at hand. Later, this became a substantial argument for Fisher against the Neyman-Pearson approach of hypothesis testing, detailed in the next chapter.[16]

The derivations of Cox (1958) made a strong argument for conditional inference in the Fisherian sense. Indeed, due to the example given by Cox (1958), the principle arose that when one of two distinct experiments is chosen randomly and performed in succession, the inference about the parameter θ of interest should be made conditional only on the chosen experiment. There is no official publication to which this principle can be rooted back, but most likely the paper of Cox (1958) can be seen as the cornerstone of the conditionality principle, which will be discussed in part IV in detail. In Cox' example, it is directly observable what happens if one ignores the conditionality principle, that is if an unconditional test is performed. In this case, when for example $x = 1.9$ is observed and $\sigma_1 = 100, \sigma_2 = 1$, under the assumption that the second

[16] After Fisher's criticism in 1956 Buehler (1959) formalized these ideas with so-called positively and negatively biased relevant subsets. These subsets are confidence sets which are called positively biased if they attain a coverage probability of $\geq 1 - \alpha$, and negatively biased if they attain a coverage probability $\leq 1 - \alpha$. Applications of these biased relevant subsets were given by Fisher (1956a) in the discussion for the Behrens-Fisher problem and also by Buehler and Feddersen (1963), who showed that for Student's t-statistic there do indeed exist positively biased subsets, doubting the validity of unconditionalized t-tests.

population was sampled, the conditional one-sided p-value can be calculated as

$$\mathbb{P}(X \geq 1.9 | X \sim \mathcal{N}(0, 1^2)) = 0.02872 \tag{3.59}$$

while the unconditional p-value is calculated as

$$\tfrac{1}{2}\mathbb{P}(X \geq 1.9 | X \sim \mathcal{N}(0, 1^2)) + \tfrac{1}{2}\mathbb{P}(X \geq 1.9 | X \sim \mathcal{N}(0, 100^2)) = \tfrac{1}{2}(0.028 + 0.492) = 0.2605 \tag{3.60}$$

So while the conditional test leads to rejection of the null hypothesis $H_0 : \theta = 0$, the unconditional frequentist test does not. The conditionality principle therefore can be traced back to Fisher (1956b) and Cox (1958), and unconditional hypothesis testing or unconditional statistical inference violate the conditionality principle. The sufficiency principle and the conditionality principle together provide the likelihood principle (see Part IV), which also can be traced back to Fisher. In his 1922 paper, Fisher (1922b) still believed that sufficiency was always provided by maximum likelihood solutions:

> "Such a method is, I believe, provided by the Method of Maximum Likelihood."
> Fisher (1922b, p. 323)

Three years later, in his 1925 paper Fisher (1925c) added:

> "When sufficient statistics exist, it has been shown that they will be solutions of the equations of maximum likelihood."
> Fisher (1925c, p. 714)

A proof of this fact is given by Rüschendorf (2014, Proposition 5.4.18). Another nine years later, Fisher (1934c) concluded that likelihood,

> "(...) when properly interpreted must contain the whole of the information respecting x which our sample of observations has to give."
> Fisher (1934c, p. 297)

This means that all inference should be done with respect to the likelihood function.[17] While Fisher's idea was conclusive at the time of publication, there was no rigorous proof. The proof followed nearly three decades later and can be attributed to Birnbaum (1962), who showed in his landmark paper *'On the Foundations of Statistical Inference'* that the likelihood principle follows from the sufficiency principle and the conditionality principle (compare Theorem 11.8 and Chapter 11 in Part IV). However, Fisher did not restrict himself strictly to following the likelihood principle. In his calculations of p-values, in the example of the 2x2 table given in Fisher (1935), he based his inference also on observations not made at all. He calculated the probabilities of 11 or more monozygotic twins in Equation (3.48), which violates the conditionality principle.

In summary, Fisher's theory of significance testing was present in full account after his 1925 publication of SMRWI, and the analysis revealed that his significance tests were intended to be used conditional on an ancillary statistic, which was also reflected in Fisher's concept of probability.

[17] This should not be confused with the statement that the maximum likelihood estimate is sufficient. In general, this is not the case.

Chapter 4

The Neyman-Pearson Theory of Hypothesis Testing

> WITHOUT HOPING TO KNOW WHETHER EACH SEPARATE HYPOTHESIS IS TRUE OR FALSE, WE MAY SEARCH FOR RULES TO GOVERN OUR BEHAVIOUR WITH REGARD TO THEM, IN FOLLOWING WHICH WE INSURE THAT, IN THE LONG RUN OF EXPERIENCE, WE SHALL NOT BE TOO OFTEN WRONG.
>
> Jerzy Neyman and Egon Pearson
> *On the problem of the most efficient tests of statistical hypotheses*

The Neyman-Pearson collaboration started in the mid-1920s as described in Chapter 2. While Egon Pearson was by then a scholar of the traditional school of statistics, which was founded by his father, Karl Pearson, he recognised the potential of the new developments achieved by Fisher's approach. Also, he was inspired by William Sealy Gosset as described in Chapter 2 to pursue his ideas on hypothesis testing further and therefore started the collaboration with Jerzy Neyman. The development of Neyman and Pearson's work is well documented. The main sources are Pearson (1966) himself as well as Constance Reid's biography of Jerzy Neyman (Reid, 1982). Also, the correspondence between Neyman and Pearson in the form of multiple letters has been archived at the Bancroft Library at the University of California in Berkeley[1].

4.1 The Beginning of a new Theory for statistical Hypothesis testing

The Neyman-Pearson collaboration started with Pearson's correspondence with Gosset, who pointed him in the right direction. Pearson (1966) recalls, that after receiving Gosset's letter, which was dated 11th May 1926

[1] The letters cited here are in collection BANC MSS 2008/250, Box 1, Folders number $1-9$; For simplicity, they are not cited separately here. Instead, the dates of the letters are given as identifiers in each case. See also: https://oac.cdlib.org/search?query=constance%20reid;idT=UCb162807442

> "...a number of new ideas must have begun to take shape. The possibility of getting a mathematical entry into the problem by specifying a class of alternative hypotheses which should be accepted as "admissible" for formal treatment; (...) the "rejection region" in the sample space; the "two sources of error." These were points which we must have discussed during autumn of 1926."
> Pearson (1966)

Pearson (1966) further stressed, that the idea of "determining the choice among possible contours in the sample space" to compare the hypothesis tested with the alternatives from first rough notes finally led to the idea of using the likelihood ratio criterion to determine these contours like it is common practice today, see Definition C.69 and C.75. In November 1926, Egon Pearson then wrote down some of his ideas and sent them to Jerzy Neyman. November 1926, therefore, marked the beginning of the Neyman-Pearson collaboration, which was to last for years. In the beginning, Jerzy Neyman somehow seemed to lack understanding for what Pearson wanted to work on. In a letter from the 9th December 1926 he stated:

> "(...) it seems to me that this principle is equivalent to the principle leading to inverse probabilities."

Moreover, concerning the likelihood ratio idea proposed by Pearson, he added:

> "What you have done can be expressed in words: wishing to test the probability of a hypothesis A we have to assume that all hypotheses are a priori equally probable and to calculate the probability a posteriori of A."

which not only shows a Bayesian influence in Neyman's thoughts but also underlines how fragile the concepts must have been at that time. While Pearson must have roughly sketched the biggest part of the key ideas like null hypothesis and alternative(s), rejection regions and the likelihood ratio statistic then, Neyman seemed to misunderstand them or at least considered solving the problem in a Bayesian manner. Pearson (1966) also mentioned that

> "...in our first joint paper (1928) we agreed to keep the door open by tackling problems in a variety of ways, one of which was based on an inverse probability approach."
> Pearson (1966)

All these ideas were more clearly expressed two years later in their first joint paper in *Biometrika*.

4.2 The Criterion of Likelihood

The first publication was named *'On the use and interpretation of certain test criteria'* (Neyman and Pearson, 1928) and published in 1928. It is divided into two parts, of which the first is much more important concerning the novelty of the proposed hypothesis testing approach. The paper starts by introducing the reader to the central problem of statistical inference:

> "One of the most common as well as most important problems which arise in the interpretation of statistical results, is that of deciding whether or not a particular sample may be judged as likely to have been randomly drawn from a certain population, whose form may be either completely or only partially specified. We may term Hypothesis A the hypothesis that the population from which the sample $\sum$ has been randomly drawn is that specified, namely $\prod$. In general, the method of procedure is to apply certain tests or criteria, the results of which will enable the investigator to decide with a greater or less degree of confidence whether to accept or reject Hypothesis A, or, as is often the case, will show him that further data are required before a decision can be reached."
> Neyman and Pearson (1928, p. 175)

After mentioning the classical and the inverse probability approach, Neyman and Pearson (1928) point out the goal of their paper:

> "What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitations, should be clearly understood, and it is because we believe this often not to be the case that it has seemed worth while to us to discuss the principles involved in some detail and to illustrate their application to certain important sampling tests."
> (Neyman and Pearson, 1928, p. 176)

Neyman and Pearson (1928) next consider two distinct approaches, which can be termed likelihood-based and Bayesian. In their words, they separate between

> "two distinct methods of approach, one to start from the population $\prod$, and to ask what is the probability that a sample such as $\sum$ should have been drawn from it, and the other the inverse method of starting from $\sum$ and seeking the probability that $\prod$ is the population sampled. The first is the more customary method of approach, partly because it seems natural to take $\prod$ as the point of departure since in practice there are often strong *à priori* grounds for believing that this is the population sampled, and partly because there is a common tendency to view with suspicion any method involving the use of inverse probability. But in fact, however strong may be the à priori evidence in favour of $\prod$, there would be no problem at all to answer if we were not prepared to consider the possibility of alternative hypotheses as to the population sampled; and we shall find that it is impossible to follow the first method very far without introducing certain ideas of inverse probability–that is to say, arguing from the sample to the population. If on the other hand we start boldly with assumptions regarding *à priori* and *à posteriori* probability, we reach by an almost simpler method sampling tests very nearly equivalent to those obtained from the first starting-point. Indeed the inverse method may be considered by some the more logical of the two; we shall consider first however the other solution."
> Neyman and Pearson (1928, p. 176)

Importantly, although the Neyman-Pearson theory of hypothesis testing later established itself as a widely used method and itself is non-Bayesian, the last sentence stresses a clear, logical preference for the Bayesian viewpoint. Still, because of Fisher's writings, the Bayesian theory of inverse probability was declared as unscientific, which probably also influenced Neyman and Pearson in favouring the likelihood approach, as both of them wanted to make a career as a statistician in academia. However, as Howie (2002) notes, "Despite Fisher's attacks, then, inverse probability was weak, but still viable, in 1930." (Howie, 2002, p. 80). A first publication ending in a hostile debate with Fisher about the appropriateness of Bayesian inference would not have helped with their goals.

After that, Neyman and Pearson explained that the sample $\sum$ is represented as a point in hyperspace and the acceptance or rejection of the hypothesis depends on a system of contours in it. These contours need to be chosen in such a way, that moving out from contour to contour the hypothesis A "becomes less and less probable." (Neyman and Pearson, 1928, p. 176), which resembles a rejection region. In a footnote on the same page, the separation between likelihood or confidence and probability seems to be clear to them, as they note:

> "the term "probability" used in connection with hypothesis A must be taken in a very wide sense. It cannot necessarily be described by a single numerical measure of inverse probability; as the hypothesis becomes less "probable", our confidence in it decreases, and the reason for this lies in the meaning of the particular contour system that has been chosen."
> (Neyman and Pearson, 1928, p. 176)

From this footnote, it seems that Neyman and Pearson (1928) differentiate between the probability of a hypothesis and the confidence in it. After further elaborations, they introduce the well known two types of error, as given in Definition C.73 and Definition C.74:

> (1) Sometimes, when hypothesis A is rejected, $\sum$ will, in fact, have been drawn from $\prod$.
>
> (2) More often, in accepting hypothesis A, $\sum$ will really have been drawn from $\prod'$.
>
> In the long run of statistical experience the frequency of the first source of error (or in a single instance its probability) can be controlled by choosing as a discriminating contour, one outside which the frequency of occurrence of samples from $\prod$ is very small-say, 5 in 100 or 5 in 1000.
> Neyman and Pearson (1928, p. 177)

where $\prod'$ is some alternative population, which has to follow a different distribution in turn. This analysis shows that next to Fisher, also Neyman and Pearson (1928) proposed a standard threshold for hypothesis testing. It is, however, interesting, that the two proposals made by Neyman and Pearson (1928) differ by a factor of ten.

After explaining the general ideas, Neyman and Pearson then turn to apply these to the problem of testing the mean of a normal distribution. Two settings are considered: In the first situation, by the null hypothesis, the mean and standard deviation are both known and by the alternative hypothesis both are unknown. In the second case, the null hypothesis specifies the mean as known and the standard deviation as unknown against the alternative hypothesis that both mean and standard deviation is unknown. What comes next, is the introduction of their 'criterion of likelihood' λ (Neyman and Pearson, 1928, p. 187), in exact terms

$$\lambda = \frac{\text{Likelihood of } \prod}{\text{Likelihood of } \prod' \text{ (max.)}} \tag{4.1}$$

which can be expressed in more modern terms as

$$\lambda = \frac{L(\prod \mid \sum)}{\max_{\prod'} L(\prod' \mid \sum)} \hat{=} \frac{\sup\limits_{\Theta_0} L(\theta|x)}{\sup\limits_{\Theta} L(\theta|x)} \tag{4.2}$$

compare Equation (C.22), where again $\prod'$ denotes any alternative hypothesis. The idea is intuitive: If the maximum likelihood is much larger under the alternative hypothesis $\prod'$, then λ will become small, indicating that is plausible to reject the null hypothesis $\prod$. In modern terms, the λ criterion is called the likelihood ratio as given in Definition C.75. After that, they derived the λ-test to test hypothesis A, and concluded:

> "Without claiming that this method is necessarily the "best" to adopt, we suggest that the use of this contour system (...) provides (...) one clearly defined method of discriminating between samples for which hypothesis A is more probable and those for which it is less probable. It is a method which takes into account the likelihood of alternative hypotheses..."
> Neyman and Pearson (1928, p. 188)

Problematically, in the above they conflate a Bayesian posterior probability with the likelihood of data given a hypothesis. When discriminating samples for which hypothesis A is more probable, a posterior distribution of A given the data is required. Their criterion, however, only provides the ratio of marginal likelihoods of the data under A and the possible alternative hypotheses, and thus does not make any statement in probability about the hypotheses under consideration.

After finishing the testing problem of hypothesis A, Neyman and Pearson then turned to the second situation. They noted immediately, that the null hypothesis is indeed a family of hypotheses, as the mean is known but the standard deviation σ has multiple possible values. Although they mentioned before that they would treat the situation via likelihood, they ended up noting:

> "But B is really a multiple hypothesis concerning the sub-universe of normal populations, $M(\prod)$, with means at a and with varying standard deviations. It only becomes precise upon definition of the manner in which σ is distributed within this sub-universe, that is to say, upon defining the *à priori* probability distribution of σ."
> Neyman and Pearson (1928, p. 189)

The Bayesian approach seems to appeal to them for treating this second case. However, as prior elicitation involves the use of inverse probability, they refrain from pursuing such a solution and move on by using their criterion of likelihood. The idea they pursued consisted of not only maximising the likelihood in the denominator under all alternatives but also maximising the likelihood in the numerator for the null hypothesis concerning the varying parameter σ. This procedure turns out to be a likelihood ratio test (see Definition C.75), and Neyman and Pearson find that their solution coincides with the well known Student's t-test.

Nevertheless, after that Neyman and Pearson eventually treat the problem via inverse probability. They propose to put a prior $\phi(a, \sigma)$ on the population $\prod$, where a is the mean and σ the standard deviation and then conclude:

> "...it is almost impossible to express ϕ in exact terms. We prefer therefore to follow a line of argument which while really equivalent to the above with ϕ assumed constant makes use of the principle of likelihood rather than the somewhat vaguer conception of *à posteriori* probability."
> Neyman and Pearson (1928, p. 193)

After noting that the inverse probability solution with a flat prior equals their MLE solution (which can be attributed to Theorem 6.7) they moved on. As the prior specification seems too arbitrary, and clear bias against the inverse probability approach was common at that time because of Fisher's influential earlier writings, they interpreted the two formally equal solutions as their MLE solution, but then noted, that likelihood

> "...as defined by Fisher is a quantity which cannot be integrated."
> Neyman and Pearson (1928, p. 194)

Neyman and Pearson proceeded by using a transformation and after finishing their derivations gave their approach via inverse probability with a flat prior a somehow crude likelihood-termed interpretation. What becomes clear from the 1928 paper is, that in the early stage of their work Neyman and Pearson both had trouble to position themselves on one side of the frequentist or Bayesian realm, or at least had different opinions on which solution to pursue first. While Pearson, of course, was motivated by his λ-test and the proposed theory of rejection regions in the form of contours, Neyman seemed to struggle with this theory and to behold to Bayesian analysis as a possible alternative.

Next to the introduction of multiple fundamental mathematical objects for their later hypothesis testing theory, the analysis of their first joint paper also shows that they had no clear probability concept when writing it. This fact becomes obvious while they proceed, and after having found that their likelihood ratio test equals the t-test derived by Gosset and Fisher they noted:

> "We may approach the problem by making use of the principle of likelihood and reaching the test given by Fisher. Suppose that we have reason to believe that two samples have been drawn from normal populations with the same standard deviation σ, but that it is necessary to compare the relative probability of two hypotheses ..."
> Neyman and Pearson (1928, p. 206)

The phrasing of relative *probability* in the context of their developed likelihood ratio test is, of course, misleading, as no statements in terms of probability are made at all by their procedure. While the Bayesian solution via inverse probability would have yielded such a statement, the likelihood ratio test only makes statements in terms of likelihood, or better, plausibility. This fault is in contrast to their earlier paying of attention to the distinction between confidence in a hypothesis and probability of a hypothesis as described previously. Again, it shows how fragile the concepts must have been at that time, neither clearly located in the frequentist philosophy nor in the Bayesian one.

In the second part of the paper, multiple issues are addressed, mainly the extension of their theory to the concepts of simple and composite hypotheses and adapting the likelihood ratio to this extension. Also, a goodness of fit test regarding the multinomial distribution is discussed, and the likelihood ratio test for this scenario is derived, which turns out to be exactly the well known Pearson χ^2 test.

Maybe the quintessence of the paper is summarised by Neyman and Pearson in the following words, which encompass what should become the new standard for hypothesis testing across science for the next decades and until today:

> "The system adopted will provide a numerical measure, and this must be coordinated in the mind of the statistician with a clear understanding of the process of reasoning on which the test is based. We have endeavoured to connect in a logical sequence several of the most simple tests, and in so doing have found it essential to make use of what R. A. Fisher has termed "the principle of likelihood." The process of reasoning, however, is necessarily an individual matter, and we do not claim that the method which has been most helpful to ourselves will be of greatest assistance to others. It would seem to be a case where each individual must reason out for himself his own philosophy."
> Neyman and Pearson (1928, p. 230)

4.3 Optimality Results and the Neyman-Pearson Lemma

While the first paper of Neyman and Pearson in 1928 was already remarkable, the collaboration went on and produced further results. In a letter dated 1st February 1930 from Neyman to Pearson, Neyman sketched ideas on how to improve their theory. In the letter, he proposed an idea of an experimental proof of the principle of likelihood. However, although only an idea, it includes as a central argument that the α level (type I errors) for their tests is fixed in advance, while simultaneously minimising the β level (type II errors). Neyman formulates the central idea as follows:

> "If we show that the frequency of accepting a false hypothesis is minimum when we use λ tests, I think it will be quite a thing!"
> Letter from J. Neyman to E.S. Pearson, dated 1st February 1930

Much later it turned out, that this indeed came as enlightenment to Neyman, as Reid (1982) noted:

> "The first real step in the solution of the problem of what today is called "the most powerful test" of a simple statistical hypothesis against a fixed simple alternative came suddenly and unexpectedly in a moment which Neyman has never forgotten. Late one evening in the winter of 1930, he was pondering the difficulty in his little office. Everyone else had gone home, the building was locked. He was supposed to go to a movie with Lola and some other friends, and about eight o'clock he heard them outside calling for him to come. It was at that moment that he suddenly understood."
> Reid (1982, p. 92)

After making his discovery, in a letter to Pearson, dated at 20th February 1930, Neyman introduced his colleague to his findings and for the first time summarised the scaffold of what later became one of the most influential hypothesis testing theories in science:

> "We test a simple hypothesis H concerning the value of some character $a = a_0$, and wish to find a contour $\varphi(x_1, ..., x_n) = c$ such that

> (1) the probability $P(\varphi_0^a)$ of a sample point lying inside the contour (which probability is determined by the hypothesis H) is equal
>
> $$P(\varphi_0^a) = \varepsilon \tag{4.3}$$
>
> where ε is a certain fixed value, say 0.01. (This is for controlling the errors in rejecting a true hypothesis) and
>
> (2) that the probability $P(\varphi_1^a)$ determined by some other hypothesis H' that $a = a_1 \neq a_0$ of a sample lying inside the same contour be maximum.
>
> Using such contours and rejecting H when $\sum$ is inside the contour, we are sure that the true hypothesis is rejected with a frequency less than ε, and that if H is false and the true hypothesis is, say, H', then most often the observed sample will be inside φ =const. and hence the hypothesis will be rejected."
> Letter from J. Neyman to E.S. Pearson, dated 20th February 1930

Here, again $\sum$ is the sample point observed, H the null hypothesis tested and H' the alternative. The first form of the modern approach of the Neyman-Pearson theory of hypothesis testing, therefore, was formulated in this letter, and another letter dated on the eight March written by Neyman expressed it even more clearly:

> "To reduce for a given level the errors of rejecting a true hypothesis, we may use any test. Now we want to find a test which would 1) reduce the probability of rejecting a true hypothesis to the level $\leq \varepsilon$ and 2) such that the probability of accepting a false hypothesis should be minimum. – We find that if such a test exists, then it is the λ-test."
> Letter from J. Neyman to E.S. Pearson, dated 8th March 1930

Two weeks later, in a letter dated on 24th March, Neyman sent Pearson his proof of what today is known as the Neyman-Pearson Lemma as given in Lemma C.76.

It took Neyman and Pearson another three years to work out the details until they finally published their paper *'On the Problem of the Most Efficient Tests of Statistical Hypotheses.'* (Neyman and Pearson, 1933) in 1933. It was a persuasive paper in which Jerzy Neyman and Egon Pearson described the results of their collaboration. In the introduction, they presented their approach, which consists of searching for rules, which govern the behaviour of the researcher, and which would become the new standard for statistical hypothesis testing:

> "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if $x > x_0$ reject H, if $x \leq x_0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false."
> Neyman and Pearson (1933, p. 291)

While Neyman and Pearson clearly outlined that for individual cases, no statements can be made at all, the appeal of a clear rule to decide between rejection and acceptance of a given hypothesis was something Fisher's theory – which was the predominant theory by that time – lacked. Indeed, as described in Chapter 3, Fisher's theory of significance testing was built upon reasoning case-based and individually, taking into account the knowledge and experience of the researcher and minute experimental design. While formally not wrong, Fisher's theory implied much more work and no easy solutions for researchers, in contrast to the Neyman-Pearson theory. Neyman's and Pearson's work was the ideal construct to sacrifice Fisher's complex and holistic theory of significance testing for the much clearer behavioural-oriented guidelines the Neyman-Pearson theory offered.

The following section *'Outline of a General Theory'* in their paper presented the various concepts such as simple and composite hypotheses as well as the two types of errors, and rejection regions. Neyman and Pearson (1933) then formulated the criterion for the best rejection region among the multitude of available rejection regions:

> "We need indeed to pick out from all possible regions for which $P_0(w) = \varepsilon$, that region w_0, for which $P_1(w)$ is a maximum (...); this region (or regions if more than one satisfy the condition) we shall term the Best Critical Region for H_0 with regard to H_1. There will be a family of such regions, each member corresponding to a different value of ε. The conception is simple but fundamental."
> Neyman and Pearson (1933, p. 297)

Here, P_0 and P_1 can be interpreted as the probability measures belonging to the Radon-Nikodym derivatives of the probability densities corresponding to H_0 and H_1. The plan of the paper was then summarised as follows by them:

> "...it will be shown below that in certain problems there is a common family of best critical regions for H_0 with regard to the whole class of alternative hypotheses Ω^*. In these problems we have found that the regions are also those given by the principle of likelihood, although a general proof of this result has not so far been obtained, when H_0 is composite."
> Neyman and Pearson (1933, p. 297)

For the case in which there are different best critical regions for H_0 with regard to each of the alternatives constituting the set of all alternative hypotheses, Ω, Neyman and Pearson (1933, p. 298) stress that it is "not clear that it has the unique status of the common best critical region of the former case.".

The rest of the paper deals with the above. It introduces the Neyman-Pearson lemma as given Lemma C.76. In their formulation it states that when testing a simple against an alternative hypothesis, for a given level α, the test maximising the probability of rejection is the likelihood ratio test at that level as given in Definition C.75. By doing so, Neyman and Pearson (1933) introduced the idea of searching optimal testing procedures. After multiple illustrations of the developed concepts, the next section deals with testing composite hypotheses and introduces the condition that for every simple hypothesis included in the composite hypothesis the probability of a type I error must be fixed at ε. If this condition is satisfied, Neyman and Pearson (1933) speak of the corresponding rejection region w of the composite hypothesis as a size ε region. The remainder of the paper then addresses characterising similar regions and illustrations of the concepts. In total, the 1933 paper is a landmark for the Neyman-Pearson theory of hypothesis testing, in that it brings together all of the central concepts and presents their theoretical results to the readership.

4.4 The Final Steps

After the 1933 paper, there was still a missing point which Neyman and Pearson had to address, as gets clear from their letters. Already two years earlier, Neyman wrote a letter dated August, 17th 1931 to Pearson in which he addresses the issue:

> "I am considering the question when there is no best critical region with regard to a given class of admissible hypotheses. What region should we then choose? I take the most simple case when the whole set of admissible hypotheses can be divided into two classes such that to each of them corresponds a 'best critical region'."
> Letter from J. Neyman to E.S. Pearson, dated 17th August 1931

To illustrate his point, he tests $\sigma^2 = 1, \mu = 0$ against $\sigma^2 \neq 1, \mu = 0$ for a normal distribution. He shows that a best critical region against the alternatives $\sigma^2 > 1$ can be found, where the variance is bigger than a calculated χ^2 threshold. Also, against the alternatives $\sigma^2 < 1$ a best critical region can be found for which a χ^2 threshold is smaller than some calculated quantity. After discussing the problem, he comes up with the option to choose the thresholds in a way such that each tail has $\varepsilon/2$ probability, which leads him to the solution:

> "Suppose we have to test a simple hypothesis H_0 with regard to a class of alternatives C with no common B.C.R. It would be no good to use a critical region w, having the following property: the class of alternatives contains a hypothesis, say H_1, such that ε_1 is $< \varepsilon$. In fact, doing so we shall accept H_0 with larger frequency when it is false (and the true hypothesis is H_1) than when it is true. (...) Therefore, the good critical region w_0 should be chosen in such a way that the probability of rejection under H_1 is $\geq$ than that under H_0."
> Letter from J. Neyman to E.S. Pearson, dated 17th August 1931

Here, B.C.R. stands for the best critical region and ε_1 is the probability of rejection under H_1. This is exactly the idea of an unbiased test as given in Definition C.77. Translating this into the current notation yields:

$$\begin{aligned} \text{Probability of Rejection under } H_1 &= \mathbb{P}(\text{reject } H_0|H_1) = 1 - \mathbb{P}(\text{accept } H_0|H_1) && (4.4)\\ &= 1 - \mathbb{P}(\text{Type II Error}) = \mathbb{E}_\theta[\varphi], \text{ if } \theta \in \Theta_0^c && (4.5)\end{aligned}$$

as is the standard notation in terms of a power function in Definition C.72. Therefore, this letter marks the introduction of the power function or power as an important concept in statistical hypothesis testing. While not clearly defined in 1931, the principal ideas were already there as the correspondence shows.

The ideas appearing in the letter dated 17th August 1931 did not get published until 1936. While it is unclear why Neyman and Pearson waited so long to publish their work, the 1936 paper *'Contributions to the theory of testing statistical hypotheses'* (Neyman and Pearson, 1936) can retrospectively be regarded as the full account of the Neyman-Pearson theory of hypothesis testing. In it, Neyman and Pearson (1936) introduce the power function for a test as in Definition C.72 and define a test to be unbiased if its power against the possible alternatives is greater or equal to the power under the null hypothesis like in Definition C.77. They also came up with the concept of uniformly most powerful tests as given in Definition C.78 and a big part of the paper deals with finding a UMP level α test. Therefore, they first looked for an unbiased, critical region with the maximum local power, which is similar to a search for an unbiased confidence set (compare Definition C.90). Neyman and Pearson (1936) proceeded by using the fact that the power function for an unbiased test of a hypothesis $H_0 : \theta = \theta_0$ has a minimum at θ_0. Therefore, the first derivate at this point is zero, and the test they were looking for can be found by maximising the second derivative with respect to θ. After illustrating this method by some applications, Neyman and Pearson (1936) moved on by using their maximisation idea and adding the constraint that the test needs to maximise the power against all possible alternative hypotheses, too. It is shown that this property is then close the notion of a UMP level α test, but not identical.

The paper is a great achievement in terms of the impact of the Neyman-Pearson theory. While in Neyman and Pearson (1933), a general and already remarkable theory was put forward, the 1936 paper can be seen as a completion of the theory of the 1933 paper via the introduction of the power function and UMP tests. From a mathematical perspective, at this point, the Neyman-Pearson theory of hypothesis testing had grown out of its early beginnings and become a serious alternative to Fisher's significance tests. Its appeal of a decision-theoretic testing procedure and the striving for optimal procedures was something that Fisher's significance tests via p-values were lacking.

In 1938 then, Neyman and Pearson (1938) published another paper which included part II and part III as an extension to their 1936 paper. In it, the structure of UMP level α tests and unbiased tests is investigated further, and it is also concerned with composite hypotheses which include more than one parameter, building upon a paper which Neyman (1937) published alone.

In summary, the Neyman-Pearson collaboration can be separated in two parts: The first parts include the early beginnings, the rough sketching of ideas, putting forward the principle of likelihood[2] as a solution to testing hypotheses and also frequent misunderstandings between both statisticians. In this phase of their collaboration, Neyman and Pearson considered both frequentist and Bayesian perspectives. In the second part of their collaboration, they dominantly focussed on a frequentist point of view. The early stages are steered by Egon Pearson, who comes up with the main ideas, and Jerzy Neyman is the more reluctant of both. Also, Neyman is not as much influenced by the writings of Fisher as Pearson seems to be, and often considers Bayesian solutions as a possible alternative. In the second stage of their collaboration, things turn around: While Pearson put up the basic concepts in the beginnings, Neyman is the one refining the whole theory later by proving the Neyman-Pearson lemma and introducing the notion of power functions, unbiased and uniformly most powerful tests into their joint theory. In total, the collaboration of Egon Pearson and Jerzy Neyman produced ten papers, of which the most important were described above. In 1938 then, Neyman left

[2] That is, their λ criterion which uses the theory of maximum likelihood of Fisher. This is not to be confused with the likelihood principle, which is detailed in Part IV and is attributed to Birnbaum (1962). Indeed, the Neyman-Pearson theory of hypothesis testing violates the likelihood principle, the reasons of which will be discussed in Chapter 10.

England for California where he was offered a statistics professorship, and the collaboration ended. In light of today's everyday work in scientific practice, the importance of their collaboration cannot be overestimated. The 1928 and 1933 papers of them had an enormous influence on statistical hypothesis testing. Their theory started a shift from Fisher's holistic significance testing towards their behavioural approach, which was justified by the long-term error control guaranteed by the Neyman-Pearson fundamental lemma. Also, as the majority of Fisher's tests could be justified by their likelihood ratio λ criterion, researchers could seamlessly shift to their approach without invalidating their previous test results. As Lehmann (2011, p. 44) notes, the Neyman-Pearson theory "continues even today to be the most commonly used approach.", although Neyman and Pearson themselves stressed that their theory can not provide any statements about the truth of a given hypothesis in a single, isolated case. Problematically, this is the most common situation in scientific research, as studies or experiments are performed once, and are seldom repeated a large number of times. Nevertheless, their theory "became one of twentieth century's most influential pieces of applied mathematics" (McGrayne, 2011, p. 49).

Chapter 5

The modern hybrid Approach

We are quite in danger of sending highly trained and highly intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be.

Ronald Aylmer Fisher
The Nature of Probability

The preceding Chapter 3 and Chapter 4 detailed the development of Fisher's theory of significance testing and the competing Neyman-Pearson theory of (uniformly most powerful) hypothesis tests.[1] Fisher developed his theory first, but most of his significance tests either assumed normality or had other specific assumptions about the statistic, for which the distribution was subsequently derived to perform the calculation of a p-value. On the contrary, the Neyman-Pearson theory offered a more structured and situation-independent approach to hypothesis testing via the use of the likelihood ratio criterion λ. This chapter details the emerging debate between both parties about which theory had to be preferred.

5.1 The Fisher-Neyman-Pearson Dissens

Fisher's early reaction to the Neyman-Pearson theory was friendly, and he was interested in the approach. Neyman asked Fisher in a letter dated 9th February 1932 to review his joint work with Pearson:

> "Presently Dr. Pearson is putting all the results in order. They will form a paper of considerable size. We would very much like to have them published in the Philosophical Transactions, but we do not know whether anybody will be willing to examine a large paper and eventually present it for being

[1] Uniformly most powerful tests, in modern notation, correspond to tests which control the type I error rate at a prespecified level $\alpha > 0$, while simultaneously minimising the type II error rate, see Rüschendorf (2014, Chapter 6) and Schervish (1995, Chapter 4.3). The λ criterion of Neyman and Pearson led to these tests, which provide long-term error guarantees under infinite repetition of an experiment.

> printed. The paper contains much of mathematics and not all the statisticians will like it just because of this circumstance. We think that the most proper critic are you, but we don't know whether you will be inclined to spend your time reading the paper..."
> Jerzy Neyman in (Bennett, 1990, p. 189)

Fisher's reply on 12th February 1932 showed his interest and that his focus already had shifted to hypothesis testing in the preceding years. Unfortunately, Fisher's review of Neyman and Pearson's paper is not available anymore, but based on the fact that the paper itself was received on August, 31st, in 1932 and published in print on February, 16th in 1933, Fisher must have been quite positive about the content [2]. After the paper of Neyman and Pearson was published (Neyman and Pearson, 1933), Fisher himself published a paper called *'Two new properties of mathematical likelihood'* (Fisher, 1934c). There, he derived the factorization lemma for sufficient statistics.[3] He showed that the existence of a real-valued sufficient statistic implies that the probability distribution has the form of a one-parameter exponential family.[4] This aspect was crucial for the Neyman-Pearson theory: Based on this fact Fisher (1934c) showed that for a UMP level α test to exist[5], a necessary criterion is the existence of such a real-valued sufficient statistic. As the Neyman-Pearson theory aims at finding such UMP level α tests, Fisher's 1934 paper was directly related to Neyman's and Pearson's theory. His results provided a criterion for the existence of a UMP level α test.

After this initial interest in Neyman's and Pearson's work, Fisher's attitude changed over time. The conflict started with Neyman rejecting Fisher's proposal that Neyman should lecture only with his book (McGrayne, 2011, p. 50), and by 1936 the quarrel between, in particular, Neyman and Fisher was becoming open hostility: According to McGrayne (2011, p. 50), "The two groups occupied different floors of the same building at University College London but they never mixed. Neyman's group met in the common room for India tea between 3:30 and 4:15 p.m. Fisher's group sipped China tea from then on." Fisher envisioned the theory of Jerzy Neyman and Egon Pearson more and more as a direct competitor to his theory of significance testing. This change in attitude becomes clear in the personal communications of Fisher. As detailed in Bennett (1990, p. 144), Fisher wrote about the Neyman-Pearson hypothesis tests in a letter to William Hick in 1951, that

> "in fact, I and my pupils throughout the world would never think of using them."
> Fisher (1951), in (Bennett, 1990, p. 192)

There are two substantial reasons why Fisher clashed with Neyman and Pearson. First, while both theories "tend to lead to the same numerical results" (Howie, 2002, p. 178), there are a few important cases in which the results produced by each theory differ (like Fisher's exact test and the traditional χ^2-test). Fisher's exact test was already briefly discussed in Section 3.1.4 in Chapter 3, where Fisher argued fervently to condition all inference on the table margins in a 2×2 contingency table, while the traditional χ^2 test of Karl Pearson ignored this information. The differences between both approaches

[2] https://royalsocietypublishing.org/doi/10.1098/rsta.1933.0009

[3] Compare Theorem C.51 in Appendix C.

[4] For a proof see Rüschendorf (2014, Theorem 4.1.21).

[5] See Appendix C, Definition C.78.

arose out of theory and not because of differences in applied work. However, both parties quickly understood that because of the differences in theory, the results obtained in practical applications – e.g. when using Fisher's exact test versus the χ^2 test, which can be shown to be a likelihood ratio test in the Neyman-Pearson theory – could differ. Clearly, this was unsatisfying for both Fisher and for Neyman and Pearson, although this happened just in a limited number of situations.

Second, Fisher saw his method of maximum likelihood as a well-founded theory of scientific inference which had to be preferred over the purely mathematical approach of Neyman and Pearson. On the other hand, Neyman and Pearson discredited Fisher's battery of significance tests as lacking a solid mathematical foundation which offered any optimality properties like the fundamental lemma which demonstrated the optimality of their hypothesis tests based on the likelihood ratio λ. However, it should be stressed that also Fisher's theory was backed up by sufficient mathematical rigour, but there was no notion of optimality to his significance tests with regard to making a type I or II error. As a consequence, his formulation of significance tests seemed less objective than the new Neyman-Pearson tests and his complex writing style did not contribute to his tests being favoured by practitioners, compare Chapter 3. Fisher's approach and his targeted audience also differed from the one of Neyman and Pearson. While both parties claimed to have given the preferable theory for statistical hypothesis testing, the differences become apparent when considering each theory from the context it evolved in. The next three subsections show that 1) different results, 2) different contexts and 3) different probability concepts can be seen as the reasons why both parties clashed.

5.1.1 Different Results

The first reason for Fisher's dissent were two cases in which both theories produce different results. One of these cases appeared to be Fisher's exact test as detailed in Section 3.1.4. Fisher (1935) tackled the problem of the 2 × 2 table and the exact Fisher test in his 1935 paper *'The logic of inductive inference'*. There, he advocated the restriction to the conditional distribution based on the (ancillary) table margins:

> "Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies."
> Fisher (1935, p. 48/49)

In 1947, Barnard G.A. (1947) proposed an unconditional test and opposed it to Fisher's solution. As described in Bennett (1990, p. 2-4), Barnard employed the Neyman-Pearson theory to guarantee the desired power of the test in the long run, that is, under hypothetical infinite repetition of the experiment. In a reply, Fisher (1948) showed that the unconditional procedure leads to false probabilities in contrast to his conditional inference via the exact Fisher test. His argument was based on the fact that one could improve upon fixing the level of significance in small samples and controlling the error rate via the Neyman-Pearson theory. He showed that his proposed test was better than the unconditional solution of Barnard G.A. (1947) for small samples, and a test whose power could be improved clearly was in contrast to the asserted optimality guarantees by the Neyman-Pearson theory:

> 'Je n'aimerais pas rehausser la signification du résultat qui a été obtenue, en raison du fait qu'une répétition du test pourrait donner une évidence moins impérative que celle effectivement obtenue. La distribution marginale dans le premier problème lu'apparaît ainsi analogue au nombre de souris classées dans le second, et devoir être acceptée comme partie des données du problème statistique correspondant, indépendamment de sa fréquence de réalisation comme résultat d'une répétition physique.'
> (Fisher, 1948, p. 213)

Neyman and Pearson attacked Fisher, too, and argued that his theory of significance testing did not include the type II error probability, which can be much more important than the type I error probability (which is gauged by Fisher's p-value) depending on the application context. Examples include diagnostic tests for a disease, where false-positive results are quickly revealed by subsequent tests and diagnostics, the associated costs of which are often moderate. False-negative results (a type II error) are more harmful as a patient with a disease receives no treatment and future costs (personal damage, economic costs for future treatments and medication) will, in general, be much larger. Also, they criticised that Fisher's theory of significance testing stood mathematically on much shallower grounds than the Neyman-Pearson theory, which appealed with its optimality results.

The second example in which both theories provide different results is the famous Behrens-Fisher-problem Fisher (1935) put forward in 1935. The Behrens-Fisher problem arose from the task of extending the Student's t-test for situations in which the group variances differ in both groups. Student's t-test assumed data in two groups were distributed as $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ with $\sigma_1^2 = \sigma_2^2$. The null hypothesis $H_0 : \mu_1 = \mu_2$ was tested against its alternative $H_1 : \mu_1 \neq \mu_2$. However, in practice, the assumption of equal variances in both groups often is unrealistic, and the Behrens-Fisher problem corresponds to a statistical test when the situation is generalized to $\sigma_1^2 \neq \sigma_2$. However, this leads to the problem that the degrees of freedom k of the resulting test statistic's distribution (which is a t_k-distribution) are dependent on the group variances σ_1^2 and σ_2^2. As these are unknown, different solutions were presented by both parties. The details are well documented by now and can be found in Lehmann (2011, Chapter 4).

As discussed in Section 3.2.3, Fisher (1955) and Cox (1958) made a strong argument for conditional inference even years later. In the fifth edition of *Statistical Methods for Research Workers*, Fisher also included his idea of conditional testing in the exact Fisher test by conditioning his inference on the marginal totals of the 2×2 table under study.

In total, the dissent of Fisher was due to the different results provided by each theory and can partially be explained by the fact that the frequentist interpretation of the Neyman-Pearson theory offered no way to respect conditional inference. As Fisher himself had also no solution for this problem, he rejected the Neyman-Pearson theory and stuck to his theory of significance testing. This attitude did not change even years later when Fisher wrote about the success of the Neyman-Pearson theory of hypothesis testing:

> "We are quite in danger of sending highly trained and highly intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In

> this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort."
> Fisher (1958b, p. 274)

5.1.2 Different Application Contexts

The second reason why both parties clashed can be attributed to the different contexts in which each theory was developed. As Howie (2002) noted, both theories are separated by a "fundamental difference in philosophy." (Howie, 2002, p. 178). Fisher saw his theory as a self-contained theory for scientific inference and was a practitioner whose work was influenced by agricultural work and scientific experimentation. He had the talent to balance mathematical theory and experimental practice. The strength of Fisher's significance tests consisted in the idea of combining his selection of likelihood-based tests with careful experimental design and domain-specific knowledge as highlighted for example by the Latin Square design in agricultural experiments or Fisher's exact test in medical statistics.

On the other hand, Jerzy Neyman and Egon Pearson developed a mathematical theory by pencil and paper. Howie (2002) underlines the important aspect that the Neyman-Pearson theory was designed to "govern behaviour: it gives a self-contained decision strategy between courses of action." (Howie, 2002, p. 178) and "is particularly suited to practical applications, for which the benefits and penalties associated with various well-defined hypotheses can often be quantified." (Howie, 2002, p. 178).[6]

In the following, the different application contexts are exemplified by considering two examples: First, a medical test for a disease which is conducted routinely is considered. A type I error happens, when a healthy patient gets a positive test result, and a type II error happens if a sick patient gets a negative test result. The Neyman-Pearson theory is a perfect match for such a situation, as it exactly resembles the idea behind the theory. The test is conducted repeatedly under (approximately) identical conditions (a large number of patients, all of which are assumed to be members of a homogeneous population), and it is assumed that the costs of type I and II errors are known or at least can be estimated roughly. From a medical and economic perspective, the long-run consequences in the form of costs thus can be evaluated. These may be the individual costs of suffering due to an undetected disease or the financial costs of treatment, or any other quantity previously defined.[7] On the other hand, Fisher's significance testing would consider each test separately, weighing the evidence from case to case and incorporating expert knowledge as well as a minute experimental design. While formally not wrong, the goal in Fisher's theory is not to control or minimise the long-run error frequencies and the associated loss. In the above context, however, minimising the long-term loss can be seen as the primary goal of the diagnostic test. As a consequence, in this first example, the Neyman-Pearson theory is an appropriate choice.

[6] Also, Neyman-Pearson hypothesis tests can be formalised quite easily from a modern decision-theoretic perspective as shown in Appendix C via the Neyman-Pearson loss function. This presented another strong justification of their theory through the later work of Wald (1939, 1949). However, quantification of the associated loss of a type I or II error is, in almost all realistic research situations, nearly impossible (Robert, 2007), which weakens this argument.

[7] Such modelling can easily be incorporated by adapting the used loss function, compare Appendix C. For example, one can use different losses L_0 and L_1 in the Neyman-Pearson loss function to quantify the different losses implied by a type I or II error.

As a second example, consider a situation in which a new drug is tested and compared to the standard treatment. The study involves a multitude of complex and interdependent factors, including individual properties of the participating patients, the geographical region where the study is conducted, the associated environmental impacts, and the experimental design. While formally a repetition of a study is possible, a repetition under *exactly* the same conditions is extremely difficult if not impossible to produce. Even an approximate repetition of a study is challenging and often not possible.[8] This also is the case for agricultural studies, in which the soil type, geographical region, plant type or fertilizer used play an important role. Such studies were the everyday work for Fisher in Rothamsted, which may be seen as another reason why Fisher rejected the Neyman-Pearson theory.[9] Also, the costs of a type I error – that is, concluding that the drug works although it barely has any effect – are much more difficult to estimate than the costs of a false-positive outcome for a single patient. For example, while the economic and individual costs for a false-positive diagnosis of a single patient can be roughly estimated (although even this task can quickly become challenging depending on what the test diagnoses), the costs of a false-positive diagnosis in the context of new drug development are more difficult to estimate. These depend on the context and number of patients which are treated with it, the time span the drug will be used until it is eventually noticed that it has no effect, and the economic costs which then depend on the previous variables. Therefore, in such cases, Fisher's theory of significance testing is the appropriate choice, and the long-term oriented Neyman-Pearson theory is of limited use. Howie (2002) puts it this way concerning the Neyman-Pearson theory and hypothesis testing:

> "Decisions concern the rational way to behave: what one wants to know is not whether a given hypothesis is true, but whether one should act as if it is."
> Howie (2002, p. 178)

On the long run, when acting this way, the Neyman-Pearson theory guarantees that one does not err too often. Problematically, no statement about the hypothesis tested in the current study can be made. This questions the usefulness of a Neyman-Pearson hypothesis test, in particular, for scientific research. For Fisher, every experiment had the goal of revealing new knowledge to the experimenter, and decisions should be made to reveal the truth about a hypothesis. For Neyman and Pearson, the goal was to guide the behaviour of scientists to produce predictable results on the long run. As they stressed:

> "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong."
> (Neyman and Pearson, 1933, p. 291)

The usefulness of the Neyman-Pearson theory, therefore, is dependent on two factors: First, determining the costs of a type I or II error must be possible. Statistically, the

[8] Compare the recent replication attempts of various studies in the biomedical and cognitive sciences (Wagenmakers and Pashler, 2012; Pashler and Harris, 2012) and Chapter 1.

[9] Regarding Rothamsted, McGrayne (2011) noted that "Fisher's job was analyzing volumes of data compiled over decades about horse manure, chemical fertilizers, crop rotation, rainfall, temperature, and yields."

associated loss with a false decision thus needs to be quantified, and the reliability of this loss estimate determines the validity of the whole theory. This might be a difficult if not impossible task, as Howie (2002) notes concerning the complexity in biomedical research:

> "By how much is it better to mistake a genetic factor than to risk defying publication regarding a possibly remediable disease?"
> Howie (2002, p. 179).

Second, the experiment has to be repeatable under the same conditions, which is problematic if not impossible for studies conducted in the medical, psychological and social sciences. Exceptions are given especially in the area of quality control, which was an active interest of Egon Pearson's research (Pearson, 1933). For example, the Neyman-Pearson theory is a perfect match for controlling the number of defect items produced by a machine. This shows the separate areas of application both Neyman and Pearson as well as Fisher came from, which attributed to the dissent and made the arguments of the other party difficult to understand for the other side. In contrast, in Fisher's everyday work at Rothamsted, experiments were at best approximately repeatable under the same conditions, and quantification of the loss when making a type I or II error was also difficult, although not impossible.[10]

Another objection of Fisher to the Neyman-Pearson theory was given by the fact that in some cases neither the null nor the alternative hypothesis is true, and therefore none of both hypotheses should be accepted. Both hypotheses can be a bad description of the exact situation at hand, and Fisher's significance testing would, in this case, be consistent in just rejecting the null, but not accepting any alternative (Fisher, 1939). Of course, this interpretation is only advantageous, if one is concerned with a single study or experiment at hand. When the long-term error rate needs to be controlled, the single study or experiment at hand barely matters and as a consequence, accepting the alternative is perfectly fine to minimise the incurred loss under infinite repetition. Also, from Neyman's and Pearson's perspective, not rejecting the null hypothesis via Fisher's p-value does not imply that the alternative is true, so no knowledge is gained by conducting Fisher's significance test when a result is not judged to be significant. However, in contexts like quality control a sample of produced items out of which a fraction is defect requires to take some action like improving the machine when too many items are defect, or acting as if the machine produces at most a specific percentage of defect items.

Summing up, different application contexts can be seen as the second cause why both parties rejected the other theory.

[10] Returning to the agricultural experiments Fisher conducted at Rothamsted station, it would have been possible to use literally the loss in crop yield when using a different fertilizer, soil type or plant. However, when testing the efficacy of a fertilizer, the loss of interest associated with a false decision would be: What is the loss when we reject the hypothesis of increased efficiency and do not use this alternative fertilizer from now on? This loss depends on a variety of aspects: Where would it have been applied? How large is the true increase in efficacy? What is the difference in crop yield that is lost by not using the alternative fertilizer? This shows how quickly it becomes impossible to estimate the loss with a decision, which in this case would be a false-negative one.

5.1.3 Different Probability Concepts

Another reason for the dissent between Neyman, Pearson and Fisher can be found in the different probability concepts both parties had. Fisher stated his concept of probability quite early, in the 1922 paper *On the mathematical foundations of theoretical statistics* (Fisher, 1922b):

> "When we speak of the probability of a certain object fulfilling a certain condition, we imagine all such objects to be divided into two classes, according as they do or do not fulfill the condition. This is the only characteristic in them of which we take cognisance. For this reason probability is the most elementary of statistical concepts. It is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit. For example, when we say that the probability of throwing a five with a die is one-sixth, we must not be taken to mean that of any six throws with that die one, and one only will necessarily be a five; or that of any six million throws, exactly one million will be fives; but that of a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives. Our statement will not then contain any false assumption about the actual die, as that it will not wear out with continued use, or any notion of approximation, as in estimating the probability from a finite sample, although this notion may be logically developed once the meaning of probability is apprehended."
> Fisher (1922b, p. 312)

Thus, Fisher's probability concept, while vague and involving quite unwieldy constructs like a hypothetical infinite population, was at its heart a frequentist one following the Laplacian tradition (Salsburg, 2001). Contrary to the epistemic probability concepts of proponents of the approach of inverse probability, Fisher thought of probability as an objective quantity which could be measured by precise statistical procedures like maximum likelihood and rigorous experimental design.

In a much later paper in 1958 titled *The Nature of Probability*, Fisher gives more insight about his precise definition of probability:

> "Probability is, I suggest, the first example of well specified state of logical uncertainty. Let me put down a short list of three requirements, as I think them to be, for a correct statement of probability, which I shall then hope to illustrate with particular examples. I shall use quite abstract terms in listing them.
>
> (a) There is a measurable reference set (a well-defined set, perhaps of propositions, perhaps of events).
>
> (b) The subject (that is, the subject of a statement of probability) belongs to the set.
>
> (c) No relevant sub-set can be recognized."
>
> Fisher (1958b, p. 263)

It is important to note that Fisher was concerned with probability as the quantity which a statistician needs to use for making inferences, and not with probability from a formal mathematical point of view. However, one can reformulate the above three points from a modern perspective. As an example of the measurable reference set[11], Fisher used the set of possible throws with a die in his 1922 paper. In modern notation, the measurable reference set is simply the power set $\mathcal{P}(\Omega)$ of the event space $\Omega := \{1, ..., 6\}^{\mathbb{N}}$. His second condition then makes it possible to make a probability statement by requiring the subject to belong to the measurable reference set, which in modern notation requires an event A to be a subset $A \subset \mathcal{P}(\Omega)$ of the σ-algebra associated with the event space Ω to use any form of probability statements about the event A. The third requirement states that no relevant subset of such a set may exist having a *different* probability. When a relevant subset exists and yields a different probability, Fisher's conditional inference mandates to condition the inference on this subset before providing any probability statement about the set A. The last statement is tied closely to his earlier ideas about conditional inference, as the existence of a relevant subset means that from the statistical inference perspective, the probability statement made is ignoring relevant information when a relevant subset exists and is ignored. A correct statement of probability in Fisher's sense, therefore required to make use of conditional inference. This third requirement shows how strongly Fisher was concerned with hypothesis tests when thinking about probability. Also, as conditioning a (Maximum-Likelihood-)estimator for a parameter in a statistical model on a sufficient statistic could reduce the variance of the estimator according to the Cramér-Rao inequality[12], Fisher was strongly convinced of conditional inference even in contexts like parameter estimation. In modern terms, while Fisher's concept of probability was based on frequencies, it still can be called epistemic, as it refers to single events and a measure of *rational* uncertainty.[13]

For Neyman (and Pearson), probability had a different definition. Neyman referred to probability as the frequency of *future events*, which led to a different hypothesis testing theory. More specific, Neyman detailed his concept of probability in *'Outline of a theory of statistical estimation based on the classical theory of probability'* (Neyman, 1937). He first introduced confidence intervals and by doing so, gave insights about his probability concept. Neyman first noted:

> "... we shall need to define the terms probability, random variable, and probability law. These definitions are needed not because I introduce some new conceptions to be described by the above terms, but because the theory which is developed below refers only to some particular systems of the theory of probability which at the present time exist,* and it is essential to avoid misunderstandings."
> Neyman (1937, p. 336)

Neyman continued by

> "I want to emphasize at the outset that the definition of probability as given below is applicable only to certain objects A and to certain of their properties

[11] Fisher's definition of measurable is not to be confused with the modern measure-theoretic definition of measurability.

[12] Compare Chapter 3, and for a modern proof of the Chapman-Robbins inequality, of which the Cramér-Rao inequality is just a special case, see Rüschendorf (2014, Chapter 5).

[13] This is not to be conflated with epistemic probability in a sense it often is attributed to the inverse probability approach, as defended by Jeffreys (1931). This concept of probability, while also called epistemic, is not based on a frequency definition. For Jeffreys (1931), probability measured belief in a specific proposition relative to the given data and involved no frequencies.

> B – not to all possible. In order to specify the conditions of the applicability of the definition of the probability, denote by (A) the set of all objects which we agree to denote by A. (A) will be called the fundamental probability set. Further, let (B) denote the set of these objects A which possess some distinctive property B and finally, $((B))$, a certain class of subsets (B'), (B''), . . ., corresponding to some class of properties B', B'', etc.
> It will be assumed†
> (1) that the class $((B))$ includes (A), so that $(A)\ \varepsilon\ ((B))$ and
> (2) that for the class $((B))$ it was possible to define a single-valued function $m(B)$, of (B) which will be called the measure of (B). The sets (B) belonging to the class $((B))$ will be called measurable."
> Neyman (1937, p. 336-337)

In assumption (1), ε can be read as $\subseteq$. The above shows that Neyman already had a quite modern definition of probability. The quantity (A) may be expressed in modern terms as the event space Ω, the set $((B))$ as the σ-algebra on Ω, and m as the probability measure on the σ-algebra. Neyman also references Kolmogorov in a footnote on the same page, and notes that "A systematic outline of the theory of probability based on that of measure is given by Kolmogorov (1933). See also Borel (1925-1926); Lévy (1925); Fréchet (1937)." (Neyman, 1937, p. 336). Then, Neyman specifies the properties of the measure m:

> "The assumed properties of the measure are as follows:
> (a) Whatever (B) of the class $((B))$, $m(B) \geq 0$.
> (b) If (B) is empty (does not contain any single element), then it is measurable and $m(B) = 0$.
> (c) The measure of (A) is greater than zero.
> (d) If (B_1), (B_2) ... (B_n) ... is any at most denumerable set of measurable subsets, then their sum, $\sum(B_i)$, is also measurable. If the subsets of neither pair (B_i) and (B_j) (where $i \neq j$) have common elements, then $m(\sum B_i) = \sum_{i=1}^{\infty} m(B_i)$.
> (e) If (B) is measurable, then the set $(\overline{B})$ of objects A non-possessing B is also measurable and consequently, owing to (d), $m(B) + m(\overline{B}) = m(A)$.
> Under the above conditions the probability, $P\{B|A\}$, of an object A having the property B will be defined as the ratio $P\{B|A\} = \frac{m(B)}{m(A)}$. The probability $P\{B_1|A\}$ or $P\{B_1\}$ for short, may be called the absolute probability of the property B. Denote by B_1B_2 the property of A consisting in the presence of both B_1 and B_2. It is easy to show that if (B_1) and (B_2) are both measurable then (B_1B_2) will be measurable also. If $m(B_2) > 0$, then the ratio, say $P\{B_1|B_2\} =: m(B_1B_2)/m(B_2)$, will be called the relative probability of B_1 given B_2. This definition of the relative probability applies when the measure $m(B_2)$ as defined for the fundamental probability set (A) is not equal to zero."
> Neyman (1937, p. 337)

The above quote shows that although properties like σ-additivity were already required by Neyman and from his definition $P\{B|A\} = \frac{m(B)}{m(A)}$ one obtains $P\{A|A\} = 1$, so that the definition of Kolmogorov for a probability measure is recovered from Neyman's specification. Summing up, Neyman had a frequentist probability concept which already contained modern measure-theoretic concepts like measurability and which adhered to Kolmogorov's axiomatic. However, at the core this concept was just a Laplacian one like Fisher's concept of probability, and the important but subtle differences are clarified when investigating the application of the probability concept in statistical contexts. In what follows, Neyman's derivation of confidence intervals is outlined to illustrate the differences between Neyman's and Fisher's idea of applying a frequentist probability concept in practice.

In the same paper, Neyman considered random variables $X_1, ..., X_n$ following the probability density $p(x_1, ..., x_n|\theta_1, \theta_2, ..., \theta_l)$ which depends on the parameters $\theta_1, \theta_2, ..., \theta_l$, "which are constant (not random variables), and that the numerical values of these parameters are unknown." (Neyman, 1937, p. 347). After stating his frequentist interpretation of the observed data (which are the realisation of a random variable) and the unknown parameter (which is a fixed but unknown constant), Neyman stated the goal:

> "It is desired to estimate one of these parameters, say θ_1. By this I shall mean that it is desired to define two functions $\bar{\theta}(E)$ and $\underline{\theta}(E) \leq \bar{\theta}(E)$, determined and single valued at any point E of the sample space, such that if E' is the sample point determined by observation, we can (1) calculate the corresponding values of $\underline{\theta}(E')$ and $\bar{\theta}(E')$ and (2) state that the true value of θ_1, say θ_1^0, is contained within the limits $\underline{\theta}(E') \leq \theta_1^0 \leq \bar{\theta}(E')$"
> Neyman (1937, p. 347)

Interestingly, Neyman (1937) first considered the Bayesian approach to such a task as more appropriate, similar to his earlier preference of the Bayesian approach for hypothesis testing in the correspondence with Egon Pearson at the beginning of their collaboration. He noted:

> "...under the influence of Bayes Theorem, we could ask that, given the sample point E', the probability of θ_0^1 falling within the limits (...) should be large, say $\alpha = .99$, etc. If we express this condition by the formula
>
> $$P\{\underline{\theta}(E') \leq \theta_1^0 \leq \bar{\theta}(E')|E'\} = \alpha$$
>
> we see at once that it contradicts the assumption that θ_1^0 is constant."
> Neyman (1937, p. 347-348)

Although the Bayesian approach seemed to be ideally suited to the problem at hand, the philosophical assumption that the unknown parameter θ is fixed and the observed data E are a random variable[14] permitted to proceed. Specifically, when θ_1^0 is an unknown, fixed constant, it follows that

$$P(1 \leq \theta_1^0 \leq 2) = \begin{cases} 1, \text{ if } 1 \leq \theta_0^1 \leq 2 \\ 0, \text{ if } \theta_0^1 > 2 \text{ or } \theta_0^1 < 1 \end{cases}$$

where in the above, the left and right boundaries 1 and 2 can be replaced with any range of values. This demonstrates that confidence intervals of the Neyman-Pearson theory do not allow for any probability statements about the parameter except for the trivial ones that the parameter is either located in the interval or not. In this example, Neyman

[14] See Appendix C.

argued that one cannot say that the probability of the true value θ_0^1 falling between one and two is equal to α. Either, the fixed constant θ_0^1 is between one and two in which case the probability is one (first case above), or θ_0^1 is not, in which case the probability is zero (second case above).

After finding that it is not possible to proceed with the Bayesian approach when starting from a frequentist perspective, he started again by considering the accuracy of an estimate T for the unknown parameter θ and noticed that the following two estimates could quantify the accuracy

$$\underline{\theta} = T - k_1 S_T \text{ and } \bar{\theta} = T + k_1 S_T$$

which indicate "the limits between which the true value of θ presumably falls." (Neyman, 1937, p. 347). In the above, $k_1, k_2 \in \mathbb{R}$ are constants which have to be chosen appropriately and S_T is the sample's standard deviation. Neyman then underlined that it is possible to consider this probability at all because from a frequentist perspective, the true unknown value θ_1^0 of the parameter θ is assumed to be fixed. The functions $\underline{\theta}$ and $\bar{\theta}$ from a frequentist perspective are random variables, as they are functions of the randomly observed data E'. Neyman then required the functions $\underline{\theta}$ and $\bar{\theta}$ to be chosen so that the probability of $\underline{\theta}(E') \leq \theta \leq \bar{\theta}(E')$ is constant, that is equal to $\alpha \in \mathbb{R}$. Neyman called the values $\underline{\theta}$ and $\bar{\theta}$ lower and upper confidence limit and the range between them confidence interval. The constant α he called the confidence coefficient:

> "The functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$ satisfying the above conditions will be called the lower and the upper confidence limits of θ_1. The value α of the probability (...) will be called the confidence coefficient, and the interval, say $\delta(E)$, from $\underline{\theta}(E)$ to $\bar{\theta}(E)$, the confidence interval corresponding to the confidence coefficient α."
> Neyman (1937, p. 348)

Neyman then moved on to sketch the three steps to calculate these quantities:

> "We can then tell the practical statistician that whenever he is certain that the form of the probability law of the X's is given by the function $p(E|\theta_1, \theta_2, ..., \theta_l)$ which served to determine $\underline{\theta}(E)$ and $\bar{\theta}(E)$, he may estimate θ_1^0 by making the following three steps : (a) he must perform the random experiment and observe the particular values $x_1, x_2, ..., x_n$ of the X's ; (b) he must use these values to calculate the corresponding values of $\underline{\theta}(E)$ and $\bar{\theta}(E)$; and (c) he must state that $\underline{\theta}(E) < \theta_1^0 < \bar{\theta}(E)$, where θ_1^0 denotes the true value of θ_1. How can this recommendation be justified?"
> Neyman (1937, p. 348)

After giving the justification, which is the strong law of large numbers, Neyman (1937, p. 349) noted that "it follows that if the practical statistician applies permanently the rules (a), (b) and (c) for purposes of estimating the value of the parameter θ_1 in the long run he will be correct in about 99 per cent. of all cases.", where $\alpha = 0.99$. Neyman also noted that due to the strong law of large numbers, if the upper and lower limits $\underline{\theta}(E)$ and $\bar{\theta}(E)$ are calculated properly, "the frequency of actually correct statements will approach α" (Neyman, 1937, p. 349). Therefore, the confidence intervals as introduced by Neyman do *not* attain this probability α of correct statements for every sample size. After that, Neyman finally gave his definition *how* the statistician should apply his frequentist probability concept detailed above in practice:

> "It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results tend to α.* Consider now the case when a sample, E', is already drawn and the calculations have given, say, $\underline{\theta}(E') = 1$ and $\bar{\theta}(E') = 2$. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α? The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made ... "
> Neyman (1937, p. 349)

This passage not only gives Neyman's personal view on probability as the basis for his hypothesis testing theory but also provides the reader with a direct warning of the probably most common misinterpretation of frequentist confidence intervals.[15] Somehow, these warnings have been overlooked by too many researchers over time, so that a Bayesian credible interval perspective is attributed most often to frequentist confidence intervals, leading partially to the problems observed in the replication crisis. The most severe problem of confidence intervals in the interpretation of Neyman may be given by the fact, that while they provide the guarantee that one is correct in α percent of the cases when stating that the parameter is included inside a confidence interval in a large succession of such intervals, one cannot provide a probability that the parameter lies inside any single confidence interval with probability α. Thus, inference for the study at hand is not possible.

The rest of Neyman's paper is concerned with finding the best – that is, shortest – confidence intervals which minimize the probability of false coverage. Neyman found that these intervals are uniformly shortest – also called Neyman-shortest – if the corresponding hypothesis test is a UMP level α test, see Casella and Berger (2002).[16]

The analysis of the probability concepts of both parties shows, why Fisher, Neyman and Pearson clashed bitterly. Fisher had an epistemic probability concept which, although at its core a frequentist one, referred to probability as the empirical probability of singular events. Jerzy Neyman opposed this view with his frequentist concept of probability which in applications was interpreted as providing statements about *future events*. This perspective settled into the Neyman-Pearson theory of hypothesis testing, which controls the α level only on the long-run, that is, for future events.[17] Of course, this position was not reasonable for Fisher, who treated probability as a concept referring to the *current singular event*. For singular events like the study or experiment currently conducted, the Neyman-Pearson theory could not make any statement, no matter if the goal were to test a hypothesis or quantify the uncertainty in a parameter estimate. As noted by Halpin and Stam (2006):

> "... unlike Fisher, Neyman considered statistical testing to make no contribution to the problem of inductive reasoning (Neyman, 1942, 1950, introduction). This foundational distinction between Fisher and Neyman regarding

[15] A review of some more misinterpretations and pitfalls when interpreting frequentist confidence intervals in the spirit of Neyman and Pearson are given in Morey et al. (2016).

[16] This follows immediately from the duality between Neyman-Pearson-tests and Neyman's confidence intervals, compare Appendix C.

[17] Notice that a hypothesis test φ for level α in the frequentist interpretation controls the type I error rate only in expectation. That is, $\mathbb{E}_\theta[\varphi] \leq \alpha$, compare Definition C.71. Thus, it is *not* possible to answer the question whether in the current study or experiment, a type I error has happened or not.

> the purpose of statistical testing can be traced to their respective conceptions of probability."
> Halpin and Stam (2006, p. 632)

Therefore, Fisher preferred his case-based procedure of significance testing, which additionally incorporated conditional inference. Halpin and Stam (2006) further noted:

> "The Neyman-Pearson theory generally has been interpreted as a decision theory rather than a theory of inference and has found its least problematic applications in quality control in industry (see Seidenfeld 1979, for a critique of the Neyman-Pearson approach as an unintended theory of inference)."
> Halpin and Stam (2006, p. 632)

Noticeably, the above analysis also reveals that confidence intervals emerged out of the inapplicability of the Bayesian approach to the frequentist assumptions about the state of nature concerning the unknown parameter θ and the observed data E. Had Neyman started with the Bayesian assumptions outlined in Appendix C, which regard the observed data E as fixed and the unknown parameter θ to be a random variable, he could have proceeded perfectly fine. It is remarkable that Neyman seems to have had a strong preference for the Bayesian approach both when publishing his likelihood ratio criterion, the fundamental lemma and also when introducing frequentist confidence intervals. The only reason that prevented him to proceed in a Bayesian way seemed to be the general rejection of inverse probability at that time due to Fisher's earlier writings.

5.2 Prespecified Test Levels α versus p-values

As described in the preceding section, a major argument for the debate between Fisher, Neyman and Pearson was that Fisher conducted inference conditionally, while Neyman and Pearson did not. This difference is directly tied to the use of p-values by Fisher and the use of fixed α levels by Neyman and Pearson. In addition to the advocation of the conventional .05 threshold in *Statistical Methods for Research Workers* as described in Section 3.2.1 and Section 3.2.2, Fisher gave deeper insights into his attitude towards p-values in a 1926 paper titled *'The arrangement of field experiments'* (Fisher, 1926), where he stated that

> "...for it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." This level, which we may call the 5 per cent. point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials...
> If one in twenty does not seem high enough, we may, if we prefer, draw the line at one in fifty (2 per cent. point), or one in hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard at the 5 per cent. point, and ignore entirely all results which fail this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance."
> Fisher (1926, p. 85)

It is important to note that Fisher here implicitly required successful replications of significant study results, where the replication studies are required to be conducted under the same conditions. Also, proper experimental design is needed, and if these two requirements are fulfilled, the scientific fact of interest is regarded as experimentally established.

From the above writings, it becomes also clear that the p-value was a continuous quantity for Fisher, and the advocation of the .05 per cent threshold should not be used carelessly. Continuous is interpreted here not in the mathematical sense, but indicates that the p-value is interpreted as quantifying the evidence against the null hypothesis directly. For example, a p-value $p = 0.049$ is interpreted different from a p-value $p = 0.001$ in the continuous interpretation, while in the Neyman-Pearson interpretation, which is binary, both of the above p-values are simply significant, when a test level $\alpha = 0.05$ is chosen. While formally, the Neyman-Pearson theory does not include p-values, the test levels in the Neyman-Pearson theory are in practice equal to Fisher's p-values, as both objects are computed via the same probability, see Appendix C.

Nevertheless, Fisher kept the wording of the 5% threshold in *Statistical Methods for Research Workers* until the twelfth edition, which was published in 1954. In the thirteenth edition, published in 1958, four years before Fisher died, this wording was changed into:

> "The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded."
> Fisher (1958a, p. 80)

Here again, the continuous interpretation of p-values as evidence against the null hypothesis H_0 in contrast to the binary interpretation of the p-value in the Neyman-Pearson theory becomes clear. In the thirteenth edition, Fisher warned the reader of a common misinterpretation of his p-values:

> "The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified."
> Fisher (1958a, p. 80)

Goodness of Fit here refers to the χ^2 test for the goodness of fit of an empirical distribution, often listed in a contingency table for a theoretically assumed distribution (Pearson, 1900). As the above quote shows, Fisher denoted p-values with a capital P, which can be read as probability. This is reasonable, as the p-value is the probability of obtaining a result equal to or more extreme than the one observed under assumption of the null hypothesis H_0. Conceptually, there is no difference between p in the modern notation and P in Fisher's writings. However, Fisher's warnings have remained widely unheard, as the official ASA statement more than half a century later shows when reporting some of the most frequent misuses of p-values in today's research[18]:

> "misconceptions and misuse of the p-value, are the following: (...) p-values do not measure the probability that the studied hypothesis is true"
> Wasserstein and Lazar (2016, p. 2)

[18] Gigerenzer (2004) notes: "Early authors promoting the error that the level of significance specified the probability of hypothesis include Anastasi (1958, p.11), Ferguson (1959, p.133), and Lindquist (1940, p.14). But the belief has persisted over decades: for instance, in Miller and Buckhout (1973; statistical appendix by Brown, p. 523), and in the examples collected by Bakan (1966), Pollard and Richardson (1987), GIgerenzer (1993), Mulaik et al. (1997), and Nickerson (2000)." (Gigerenzer, 2004, p. 597)

What can be learned further from the various editions of *Statistical Methods for Research Workers* is that Fisher was rarely interested in p-values themselves, but much more in deciding whether or not the results could be deemed *significant*. This attitude can be witnessed by checking his examples, in which he nearly always used his 5% level. In the thirteenth edition, examples 8, 11, 12, 27, 28, 35 and 37 show that he always compared his calculated p-value to the 5% threshold. Below, Fisher's conclusions regarding example 8, 12 and 28 are given:

> "... $\chi^2 = 10.87$, the chance of exceeding which value is between .01 and .02; if we take $P = .05$ as the limit of significant deviation, we shall say that in this case the deviations from expectation are clearly significant."
> Fisher (1958a, p. 81)
>
> "For $n = 9$, the value of χ^2 shows that P is less than .01, and therefore the departures from proportionality are not fortuitous."
> Fisher (1958a, p. 89)
>
> "Calculating t from t as before, we find $t = 2.719$, whence it appears from the table that P lies between .02 and .01. The correlation is therefore significant."
> Fisher (1958a, p. 196)

The examples illustrate that although Fisher advocated to report the precise p-value to quantify the evidence against the null hypothesis H_0, in his most influential textbook he focussed mostly on the *significance* of the p-value compared to a prespecified threshold like 0.05. This resembles a binary interpretation similar to the Neyman-Pearson interpretation of p-values, and is contradictory to the case-by-case oriented evidential perspective Fisher was a fervent proponent of at the same time. Together, this can be seen as one reason why readers had difficulties in separating Fisher's from Neyman's and Pearson's methodology and eventually an inconsistent blend of both theories in form of a hybrid approach emerged. Although Fisher advertised to quantify the p-value continuously, in nearly all examples he presented the decisions of a significance test were based on a fixed threshold like 0.05, which strongly resembled the prespecified test level α in Neyman and Pearson's hypothesis testing theory.

Chapter 4 already described the Neyman-Pearson theory of hypothesis testing in detail, and Table 5.1 provides an overview of the differences between it and Fisher's theory of significance testing. In quintessence, the Neyman-Pearson theory tries to provide the researcher with rules governing the behaviour regarding the acceptance and rejection of hypotheses[19], which on the long run – that is, when the experiment is hypothetically repeated an infinite number of times – leads to the test level α.[20] This is not to be confounded with the asymptotic distribution of certain tests, like LRT tests, compare Appendix C, where the level α is attained asymptotically for convergence in distribution. These properties mainly address the sample size and are inherent to the asymptotic distribution of the tests, not to the inferential method itself. The Neyman-Pearson theory

[19] The emphasis on acting as if the hypothesis were true stems from the decision-theoretic foundation of the Neyman-Pearson theory, see Appendix C: In statistical decision-theory, the incurred loss is quantified when deciding for an action, which in the context of hypothesis testing is either accepting or rejecting H_0. Therefore, the expected loss is minimised when one *acts* as if H_0 (or H_1) were true whenever accepting it. If H_0 (or H_1) *is* true, cannot be answered by the Neyman-Pearson theory.

[20] The idea of long-run error control manifests itself mathematically in the Neyman-Pearson theory in the form that statements about the error guarantees are made *in expectation*, see Rüschendorf (2014, Remark 2.1.8 (d)).

	Fisher's Significance Testing	Neyman-Pearson Hypothesis Testing	Hybrid approach
Setup	Set up a single statistical null hypothesis H_0. This hypothesis serves for the interpretation of experimental results and specifies the distribution under which the experimental data is assessed mathematically.	Set up two statistical hypotheses H_0 and H_1, set a fixed level α, β and sample size n before conducting the experiment, based on the costs associated with a type I and II error, and the time and costs required to collect n observations. These values define a rejection region for H_0.	Set up a statistical null hypothesis H_0 which supposes that there is "no effect".
Analysis	A *significance test* is conducted, and the exact level of significance is reported, which is the probability of obtaining a result equal to or more extreme than the one observed under assumption of H_0, that is, the p-value. No conventional threshold like 0.05 is used, but subject-domain knowledge must be incorporated. If the p-value is small enough, the result is *significant*, and H_0 is rejected. Otherwise, no conclusions are drawn unless more data is accumulated.	A (decision) *rule of inductive behaviour* is applied, which is a hypothesis test in the Neyman-Pearson theory. If the data fall into the rejection region of H_0 (which is the case when $P(D\|H_0) < \alpha$)), reject H_0 and accept H_1. Else, reject H_1 and accept H_0. Accepting (or rejecting) an hypothesis means *not* to believe in it (or not to believe in it), but only to *act* as if it were true (or false).	Conduct a Neyman-Pearson test with test level α (often, $\alpha = 0.05$). Compare the calculated p-value to α. If $p < \alpha$, the p-value is significant, reject H_0 and report $p < \alpha$ as well as the p-value as the continuous quantification of strength against H_0.
Interpretation of results	Use the procedure only if little is known about the situation to be investigated, and only to draw conclusions as an attempt to understand reality. If the result is significant, either the null hypothesis H_0 is false, or an unlikely event has occurred.	Adopt a specified course of action corresponding to which hypothesis has been accepted. The procedure is used to control the long-term error rates of hypothesis tests. Therefore, α, β and n need to be selected and balanced carefully.	Interpret the calculated p-value as the strength of evidence against H_0.

Table 5.1: Fisher's Significance Testing and Neyman-Pearson hypothesis testing

itself can only make statements regarding α under a hypothetically infinite repetition of the experiment. The advantages, on the other hand, are the direct applicability to a multitude of situations by the use of the likelihood ratio test as given in Definition C.75.

In contrast, Fisher's significance tests use only a null hypothesis instead of a null and an alternative hypothesis. Additionally, the rejection or acceptance of a hypothesis in a decision-theoretic way as in the Neyman-Pearson theory is not allowed in Fisher's significance testing. His theory relies on falsification of the null hypothesis H_0 by declaring *significance* of the observed results. Depending on the size of the p-value, "the strength of the evidence against the hypothesis." (Fisher, 1958a, p. 80) is quantified by Fisher's significance tests. Also, rigorous experimental design and conditional inference need to be incorporated for producing reliable conclusions. It is of paramount importance to underline that a *significant* result in Fisher's interpretation is only a provisional insight into the situation at hand. The ultimate goal of experimentally establishing scientific facts requires successful replications of an experiment or study according to Fisher:

> "A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance." Fisher (1948, p. 85)

In summary, his testing methodology can be regarded as a self-contained scientific theory. The Fisherian methodology also had problems: It was not allowed to accept or confirm a hypothesis, which is desirable in a variety of research settings. Also, the concept of conditional inference was difficult to implement for practitioners, and the vague recommendations of rigorous experimental design complicated things even further. In contrast, the Neyman-Pearson theory offered a clear rule for acceptance or rejection of a hypothesis without requiring mathematical subtleties like conditional inference or incorporating experimental design. As Lehmann (1993) stressed, the differences between the Neyman-Pearson theory and Fisher's theory condense into the question "What is the relevant frame of reference? It seems clear to me that even in the situations (...), no universal answer is possible. In any specific case, the solution will depend on contextual considerations that cannot easily be captured by a general theory." (Lehmann, 1993, p. 1247). Over time, a hybrid approach evolved out of both theories.

5.3 The Evolution of the modern Hybrid Approach

When Neyman left Great Britain for the United States in 1938, both Fisher's significance testing as well as the Neyman-Pearson theory were fully available and provided a solid theoretical foundation for hypothesis testing. While both approaches differed substantially in 1) their application context, 2) the intended use of the underlying frequentist probability concept and sometimes even in 3) the resulting tests and their results for identical data, the differences were subtle and difficult to grasp for non-specialists at that time. Researchers were soon confronted with the choice between both approaches. As the mathematical level of the content was quite demanding, both concepts were hybridized: Elements of the Neyman-Pearson theory, which appealed with its mathematical optimality properties, made their way into scientific practice and were mixed with Fisher's significance tests. After Neyman left for the United States interested researchers were left on their own to make sense of the existing theories, and in particular, to make sense of what authors of statistical textbooks made of both theories.

Huberty (1993) analysed 28 of the most popular statistical textbooks in the years from 1910 to 1949, and reviewed them in terms of presentations of statistical testing. He investigated the textbook coverage of the p-value (i.e., Fisher) and fixed α level (i.e., Neyman-Pearson) approaches to hypothesis testing. His results show that some of the textbook presentation can be seen as the cause why the hybrid theory evolved.

Halpin and Stam (2006) showed in a detailed analysis, that writers of statistical textbooks at that time merged both approaches into one, ignoring the subtle differences, which led to a hybrid hypothesis testing approach. Fisher (1958b) himself also took note of the inconsistent hybrid approach four years before his death, and argued that it presented a danger for scientific progress.[21] Halpin and Stam (2006) investigated whether both Fisher's and Neyman and Pearson's approach were coexisting in the literature and amalgamated later, or whether Fisher's theory was succeeded by the Neyman-Pearson theory. Their analysis shows that:

> "Technical innovations such as small-sample testing distributions, random assignments of experimental treatments, ANOVA designs (with their corresponding tests of significance), and his attempts to make these advances accessible to the research worker in the form of pedagogical texts led to a wide reception for "Fisher's methods" in various applied sciences (Hotelling, 1951; Yates, 1951; Youlden, 1951; see Lovie, 1979, for a more critical discussion of this reception in psychology). Even Neyman, Fisher's staunchest critic, credited him with founding "the theory of experimentaion" as a domain of study (Neyman, 1967, p. 1456)."
> Halpin and Stam (2006, p. 629)

The history of the hybridization of both theories was documented by Halpin and Stam (2006) through the analysis of two sources: First, they analysed popular statistical textbooks in the years between 1940 and 1960. Second, they analysed the use of statistical hypothesis tests in the journal literature of that time.[22] The textbooks analysed by Halpin and Stam (2006) are Lindquist (1940), Lindquist (1953), Edwards (1950), Edwards (1954), McNemar (1949) and McNemar (1955), which were the most cited statistical textbooks in the years 1940-1960 in research articles in the *Journal of Experimental Psychology* and the *American Journal of Psychology*.[23] The results showed that Lindquist (1940) provided no bibliography or references at all, but presented Fisherian concepts like the null hypothesis and tests of significance (Lindquist, 1940, p. 15-16), but simultaneously elements of the Neyman-Pearson theory like the two types of error (Lindquist, 1940, p. 16-17). Therefore, "already in 1940, the Fisher and Neyman-Pearson approaches to statistical testing were hybridized in textbooks." (Halpin and Stam, 2006, p. 635). In his second textbook *Design and analysis of experiments in psychology and education*, Lindquist (1953) also "provides incomplete interpretations of testing outcomes from both approaches in that it discusses neither Fisher's inductive logic nor

[21] "We are quite in danger of sending highly trained and highly intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be." (Fisher, 1958b)

[22] A limitation of their results is that the journal literature analysis comprises only the *Journal of Experimental Psychology* and the *American Journal of Psychology*. However, they analysed 678 research articles, which provides a relatively detailed picture at least of the use of hypothesis tests in the cognitive sciences.

[23] The analysis of Halpin and Stam (2006) also reveals that George Snedecor's textbook *Statistical Methods* – which can be seen as the accessible version of Fisher's Statistical Methods for Research Workers, compare Chapter 3 – was among the top three of the most cited statistical textbooks together with Fisher's own textbook's and the ones of Lindquist. Crucially, Halpin and Stam (2006) note that "the influence of the Neyman-Pearson theory cannot be discerned in his *Statistical Methods*" (Halpin and Stam, 2006, p. 634), so that Snedecor did not hybridize both theories.

Neyman and Pearson's mathematics of error." (Halpin and Stam, 2006, p. 636). The conclusions for the textbooks of Edwards (1950, 1954) are not much better, as also no citations to the original sources were provided and even worse, the "text propounds a hybridized version of statistical testing in that both the Fisher and Neyman-Pearson approaches are presented under a single model" (Halpin and Stam, 2006, p. 639). The books of McNemar (1949, 1955) show even more precisely how the hybridization has happened: While in the first textbook, McNemar (1949) uses concepts from both approaches without citations, in the second textbook, McNemar (1955) notes in the preface that the text has been revised to include the Neyman-Pearson principles of hypothesis testing. However, as noted by Halpin and Stam (2006, p. 639), McNemar (1955) makes the crucial mistake to interpret the Neyman-Pearson theory as being able to make statements about the truth of a single research hypothesis under consideration (that is, gives it an evidential interpretation in the Fisherian sense):

> "An experiment is carried out which yields sample values, $p1$ and $p2$, and the difference we get between $p1$ and $p2$ is used to test $H0$ against $H1$; that is, on the basis of the obtained difference we are to make a decision as to whether $H0$ or $H1$ is *true*."
> McNemar (1955, p. 61-62)

In the above quote, italics have been added to emphasize the difference to the original interpretation of Neyman and Pearson: No statement about the truth of a single hypothesis can be made at all by the Neyman-Pearson theory, and the only thing which is guaranteed is, that the long-term loss incurred when acting as if H_0 (or H_1) were true (depending on which hypothesis is accepted in each case), is minimised. In the context of hypothesis testing this equals the control of type I and II error rates. Formally, the statement is even false when trying to summarise Fisher's significance tests, as they also do not allow for probabilistic statements about a hypothesis, but only can quantify the plausibility of a hypothesis via the p-value. For a probabilistic statement about a hypothesis, the only option is provided by Bayesian inference.[24] In summary, researchers "were left in the lurch with regard to the interpretation of testing outcomes." (Halpin and Stam, 2006, p. 641). The analysis of Halpin and Stam (2006) shows that the newly introduced elements of the Neyman-Pearson theory like the two types of error, power analysis and statements about the confidence about a procedure were incorporated and integrated into the much simpler theory of Fisher by writers of statistical textbooks in the years between 1940 and 1960.

Switching to the use of hypothesis tests in the journal literature in these years, Figure 5.1 summarizes the analysis of Halpin and Stam (2006): The number of research articles in the 637 articles submitted to the *Journal of Experimental Psychology* and the *American Journal of Psychology* between 1940 and 1960 is shown, and the category non p-value testing refers to older methods of hypothesis testing like critical ratios or probable errors. Although no explicit use of Neyman-Pearson tests can be observed, the data show that "an inference revolution occurred" (Halpin and Stam, 2006, p. 643).[25] The analysis

[24] This shows how strong the natural appeal of researchers was to Bayesian concepts without any formal reference to them, similarly as the Bayesian approach appealed to Neyman when he introduced his confidence intervals.

[25] Halpin and Stam (2006) stress: "It can be seen that the proportion of articles using exact p value methods of statistical testing substantially increased from 1940 to 1960 and that this was accompanied by a decrease in the older methods." (Halpin and Stam, 2006, p. 643).

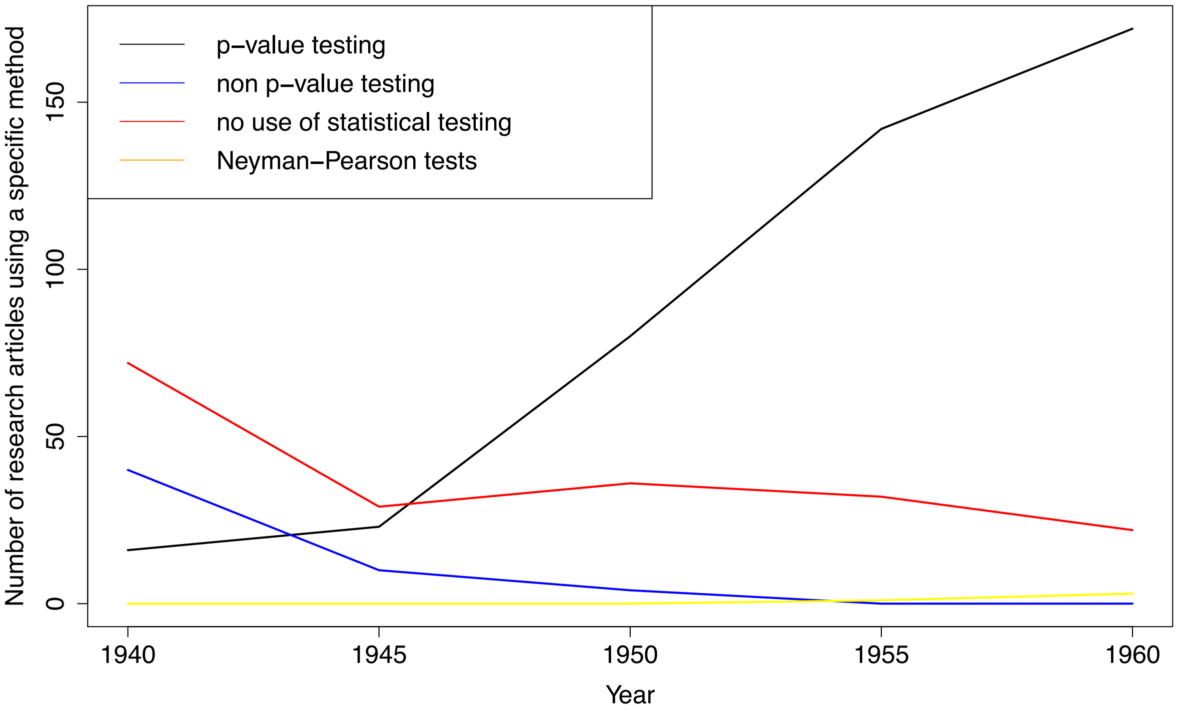

Figure 5.1: Frequencies of different hypothesis testing methods between 1940 and 1960 according to the analysis of Halpin and Stam (2006)

shows that the use of p-values increased strongly. However, it also demonstrates that Neyman-Pearson tests in their original formulation with test levels were rarely if ever used. Instead, the p-value became the preferred solution. The analyses of Hubbard et al. (1997) and Hubbard and Ryan (2000) show that in addition to textbooks, the hybridization also happened in the journal literature with some delay: Hubbard and Ryan (2000) showed that the inference revolution came to a stop in 1955. Still, based on their data, the use of statistical tests in the journal literature increased from 71.7% to 86.1% between 1950-1954 and 1955-1959. In the period from 1960-1964 the use of statistical tests decreased slightly to 84.6%.

However, the increased use of p-values itself does not necessarily imply that both theories were hybridized in the journal literature, too. An answer to this question is provided in the analysis of Hubbard (2004), which reveals that the hybridized version also made its way into the journal literature later. Hubbard (2004) analysed a sample of 1645 papers from 12 psychology journals for the period 1990 through 2002. Hubbard (2004) noted:

> "The confusion arises because researchers mistakenly believe that their interpretation is guided by a single unified theory of statistical inference. But this is not so: classical statistical testing is a nameless amalgamation of the rival and often contradictory approaches developed by Ronald Fisher, on the one hand, and Jerzy Neyman and Egon Pearson, on the other. In particular, there is extensive failure to acknowledge the incompatibility of Fisher's evidential p value with the Type I error rate. (...) The distinction between evidence (p's) and errors (α's) is not trivial."

> Hubbard (2004, p. 1)

While the results of Halpin and Stam (2006) show that the hybridization did not happen in the journal literature between 1940 and 1960, the results of Hubbard (2004) reveal that from 1990 on, the hybridized version was definitely established in the research literature. Though it cannot be specified precisely when the hybridized version entered the journal literature first, the increased use of p-values already in the 1940s and 1950s shows, that hypothesis testing established itself as the new scientific standard for quantifying experimental evidence. Together with the fact that the textbooks hybridized both theories, this is highly problematic. One possible reason for the inference revolution towards Fisher's p-values as the dominating statistical tool according to Halpin and Stam (2006) is:

> "In the research literature reviewed here, the interpretation of testing outcomes (i.e. "statistical significance") was uniformly in terms of experimental evidence, and we did not find any attempts to interpret the general outcomes of an experiment in terms of long-term decision errors."
> Halpin and Stam (2006, p. 644)

Hubbard (2004) has shown that the interpretation of testing procedures in the spirit of Neyman and Pearson is still absent from the journal literature, and according to Halpin and Stam (2006), "after all, researchers are not interested in decision errors but rather in experimental evidence." (Halpin and Stam, 2006, p. 646).

However, as Neyman and Pearson provided the stronger theoretical underpinning of their theory, it appealed to most researchers (Gigerenzer, 2004).[26] Most of Fisher's tests were constructed on intuitive grounds in the first place and later justified as optimal tests when interpreted as UMP level α tests in the Neyman-Pearson theory. Therefore, the vocabulary introduced by Neyman and Pearson like the rejection region, type I and type II error established itself as the new standard over time (Halpin and Stam, 2006; Lenhard, 2006). Gigerenzer (2004) found similar results as Halpin and Stam (2006) when analysing how the hybridisation took place:

> The answer is right there in the first textbooks introducing (...) null hypothesis testing more than 50 years ago. Guilford's *Fundamental Statistics in Psychology and Education*, first published in 1942, was probably the most widely read textbook in the 1940s and 1950s.[27] Guilford suggested that hypothesis testing would reveal the probability that the null hypothesis is true. "If the result comes out one way, the hypothesis is probably correct, if it comes out another way, the hypothesis is probably wrong" (p.156).
> Gigerenzer (2004, p. 596)

Gigerenzer's analysis shows what Halpin and Stam (2006) also found: Researchers are interested in the experimental evidence provided for or against a research hypothesis by

[26] Still, as already shown in Chapter 4, the driving force of the Neyman-Pearson optimality results was Jerzy Neyman, and indeed, "Pearson seems to have distanced himself from his earlier writing (and especially from his association with Neyman's concept of inductive behavior; see Pearson, 1955, p. 207" (Halpin and Stam, 2006, p. 629).

[27] Gigerenzer (2004) provides no arguments for this statement and based on the analysis of Halpin and Stam (2006) the textbooks of Lindquist, McNemar and Edwards were the most widely established ones, at least in psychological research.

the data, which naturally is expressed in terms of probability. Note that Guilford's statement in Gigerenzer's quote above is wrong, that is, probability statements about the research hypothesis of interest (when using Fisher's significance tests or the Neyman-Pearson theory) are not possible. Fisher's significance tests only judge the plausibility of an hypothesis via the likelihood, while the Neyman-Pearson tests can only guarantee error control in expectation and also do not allow for probabilistic statements about the hypothesis. Gigerenzer (2004) concluded:

> "He [Guilford] marked the beginning of a genre of statistical texts that vacillate between the researchers' desire for probabilities of hypotheses and what significance testing can actually provide. For instance, within three pages of text, Nunally (1975, pp. 194-196; italics in the original) used all of the following statements to explain what a significant result such as 5% actually means:
>
> - "the probability that an observed difference is real"
> - "the *improbability* of observed results being due to error"
> - "the *statistical confidence* ... with odds of 95 out of 100 that the observed difference will hold up in investigations"
> - the danger of accepting a statistical result as real when it is actually due only to error
> - the degree to which experimental results are taken "seriously"
> - the degree of "faith [that] can be placed in the reality of the finding"
> - "the investigator can have 95% confidence that the sample mean actually differs from the population mean"
> - "if the probability is low, the null hypothesis is improbable"
> - "all of these are different ways to say the same thing"
>
> Gigerenzer (2004, p. 596-597)

The remarkable aspect of this collection of statements is that first, both elements of Fisher's and the theory of Neyman and Pearson are included, for example statements about the statistical confidence, or about confidence in general. Second, probabilities of hypotheses are used, which do neither appear in Fisher's methodology, nor in the Neyman-Pearson theory.

Over the years, an inconsistent terminology developed: This is observed for example in the existence of an acceptance region for the null hypothesis, although neither Fisher's significance testing nor the Neyman-Pearson theory postulates the existence of such a region. In the latter, the null hypothesis can be accepted, but formally only a rejection region exists, and acceptance of a hypothesis is not to be interpreted literally even in the Neyman-Pearson theory.

The results of Halpin and Stam (2006) and Gigerenzer (2004) show that writers of statistical textbooks specifically began combining both theories by calculating a p-value of the Fisherian significance testing in place of the Neyman-Pearson test statistic (e.g., an LRT statistic, Wald statistic or t-statistic) and testing it against the prespecified fixed Neyman-Pearson α level. The Neyman-Pearson theory thus was primarily only observed at best as a theoretical underpinning which provided explicit error bounds without a clear understanding of its original purpose and limitations (Hubbard, 2004). As often, researchers want to quantify the experimental evidence concerning their research hypothesis, the precise p-value in Fisher's interpretation was then reported. Simultaneously, the p-value was compared to the predetermined Neyman-Pearson test level α, like $p < 0.05$. This makes the impression that a Neyman-Pearson interpretation is followed, in which the result is either in the rejection region or not. However, when interpreting the p-value continuously, e.g. two p-values $p < 0.05$ and $p < 0.01$ are reported and it is argued that the latter one provides more evidence, Fisher's interpretation is taken. In the Neyman-Pearson theory both values are solely in the rejection region and thus interpreted identically. Associating a Fisherian continuous interpretation to p-values implies that the second type of error, which only exists in the Neyman-Pearson theory, vanishes, and the prior sample size and power calculations for n and β of the Neyman-Pearson theory for the prespecified test level α also become invalid as Fisher's significance test do not include these concepts. As a consequence, the long-term optimality guarantees of the Neyman-Pearson theory are lost when two such p-values are interpreted continuously. Problematically, this practice has established itself as the hybridized version, as is also confirmed in the analyses of Huberty (1993), Hubbard et al. (1997), Hubbard and Ryan (2000) and Hubbard (2004).

The hybrid procedure was called the *null ritual* by Gigerenzer (2004) – as often, the null hypothesis, which originally was a well-formulated research hypothesis, degenerates to a "no effect" hypothesis in this hybrid approach – which is shown in the third column of Table 5.1. Importantly, in the hybrid approach which developed as a mixture between both theories in statistical textbooks as shown by Halpin and Stam (2006), the researcher "ends up presenting an exact level of significance as if it were an alpha level, by rounding it up to one of the conventional levels of significance, $p < 0.05$, $p < 0.01$, or $p < 0.001$. The result is not alpha, nor an exact level of significance." (Gigerenzer, 2004, p. 594). Termed differently, the hybrid approach tries to combine both Fisher's evidential interpretation of p-values with the mathematical and decision-theoretic optimality properties of the Neyman-Pearson theory, both of which are mutually exclusive.

Thus, in the hybrid approach, researchers started to use the prespecified α level of the Neyman-Pearson theory and then used Fisher's p-values to quantify the strength against the null hypothesis H_0 continuously. At the same time, researchers thought that the long-run objectivity of the Neyman-Pearson theory would still hold, see also Loftus (1991), by rounding to the next critical value like $p < 0.05$, which is interpreted as "We are in the rejection region of the Neyman-Pearson test for a test level $\alpha = 0.05$". However, in the Neyman-Pearson theory, p-values are not defined at all, so no difference is made between the location of points in the rejection region. Therefore, in the Neyman-Pearson theory continuous interpretation of p-values as the strength against the null hypothesis H_0 is not allowed. Instead, only acceptance or rejection of H_0 based on the binary interpretation of p-values is possible. Also, even highly significant research results with tiny p-values have to be interpreted in the same way as results with moderate p-values. This renders the Neyman-Pearson theory inappropriate for scientific research[28], and the ASA also favoured a position resembling Fisher's continuous interpretation of p-values in principles one and three in their 2016 statement:

> '1. p-values can indicate how incompatible the data are with a specified statistical model. (...)

[28] See also the *Principle of Adequacy* in Part IV in Chapter 11.

> 3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.'
> Wasserstein and Lazar (2016, p. 2)

On the other hand, the ASA statement also stressed that p-values alone are problematic, questioning the usefulness of Fisher's interpretation for scientific research, too:

> '5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. (...)
> 6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.'
> Wasserstein and Lazar (2016, p. 2)

This situation becomes even worse, as the majority of studies and experiments in science is not replicated, so that the possibility of establishing scientific facts, in general, is questionable from both Fisher's and Neyman's and Pearson's perspective. When making use of the hybrid theory, which incorrectly interprets p-values continuously when using the Neyman-Pearson UMP level α tests, researchers are tempted to continuously quantify their results via p-values, although this is simply not allowed. The hybrid theory which evolved out of both theories still troubles researchers today, as recently shown by Cassidy et al. (2019) in the context of psychological research: "We examined 30 introductory-psychology textbooks, including the best-selling books from the United States and Canada, and found that 89% incorrectly defined or explained statistical significance" (Cassidy et al., 2019, p. 1). In light of this undesirable situation, the ASA noted that there are also alternatives to these theories, including approaches which

> "...emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors;"
> Wasserstein and Lazar (2016, p. 2)

and which hopefully would lead to a "more nuanced approach to interpreting, communicating, and using the results of statistical methods in research." Wasserstein and Lazar (2016, p. 2). Therefore, the next chapter analyses the evolution of one of the most promising alternatives noted in the ASA statement, namely Bayesian inference.

INTERMEDIATE CONSIDERATIONS

Part I showed that current dominating statistical methodology emerged out of both Fisher's theory of significance tests which advocated p-values, and the Neyman-Pearson theory of statistical hypothesis testing. The result is an inconsistent hybrid approach and the reconstruction showed why the current status quo of NHST is highly problematic in scientific research. Also, Part I clarified that the dominant approach to statistical hypothesis testing today was never intended by the creators of the underlying statistical theories in such an application context.

In the following Part II, the evolution of Bayesian approaches to hypothesis testing is detailed with a focus on the Bayes factor. Chapter 6 outlines the basics of Bayesian statistics and contrasts them with the frequentist approach, and Chapter 7 analyses the evolution of the Bayes factor as an alternative to frequentist hypothesis tests based on p-values. It is shown that although the Bayes factor approach did not succeed in the decades that followed, the primary reasons were mostly computational hurdles which prevented a more widespread use of Bayesian methods in scientific research. Also, the core differences between the frequentist and Bayesian approach are analyzed and it is shown that the more appropriate approach for hypothesis testing in scientific contexts is a Bayesian one.

Part II

The Evolution of Bayesian Hypothesis Testing

Chapter 6

Bayesian Statistics

> No probability is ever determined from experience alone. It is always influenced to some extent by the knowledge we had before the experience.
>
> Dorothy Wrinch & Harold Jeffreys
> *On certain fundamental principles of scientific inquiry*

This chapter introduces the fundamental concepts and definitions in Bayesian statistics, which are necessary to test hypotheses in the Bayesian approach.

6.1 Elements of Bayesian Statistics

In frequentist inference, the data X are random, and the parameter of interest θ is regarded to be fixed, and point and interval estimates are functions of this data X.[1] Examples include the sample mean or confidence intervals as introduced by Neyman (1937). In the Bayesian approach, things are opposite. Here, the observed data X are treated as fixed (they have been observed) and the parameter θ is a random variable. Measure-theoretic details are provided in Appendix C. Of course, there are multiple philosophical aspects which provide support or criticism for the frequentist or Bayesian approach, but these are discussed in Part IV.

While frequentist inference includes the theory of point estimators, confidence intervals, (approximative) pivots and significance tests, Bayesian inference derives an entire posterior distribution for the parameter θ of interest. By incorporating both the observed data $X = x$ as well as the available prior information about the parameter of interest, Bayes' theorem is used to obtain the posterior distribution of θ, given the data $X = x$.

In Bayesian inference, the prior distribution $p(\theta)$ of the parameter θ is interpreted as the uncertainty about the actual value of θ before conducting a study or experiment, that is, before data $X = x$ are observed. The shape of the prior distribution $p(\theta)$ models the available prior knowledge: Compact, narrow priors can express a considerable certainty about θ, while wide, noninformative priors can express wide ignorance of any

[1] Here, we loosely speak of random "data" X and fixed "parameter" θ, which is of course to be interpreted in the measure-theoretic sense outlined in Appendix C.

knowledge about θ. Still, one of the oldest critiques is the subjectiveness involved in the choice of a suitable prior when conducting Bayesian inference (Howie, 2002).

In the Bayesian approach, the likelihood function $f(x|\theta)$ expresses the plausibility for obtaining the data x given a specific value of θ. Thus, in total Bayesian inference updates the prior information by multiplying $p(\theta)$ with the likelihood $f(x|\theta)$, obtaining the updated posterior distribution $p(\theta|x)$.

$$\textit{Prior} \times \textit{Likelihood} \propto \textit{Posterior}$$

In the above, the sign $\propto$ means "proportional to", and in general, as likelihood is no probability density, the posterior is only proportional to the product of the likelihood and prior up to a normalization constant. The basis of this approach is Bayes' theorem (Held and Sabanés Bové, 2014, p. 168)[2]:

Theorem 6.1 (Bayes)**.** Let A and B denote two events on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ with $0 < \mathbb{P}(A) < 1$ and $\mathbb{P}(B) > 0$. Then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \qquad = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B|A) \cdot \mathbb{P}(A) + \mathbb{P}(B|A^c) \cdot \mathbb{P}(A^c)} \tag{6.1}$$

and for a general partition $A_1, A_2, ..., A_n$ of Ω with $\mathbb{P}(A_i) > 0$ for all $i = 1, ..., n$, we have that

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^n \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)} \tag{6.2}$$

for each $j = 1, ..., n$.

Bayes' theorem is a trivial consequence of the definition of conditional probability: Equation (6.1) follows from $\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A)/\mathbb{P}(B)$. The idea in Bayesian inference is to assume a prior distribution μ_ϑ with density $p(\theta)$ for the unknown parameter θ, which is a random variable. The prior $p(\theta)$ is updated by the information the data provide through $f(x|\theta)$, by means of Bayes' theorem into the posterior $p(\theta|x)$. In Bayesian inference, the posterior $p(\theta|x)$ is the central quantity of interest. As it contains all information about the parameter θ after having observed the data $X = x$, it is used to derive Bayesian point and interval estimates (Held and Sabanés Bové, 2014, p. 170).

Definition 6.2 (Posterior distribution)**.** Let $X = x$ denote the observed realisation of a (possibly multivariate) random variable X with density function $f(x|\theta)$. Let μ_ϑ a prior distribution with density function $p(\theta)$. The density function $p(\theta|x)$ of the posterior distribution $\mu_{\vartheta|x}$ is defined as

$$p(\theta|x) := \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} \tag{6.3}$$

For a discrete parameter θ, the integral in the denominator is replaced with a sum and the densities with probability mass functions. Note that while Bayes' theorem operates on a single probability space $(\Omega, \mathcal{A}, \mathbb{P})$, the posterior distribution is derived by combining the likelihood and prior, which operate on different spaces: The likelihood

[2] For a measure-theoretic formulation of Bayes' theorem see Appendix C.

is defined on the sample space $(\mathcal{X}, \mathcal{A})$, which is a measure space, compare Appendix C. The prior distribution operates on the parameter space, which is a probability space $(\Theta, \tau, \mu_\Theta)$, where μ_Θ is the corresponding probability measure associated with the density p for θ. The parameter θ is thus the realisation of a random variable $\vartheta : \Omega \to \Theta$, and Bayes theorem' is used to derive the conditional distribution $\mu_{\vartheta|X} : \tau \times \mathcal{A} \to [0,1]$ for $\vartheta|X$ – which is called the posterior distribution. After conditioning on the observed data X, the posterior $\mu_{\vartheta|X}$ is thus a regular conditional distribution (**?**), which means that $\mu_{\vartheta|X} : x \mapsto \mu_{\vartheta|X}(B, x)$ is B-measurable for all $B \in \tau$ and $B \mapsto \mu_{\vartheta|X}(B, x)$ is a probability measure on τ for almost all $x \in \mathcal{X}$, where almost all refers to the prior predictive distribution of the data X (see below and compare Appendix C). The denominator of the posterior in Equation (6.3) can be written as

$$\int f(x|\theta)p(\theta)d\theta = \int \tilde{f}(x,\theta)d\theta = f(x) \tag{6.4}$$

and does not depend on θ.[3] Equation (6.4) shows that the denominator in the posterior distribution is independent of θ. Therefore, the posterior distribution is proportional to the product of the likelihood and prior distribution (Held and Sabanés Bové, 2014, p. 170). This is usually written as

$$f(\theta|x) \propto f(x|\theta) \cdot p(\theta) \tag{6.5}$$

where '$\propto$' means 'proportional to' and the proportionality constant in this case is $1/f(x)$. The normalizing constant

$$\frac{1}{f(x)} = \frac{1}{\int f(x|\theta)p(\theta)d\theta} \overset{\text{Definition C.35}}{=} \frac{1}{\int L(\theta)p(\theta)d\theta} \tag{6.6}$$

ensures that the product of likelihood and prior distribution indeed yields a probability density again, which integrates to unity. For a simple example and minimal interpretation see Appendix C.

6.2 Bayesian Point and Interval Estimates

In Bayesian inference, point and interval estimates of θ are obtained from the posterior distribution (Held and Sabanés Bové, 2014, p. 171)[4]:

Definition 6.3 (Posterior mean). The *posterior mean* $\mathbb{E}[\theta|x]$ is the expectation of the posterior distribution:

$$\mathbb{E}[\theta|x] = \int \theta f(\theta|x)d\theta \tag{6.7}$$

Next to Bayesian point estimates like the posterior mean, the analogue to confidence sets and intervals as introduced in Appendix C.8 are *credible intervals* (Held and Sabanés Bové, 2014, p. 172):

[3] In the above, $\tilde{f}$ denotes the product-density on the product-space $(\mathcal{X} \times \Theta, \mathcal{A} \times \tau)$.

[4] For a decision-theoretic justification as detailed in Appendix C of Bayesian point-estimates see Rüschendorf (2014, Chapter 2) and Robert (2007, Chapter 2).

Definition 6.4 (Credible interval). For fixed $\gamma \in (0,1)$, a $\gamma \cdot 100\%$ credible interval is defined by two real numbers t_l and t_u, which fulfill

$$\int_{t_l}^{t_u} f(\theta|x)d\theta = \gamma \tag{6.8}$$

The quantity γ is called the *credible level* of the credible interval $[t_l, t_u]$.

In contrast to the complex and easily misunderstood interpretation of the confidence interval – see Definition C.86 and Chapter 5 – this definition indeed implies that the random variable θ is contained in a $\gamma \cdot 100\%$ credible interval with probability γ, given the data $X = x$. The easiest way to construct a credible interval is to use the $(1-\gamma)/2$-quantile of the posterior distribution as t_l and the $(1+\gamma)/2$-quantile of the posterior distribution as t_u. The posterior distribution, however, does not need to be symmetric, therefore often credible intervals are used that include the parameter values with the highest posterior density (Held and Sabanés Bové, 2014, p. 177):

Definition 6.5 (Highest Posterior Density Interval (HPD)). Let $\gamma \in (0,1)$ be a fixed credible level. A $\gamma \cdot 100\%$ credible interval $I = [t_l, t_u]$ is called a *highest posterior density interval* if

$$f(\theta|x) \geq f(\tilde{\theta}|x) \tag{6.9}$$

for all $\theta \in I$ and $\tilde{\theta} \notin I$.

For discrete parameter spaces Θ, one drawback of the definition of a credible interval is that it has to be modified, as it may be impossible to obtain an exact credible level γ due to the discreteness of Θ. A $\gamma \cdot 100\%$ credible interval $I = [t_l, t_u]$ for θ is then defined as

$$\sum_{\theta \in I \cap \Theta} f(\theta|x) \geq \gamma \tag{6.10}$$

When no information is available, a uniform prior, as already used by Bayes and Price (1763), can be used:

Definition 6.6 (Uniform prior). A uniform prior is a prior which is distributed uniformly on the parameter space.

The following result which is immediate from Equation (6.5) gives another connection between Bayesian and frequentist inference and is due to the occurrence of the likelihood function in the Bayesian approach:

Theorem 6.7. Under a uniform prior, the posterior mode equals the maximum likelihood estimator (MLE).

While Bayes' theorem offers a simple way to obtain the posterior distribution, the computation of the normalising constant in the denominator makes the computation effortful if not impossible in practice. The earliest approaches to obtain the posterior distribution were – as there were no computing resources available – limited to the investigation of *conjugate priors* (Held and Sabanés Bové, 2014, p. 180). Conjugate priors have the appealing property, that one precisely knows the form of the resulting posterior after selecting the likelihood and choosing a conjugate prior.

Definition 6.8 (Conjugate prior). Let $f(x|\theta)$ denote a likelihood function based on the observation $X = x$. A class $\mathcal{G}$ of distributions is called *conjugate with respect to* $L(\theta)$ if the posterior distribution $f(\theta|x)$ is in $\mathcal{G}$ for all x whenever the prior distribution $f(\theta)$ is in $\mathcal{G}$.

When a conjugate prior is chosen for a likelihood, the posterior is again distributed as the conjugate prior, with updated parameters. This allows for a plug-in procedure: 1) Choose the likelihood, 2) choose a conjugate prior, 3) obtain the resulting posterior by updating the parameters, where the updating often involves using the data and the prior parameters to obtain the posterior parameters, and 4) compute point or interval estimates based on the posterior with updated parameters. For multiple widely used distributions, there is good knowledge about the conjugate classes: Combining a beta prior with a binomial likelihood results in another beta posterior with different parameters. Combining a Poisson prior with a Poisson likelihood results in another Poisson posterior with different parameters.

In most realistic cases where multiple parameters are involved, or the likelihood function has a complicated form, there is no conjugate prior available. This situation remains the status quo and is the strongest motivation for the use of Markov-Chain-Monte-Carlo algorithms (Brooks, 2011; Kruschke, 2015) which are discussed in Part III.

If there is little or even no information available, a vague prior with large scale parameter is often selected. However, if the scale parameter is chosen too large, such a prior may degenerate until it becomes completely "flat". While this seems beneficial and highly objective, a flat prior like $p(\theta) = 1$ does not integrate to unity anymore. As a consequence, such priors violate the definition of a probability density and are called *improper* (Held and Sabanés Bové, 2014, p. 184).

Definition 6.9 (Improper prior). A prior distribution with density function $f(\theta) \geq 0$ is called improper if

$$\int_{\Theta} f(\theta)d\theta = \infty \qquad \text{or} \qquad \sum_{\theta\in\Theta} f(\theta)d\theta = \infty \tag{6.11}$$

for continuous or discrete parameters θ, respectively.

When using an improper prior, it is necessary to check that at least the posterior is a probability density. Otherwise, any information obtained from the posterior is not valid if the posterior has been obtained from an improper prior. Next to the fact that improper priors are no probability densities anymore, there is another problem. One can show when using the naive choice for a improper prior $f_\theta(\theta) = 1$, that transforming the parameter θ to a parameter $\phi = h(\theta)$, where h is a one-to-one differentiable transformation of θ, may lead to a non-constant prior for ϕ. As invariance under reparameterisations is desirable, often *Jeffreys' prior* is used, which is named after Sir Harold Jeffreys (1891-1989) (Held and Sabanés Bové, 2014, p. 186):

Definition 6.10 (Jeffreys' prior). Let X be a random variable with likelihood function $f(x|\theta)$ where θ is an unknown scalar parameter. *Jeffreys' prior* is defined as $p(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher-Information as given in Definition C.46.

The use of the Fisher-Information in Jeffreys' prior is another connection between frequentist and Bayesian inference. Additionally, Jeffreys' prior enjoys exactly the desired property of invariance under reparameterizations (Held and Sabanés Bové, 2014, p. 186).

In frequentist inference, point estimates are mainly judged by their variance, compare Appendix C.5. Results like the Cramér-Rao-Lower-Bound (Appendix C, Theorem C.44) or the Rao-Blackwell-Theorem (Appendix C, Theorem C.55) assist in reducing the variance of an estimator and judging its optimality. Another option is to show decision-theoretic optimality of an estimator. In Bayesian inference, a decision-theoretic approach can be used, too, and is based on different decision rules and loss functions, see Appendix C. For a decision-theoretic justification as of common Bayesian point-estimates like the posterior mean, median or mode see Rüschendorf (2014, Chapter 2).

Common Bayesian interval estimates can also be derived based on optimality with respect to specific loss function (Held and Sabanés Bové, 2014, Chapter 6). One important aspect is that in the Bayesian approach, not all point estimates necessarily lie within the obtained interval estimates. Specifically, the posterior mean may not lie within the HPD if the posterior distribution is skewed. Also, due to Jensen's inequality (Held and Sabanés Bové, 2014, Appendix 3.7), Bayes estimates are generally not invariant under one-to-one transformations, unlike the MLE.

An important question is whether Bayesian point estimates are consistent in the classical sense, that is if they converge to the true parameter value if the sample size increases. For discrete asymptotics, this is not possible, but one can show that the probability mass gets more and more concentrated around the true value for increasing sample size. In the case of continuous asymptotics, it is possible to show that under the Fisher-Regularity conditions – see van der Vaart (1998, Chapter 10) – the posterior distribution is asymptotically normal if the prior is not degenerate (Held and Sabanés Bové, 2014, Chapter 6). This fact again parallels the asymptotic normal distribution of the MLE – Theorem C.49 – and allows for a Bayesian interpretation of the MLE and its standard error. For details, see Held and Sabanés Bové (2014, Section 6.6) and the Bernstein-von-Mises theorem (van der Vaart, 1998, Chapter 10).

6.3 Bayesian Hypothesis Testing

Testing a hypothesis in the Bayesian approach is formulated as a model selection problem. The oldest and most widely established method for Bayesian hypothesis tests is the Bayes' factor (Jeffreys, 1931, 1939, 1935, 1948, 1961; Kass and Raftery, 1995; Held and Sabanés Bové, 2014; Ly et al., 2016b; Held and Ott, 2018). The two competing hypothesis H_0 and H_1 are considered as models in the Bayesian approach. From this point of view, these two models M_0 and M_1 can be assigned prior probabilities $\mathbb{P}(M_0)$ and $\mathbb{P}(M_1)$, where of course $\mathbb{P}(M_0) + \mathbb{P}(M_1) = 1$ has to hold. The hypothesis testing procedure then can be translated into the calculation of the posterior model probabilities $\mathbb{P}(M_0|x)$ and $\mathbb{P}(M_1|x)$ after observing the data x. By Bayes' theorem, the posterior odds $\mathbb{P}(M_0|x)/\mathbb{P}(M_1|x)$ can be expressed as[5]:

$$\underbrace{\frac{\mathbb{P}(M_0|x)}{\mathbb{P}(M_1|x)}}_{\text{posterior odds}} = \underbrace{\frac{f(x|M_0)}{f(x|M_1)}}_{\text{Bayes factor } BF_{01}} \cdot \underbrace{\frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}}_{\text{prior odds}} \tag{6.12}$$

This leads to the definition of the Bayes factor (Held and Sabanés Bové, 2014, p. 233), which was invented by Jeffreys (1935):

[5] This follows from applying Bayes' theorem twice as $\mathbb{P}(M_0|x) = f(x|M_0)\mathbb{P}(M_0)/f(x)$ and $\mathbb{P}(M_1|x) = f(x|M_1)\mathbb{P}(M_1)/f(x)$. The ratio $\mathbb{P}(M_0|x)/\mathbb{P}(M_1|x)$ is then given as Equation 6.12.

Definition 6.11 (Bayes factor). The *Bayes factor* BF_{01} is given by the ratio of the marginal likelihoods of the two models M_0 and M_1, that is,

$$\mathrm{BF}_{01}(x) = \frac{f(x|M_0)}{f(x|M_1)} \tag{6.13}$$

When both models M_0 and M_1 are completely specified (contain no unknown parameters), the Bayes factor equals the likelihood ratio. If there are unknown parameters $\boldsymbol{\theta}$ (which may be vector-valued), the computations become more involved as these parameters need to be marginalized out, leading to the prior predictive distribution and marginal likelihood (Held and Sabanés Bové, 2014, p. 232):

Definition 6.12 (Prior predictive distribution). The *prior predictive distribution* for the model $M_i, i = 1, 2$ is has the density

$$f(x|M_i) = \int f(x|\boldsymbol{\theta}_i, M_i) f(\boldsymbol{\theta}_i|M_i) d\boldsymbol{\theta}_i, \quad i = 1, 2 \tag{6.14}$$

where $\boldsymbol{\theta}_i$ is the unknown parameter vector for the model M_i.

Definition 6.13 (Marginal Likelihood). The *marginal likelihood* is the value of the prior predictive distribution $f(x|M_i)$ evaluated at x.

Bayesian hypothesis testing, therefore, is most often interpreted as a model selection problem of two models M_0 and M_1 corresponding to the hypotheses H_0 and H_1. The Bayes factor quantifies the change in belief from the prior odds towards either of both hypotheses. The evidence concerning H_0 and H_1 can also be obtained by calculating the *maximum a posteriori* (*MAP*) model with the largest posterior model probability. Still, the posterior model probabilities form the posterior model odds, and these odds depend strongly on the assumed prior odds. The Bayes factor is influenced only by the data x (and the prior distributions on the parameters, in case any unknown parameters need to be marginalised out for calculating the marginal likelihoods). Therefore, the Bayes factor is often preferred in practice as it quantifies the evidence in the data x regarding H_0 and H_1, no matter how the prior odds were selected, see Robert (2007) and **?**, Lemma 2.6.

Example 6.14. A simple example demonstrates the application of the Bayes factor. It contrasts it to the frequentist approach using significance levels or p-values: Let a random variable simulate the outcome of a medical disease test which produces either a 'yes' or 'no' for patients with the disease. Data are assumed to be binomially distributed with parameters n and p, and the comparison is made for

1. a Model M_0 where the probability p is assumed to be $p := 0.5$ for correctly producing a 'yes' for a patient with the disease
2. Model M_1, where p is unknown (and a uniform prior $p(\theta) = \frac{1}{1-0} = 1$ is assumed for p)

Assume a sample of $n = 200$ patients was taken and the diagnostic tests yield 115 positive and 85 negative results. The likelihood is a binomial distribution

$$f(X = 115|M_0) = \binom{200}{115} p^{115}(1-p)^{85} = \binom{200}{115} 0.5^{200} = 0.005956 \tag{6.15}$$

Table 6.1: Categorization of Bayes factors $BF_{01} \leq 1$ into evidence against H_0

Bayes factor BF_{01}	Strength of evidence against H_0			
	Jeffreys (1961)	Goodman (1999)	Held and Ott (2016)	Lee and Wagenmakers (2013)
1 to 1/3	Bare mention		Weak	Anecdotal
1/3 to 1/10	Substantial	Weak to moderate	Moderate	Moderate
1/10 to 1/30	Strong	Moderate to strong	Substantial	Strong
1/30 to 1/100	Very strong	Strong	Strong	Very strong
1/100 to 1/300	Decisive	Very strong	Very strong	Extreme
< 1/300			Decisive	

Note: Jeffreys (1961) actually used the cut points $(1/\sqrt{10})^a$ with $a = 1, 2, 3, 4$, and Goodman (1999) used the cut points 1/5, 1/10, 1/20 and 1/100 for “weak”, “moderate”, “moderate to strong” and “strong to very strong”, which have been aligned with the cut points in the left column to make comparison easier.

and

$$f(X = 115|M_1) = \int_0^1 f(x|\theta) \cdot p(\theta) dp \tag{6.16}$$

$$= \int_0^1 \binom{200}{115} p^{115}(1-p)^{85} dp = \binom{200}{115} \int_0^1 p^{115}(1-p)^{85} dp \tag{6.17}$$

$$\overset{(1)}{=} \binom{200}{115} \cdot \frac{\Gamma(116)\Gamma(86)}{\Gamma(116+86)} = \frac{1}{201} = 0.004975 \tag{6.18}$$

where (1) follows from the fact that a beta distribution has density $f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$ where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and thus

$$\int_0^1 p^{115}(1-p)^{85} dp = \int_0^1 f_{116,86}(p) \cdot B(116, 86) dp = B(116, 86) \underbrace{\int_0^1 f_{116,86}(p) dp}_{=1} \tag{6.19}$$

$$= \frac{\Gamma(116)\Gamma(86)}{\Gamma(116+86)} \tag{6.20}$$

Therefore, according to Definition 6.11, the Bayes factor results in

$$\mathrm{BF}_{01}(x) = \frac{0.005956}{0.004975} = 1.1971 \tag{6.21}$$

Multiple thresholds for interpretation of the Bayes factor were proposed over time (Jeffreys, 1961; Kass and Raftery, 1995; Lee and Wagenmakers, 2013; Goodman, 1999; van Doorn et al., 2021).

Table 6.1 provides the strength of evidence against H_0 of the Bayes factor $BF_{01} = 1.1971$. In this case, $BF_{01} \geq 1$ and therefore the scale is changed: The evidence for H_0 is given by the steps $1 \leq BF_{01} < 3$, $3 \leq BF_{01} < 10$, $10 \leq BF_{01} < 30$, $30 \leq BF_{01} < 100$ and $100 \leq BF_{01} < 300$ and $BF_{01} \geq 300$. Using this scale, the evidence for H_0 is barely worth mentioning according to Jeffreys (1961), weak according to Held and Ott (2018) and anecdotal according to Lee and Wagenmakers (2013), and M_0 is therefore not confirmed by the data x. In summary, the evidence in the data x is not substantial enough to confirm H_0, and it also does not advise against H_0.

While the Bayesian approach does not treat the data as convincing enough to accept M_1 or reject it based on the results, a classical frequentist analysis would have yielded different results. Let $\alpha = .05$ be the test level and $H_0 : p = \frac{1}{2}$ and $H_1 : p \neq \frac{1}{2}$. The probability of getting a figure as extreme as 115 or more extreme, that is $\mathbb{P}(X \geq 115|H_0) + \mathbb{P}(X \leq 85|H_0)$ (we perform a two-sided test), is 0.04, which is smaller than $\alpha = .05$. In total, frequentist inference via a p-value (no matter if in the Neyman-Pearson or Fisherian interpretation) for the usual threshold 0.05 (in Fisher's words, significance level, in Neyman-Pearson language, test level) would reject the hypothesis $H_0 : p = \frac{1}{2}$ belonging to M_0, while Bayesian inference via the Bayes factor would not.

Note that changing the prior model probabilities $\mathbb{P}(M_0)$ and $\mathbb{P}(M_1)$ does not influence the Bayes factor. In contrast, when using a different prior distribution on the model parameter p in M_1, the marginal likelihood $f(x|M_1)$ changes and therefore also the resulting Bayes factor changes, too: Under the uniform prior $f(p|M_1) \propto 1$, the marginal likelihood is given as $f(x|M_1) = \int f(x|p, M_1)f(p|M_1)dp = \int f(x|p, M_1)dp$, while under any nonuniform prior, the marginal likelihood results in

$$f(x|M_1) = \int f(x|p, M_1)f(p|M_1)dp$$

As the Bayes factor $BF_{01} = f(p|M_1)/f(p|M_0)$ is the ratio of the two marginal likelihoods under M_0 and M_1, it is influenced by the prior selection.

Chapter 7

The Evolution of the Bayes Factor

An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure.

Harold Jeffreys
Theory of Probability

This chapter outlines the evolution of the Bayes factor, which is the central quantity for Bayesian hypothesis testing. Historically, the invention of the Bayes factor for testing statistical hypotheses in a coherent Bayesian framework was first introduced by Harold Jeffreys and Dorothy Wrinch. Their methodology was invented as a statistical theory for use in *scientific* applications. As a consequence, the context of application and the underlying probability concept of Jeffreys' theory are in sharp contrast to the assumptions Fisher and Neyman and Pearson made. Also, the contributions of J.B.S. Haldane to the development of the Bayes factor will be analysed in this chapter. The influence of Haldane on the development is also discussed by Etz and Wagenmakers (2017). Here, the difference in analysis and interpretation of observed data between Jeffreys and Haldane will be focussed, and it is shown that the context of application is substantial to explain these differences.

Jeffrey's life and achievements have been analysed in multiple publications, including (Aldrich, 2006) and (Howie, 2002). Jeffreys monograph *'Theory of Probability'* (ToP) (Jeffreys, 1931, 1948, 1961) remains a foundational work of Bayesian statistics until today, although the notation has become apocryphal from a modern perspective. The achievements in and impact of ToP have been analysed by Robert et al. (2008) and Ly et al. (2016b,a). While Robert et al. (2008) analyses the monograph in full length, Ly et al. (2016a) and Etz and Wagenmakers (2017) are more focussed on the Bayes factor and Jeffrey's (respectively J.B.S. Haldane's) influence on architecting a Bayesian methodology for hypothesis testing. Up to date, only Etz and Wagenmakers (2017) have analysed the influence of J.B.S. Haldane, a Cambridge fellow of Harold Jeffreys, on the invention of the original Bayes factor. They suggest that Haldane may have had more influence on the development of the Bayes factor than previously thought, and the difference between Jeffrey's and Haldane's approach plays an especially important role for the later derivations in Part V.

The emphasis of this chapter is therefore on the timespan when the Bayes factor was first introduced, including the rivalling approach of J.B.S. Haldane. Moreover, Fisher's dissent to Jeffreys' inverse probability approach is discussed, see also Aldrich (2008), Zabell (1989a), Zellner (1980) and Howie (2002). Historically, what today is called Bayesian statistics or Bayesian inference, was at that time called *inverse probability*. In his introduction to the first edition of *Statistical Methods for Research Workers*, Fisher (1925a) wrote: "For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences are usually distinguished under the heading of *Inverse Probability*, and have at times gained wide acceptance." (Fisher, 1925a, p. 10). The phrasing inverse can thus be attributed to the procedure that in the Bayesian approach, instead of deducing inferences about the parameter from the sample directly, the inverse path is taken and a prior distribution is first specified about the unknown parameter, before any deduction of inferences takes place. Another reason for the name inverse probability can be understood by considering a simple urn model: The proportion of coloured balls in a sample taken with replacement was interpreted as caused by the ratio of colours in the urn. The strong law of large numbers guaranteed that, in the long run, such causes would manifest themselves through chance fluctuations when taking a large sample. Consequently, after observing a sample, one could calculate the ratio of coloured balls in the urn. Interpreting the sample as the observed effect, and the ratio as the unknown causes, "Inverse probability could be used to infer unknown causes from known effects." (Howie, 2002, p. 23).

7.1 Wrinch and Jeffreys's Invention of the Bayes Factor

The invention of the Bayes factor as it is used today is most often attributed to Sir Harold Jeffreys. For a detailed account of Jeffreys' personal life see Howie (2002, Chapter 4) and Aldrich (2006). Jeffreys' early work on Bayes factors is also described in Etz and Wagenmakers (2017) and Howie (2002, p. 85-92). Together with his colleague Dorothy Wrinch, Jeffreys wrote four papers which constituted the centrepiece of what was later introduced in ToP. Jeffreys (1980) later recalled the starting point of their work, which was inspired by the results of Broad (1918), see also Etz and Wagenmakers (2017, p. 11). In the early 20th century, the influence of Laplace's principle of insufficient reason was considerable, as Howie (2002) notes, and this principle built the starting point of the work of Broad (1918) years earlier:

> "Accustomed to the tradition in classical probability, in which problems were set up in terms of equally likely cases, Laplace tended to assume what became known as the Principle of Insufficient Reason – that where we have no knowledge of causes, or reason to favor one hypothesis over another, we should assign each the same probability. Thus in his study of comets, Laplace took all values of the orbit's perihelion – its nearest point to the Sun – to be initially equally likely."
> Howie (2002, p. 31)

Laplace's principle can be identified with an improper, "flat" prior distribution over the parameter space. Broad (1918) showed in 1918 using Laplace's principle of insufficient reason that when uniform prior probabilities are used in finite populations, the posterior probability of a general law paradoxically never achieves to come close to unity unless the whole population is sampled. A general law is a hypothesis like all crows are black, or all apple trees bear apples, or more generally, a hypothesis with the parameter $\theta = 1$ or $\theta = 0$.[1]

> "Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property φ, and we do not know r, the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that (...) if we take a sample of number m and find all of them with φ, the posterior probability that all n are φ's is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been sampled. We could never be reasonably sure that apple trees would always bear apples (...). The result is preposterous, and started the work of Wrinch and myself in 1919-1923."
> Jeffreys (1980, p. 452)

A detailed mathematical derivation of this result is given in Zabell (1989b, p. 309/310). Zabell (1989b) notes that an "important consequence of Broad's analysis was (...) a serious setback to the Laplacean program of justifying induction probabilistically, and was an important impetus for the early work of Jeffreys and Wrinch" (Zabell, 1989b, p. 286). The problem appears when only a fraction of m of the whole population is sampled. Indeed, when using a uniform prior, the entire population needs to be sampled to yield a posterior probability of one, if the true parameter value is $\theta = 1$, or a posterior probability of zero, if the true parameter is $\theta = 0$. That means a general law that all n population members have the property φ, like apple trees are bearing apples, can never be established until the entire population is sampled. However, when the whole population is sampled, the use of statistical inference becomes superfluous. Note that about the same time, Fisher was beginning to object to inverse probability exactly because of this problem. As Howie (2002, p. 64) observed, "it was necessary for Fisher to refute the rival of inverse probability. (...) His chief objection was to the arbitrariness and inconsistency of the Principle of Insufficient Reason."

On the other side, Wrinch and Jeffreys (1919) had a problem with the conventional frequency definition of probability by Venn or Fisher because "a statement that it is probable that the solar system was formed by the disruptive approach of a star larger than the sun" (Wrinch and Jeffreys, 1919, p. 716) of course made no sense when considering probability as a limiting ratio of events performed in indefinite repetition in the Neyman-Pearson interpretation or in Fisher's hypothetical infinite population interpretation. Wrinch and Jeffreys (1919) themselves did not define probability at all. What is more, they even stressed that probability has no definition and maybe is not even definable. For Jeffreys and Wrinch, the goal was to create a theory for quantifying the evidence about scientific hypotheses. Note the contrast to the Neyman-Pearson theory, which did not include any notion of evidence and was aimed at quality control situations. Similarly, Fisher's significance tests were inspired by his agricultural experimental work at Rothamsted and his collaboration with Sealy Gosset at the Guinness brewery. However, Fisher's theory of significance testing at least was aimed at quantifying evidence about a research hypothesis and included a concept of evidence in contrast to the Neyman-Pearson theory. However, the probability concept was, of course, a frequentist one, while Jeffreys followed the Bayesian approach.

[1] Here, it is assumed that $\theta \in [0, 1]$. The general case considers θ at the boundary of the parameter space Θ then.

In 1921, Wrinch and Jeffreys (1921) published the paper titled *XLII. On certain fundamental principles of scientific inquiry* in *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. It contained the first version of a Bayes factor and offered a solution to the dilemma that one needed to sample the entire population to obtain a sufficiently high posterior probability, as discovered by Broad (1918). It is important to note that both Wrinch and Jeffreys started their work while sharing the opinion that any scientific method must be useable in real research settings:

> "In order that a scientific method be of any value, it must satisfy two conditions. In the first place, it must be possible to apply it in the actual cases to which it is meant to be relevant. In the second, its arguments must be sound. The main object of science is to increase knowledge of the world, and if a method is not applicable to anything in the world it obviously cannot lead to any knowledge. This principle is very elementary, and it is probably for that very reason that it is habitually overlooked in theories of scientific knowledge."
> Wrinch and Jeffreys (1921, p. 369)

The biggest obstacle now was to resolve the paradox resulting from the principle of insufficient reason. With an infinite number of possible hypotheses, the prior probability of each hypothesis has to be effectively zero. Bayes' theorem then yields that the posterior, proportional to the prior cannot yield any posterior probabilities which can be distinguished from zero.[2] Of course, informative priors would solve for the problem, but the justification of those was questionable. According to Howie, "the first piece of the puzzle came to Wrinch during a picnic lunch taken together on Madingley Hill." (Howie, 2002, p. 105). Wrinch assumed that one could express every law of physics by some particular differential equation with rational coefficients, finite degree and order (Wrinch and Jeffreys, 1921, p. 386). Then, all hypotheses – including general laws – have to form an enumerable set. Ordering this set against the set of the integers and restricting the prior probabilities assigned to the elements of this ordered set to sum up to one was the solution, because then no infinitesimal probabilities were needed anymore.[3] Importantly, the number of possible general laws or hypotheses could still be infinite under the above procedure. Years later, Jeffreys (1931) noted this as a postulate in his book *Scientific Inference*:

> "Every quantitative law can be expressed as a differential equation of finite order and degree, in which the numerical coefficients are integers."
> Jeffreys (1931, p. 45)

The second strike was the simplicity postulate Wrinch and Jeffreys (1921) proposed. Referring to Broad (1918, p. 402), who had also stressed the importance of simplicity of hypotheses, Wrinch and Jeffreys (1921) stated, that simple laws have to be preferred

[2] From a strictly measure-theoretic perspective, the posterior distribution is absolutely continuous with respect to the prior predictive distribution, and the prior predictive distribution is itself absolutely continuous with respect to the prior distribution, compare **?**. Thus, any prior distribution that assigns zero prior probability to a hypothesis (which is the case when an infinite number of hypothesis is considered and each hypothesis has equal prior probability) results in a posterior that assigns zero probability to this hypothesis, too. While the example of Broad operates on discrete parameter spaces, the problem is the same when only enough hypotheses are considered.

[3] Mathematically, there is no uniform distribution on $\mathbb{R}$ or $\mathbb{N}$, so this motivated the ideas of Wrinch, as Laplace's principle of insufficient reason was not applicable.

over complex laws.[4] This way, the prior probability of each law in the sum described above was determined by the simplicity of the law itself. Simpler laws would be assigned higher prior probabilities and more complex laws smaller ones. Together, these two ideas leveraged inverse probability from the Laplacian tradition of the principle of insufficient reason to a modern interpretation, in which slightly informative priors were used, but the procedure was applied with the claim of high objectivity:

> "That scientists prefer simple laws is an empirical fact; it can provide a basis for ordering prior probabilities. Wrinch and Jeffreys announced a 'Simplicity Postulate': the simpler the law, the greater its prior probability. The Bayesian machinery can finally be cranked up."
> Howie (2002, p. 106-107)

The theory they introduced in their 1921 paper attempted to analyse the rules of logic to solve Broad's principle of insufficient reason. According to Wrinch and Jeffreys (1921), the character of science was revealed in the uncertainty of inferences made relative to a given body of data. Wrinch and Jeffreys (1921) also emphasized that prior knowledge should be incorporated into scientific inquiries. In essence, "...no probability is ever determined from experience alone. It is always influenced to some extent by the knowledge we had before the experience." (Wrinch and Jeffreys, 1921, p. 381). In order to make use of probability statements, they then introduced the following formula:

> "(...) the problem of the probability to be attached to an inference can be dealt with. If p denote the most probable law at any stage, and q an additional experimental fact, we can easily prove that
>
> $$\underbrace{\frac{P(p:q.h)}{P(\sim p:q.h)}}_{\text{posterior odds}} = \underbrace{\frac{P(q:p.h)}{P(q:\sim p.h)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{P(p:h)}{P(\sim p:h)}}_{\text{prior odds}}$$
>
> (...) Hence, even if p has not a very large prior probability, a single verification of a consequence not predicted by the contrary of p may raise the probability of p to something much greater than that of its contrary;"
> Wrinch and Jeffreys (1921, p. 387)

In the above notation, the underbraces have been added to improve readability and the formula follows from a simple application of Bayes' theorem as shown in Chapter 6. In modern terms, p and $\sim p$ can be read as the null and alternative hypothesis, while q is the observed data and h the unchanged background knowledge used to model the prior. Interestingly, while from a mathematical perspective the background knowledge is irrelevant to the above formula (one could easily omit or suppress h in each term), the fact that Wrinch and Jeffreys kept it in the formula underlines the importance of incorporating the available knowledge into a Bayesian analysis. This is in line with

[4] The analogy to Occam's razor is strong, and, remarkably, Wrinch and Jeffreys noticed this principle that early. Until today, the complexity of models, expressed in the number of parameters, is often used as a penalty in both frequentist and Bayesian model selection. Examples are regularized regression models, or information criteria (Hastie et al., 2015, 2017; Efron and Hastie, 2016; McElreath, 2020; Piironen and Vehtari, 2017).

the notion that no probability is determined from experience alone, but always is influenced to some extent by the knowledge we had before conducting the experiment.[5] Wrinch and Jeffreys (1921) did not explicitly name the quantity Bayes factor in 1921, but Etz and Wagenmakers (2017) noted, referencing Good (1988), that the centerpiece equation given above has been used in literature frequently ever since. Interestingly, credit to Wrinch and Jeffreys (1921) has been given rarely.

After the 1921 paper, two more papers followed. Wrinch and Jeffreys (1923a) published their joint paper under the name *'The theory of mensuration'* and delved further into the previous ideas of probabilistic inference from observed data. Concerning the Bayes factor, the paper added nothing new. Indeed, in 1923 the collaboration between both scientists already started to resolve. Jeffreys personal situation changed as he needed to find a permanent academic position, and he moved to Cambridge. Wrinch herself got married and left for Oxford where her husband had a position (Howie, 2002).[6] Jeffreys concentrated on geophysics from then on, and his work on seismology eventually lead to the detection that the earth's core is liquid. After the election as a fellow of the Royal Society in 1925 and appointment as lecturer at the university one year later, he finally was assigned as a geophysics reader in the year of 1931.

Having settled in his professional position, Jeffreys started from the 1930s on to reconsider his older work with Wrinch. Howie (2002) argued that the earlier work on applied geophysics caused Jeffreys to notice "the practical need for definitive numerical criteria for the evaluation of hypotheses and the probabilistic combination and reduction of data." (Howie, 2002, p. 114) This is one possible interpretation. Another one is given by the fact that Jeffrey's work in geophysics was partly motivated by the need to get out of his precarious professional situation, and applying for a permanent academic position was much easier in geophysics than in probability theory or statistics in the 1920s. After having settled in an academic position, he could spend more time on the topic of scientific inference again. Also, statistical aspects in seismology were not investigated deeply, so that it was an ideal way to apply the methods he developed earlier in collaboration with Dorothy Wrinch. Therefore, Jeffreys book *Scientific Inference* (Jeffreys, 1931) was the first step to a self-contained Bayesian methodology for scientific inference. Using the simplicity postulate as well as his and Wrinch's previous ideas, Jeffreys (1931) introduced their solutions to Laplace's principle of insufficient reason as well as their thoughts about general laws in the first part. In the second part, he went on to show that his theory was able to "account for the phenomenological development of real scientific theories." (Howie, 2002, p. 115). Without the older self-criticism about prior probabilities, Jeffreys stated that by

> "...analyzing the processes involved in our forward scientific reasoning we detect the fundamental postulate that it is possible to learn from experience. This is a primitive postulate, presumably on the frontiers between *a priori*

[5] The most drastical examples of this assumption are experiments in which people are tested for supernatural abilities like the ability to foresee the future. In these situations, the prior knowledge intuitively assigns a very low probability to the hypothesis that the tested person indeed has supernatural abilities, see Kadane (1987), Berger and Delampady (1987), Good (1981), Good (1993), Good (1994) as well as Robert (2007, p. 229-231), Rao and Lovric (2016), Zumbo and Kroc (2016), Sawilowsky (2016) for discussions about the existence and appropriateness of precise hypotheses in scientific research in general.

[6] Wrinch married John Nicholson in 1922, who was appointed at Balliol College, Oxford as a lecturer in mathematics. After Nicholson found students for her in Oxford, Wrinch moved from Cambridge to Oxford and the collaboration with Jeffreys ended (Howie, 2002).

> and empirical knowledge. The status of the laws of probability and the simplicity postulate is that of inferences from this principle."
> Jeffreys (1931, p. 47-48)

Jeffreys (1931) built upon Wrinch's idea of assigning prior probabilities to a general law according to the law's complexity. Therefore, he used the corresponding differential equation's degree, order and coefficients to form a complexity coefficient so that the prior probability of any hypothesis could be determined. Nevertheless, the impact of *Scientific Inference* remained low. Bennett (1990, p. 164) noted, that in a letter to Fisher on 5th June 1937, Jeffreys wrote that he wanted to redo the entire book again, but as it was not expected to be sold out in the next decades, there would be no use in such an effort. Also, one major drawback of *Scientific Inference* according to Howie (2002) was, that

> "Jeffreys's theory of scientific inference shared a status with the eighteenth-century doctrine of chances. In Jeffreys's case, the 'men of quality' were scientists, whose method was simply a sophisticated form of commonsense. The obligation on these scientists to adopt his assessments of prior distributions was to ensure consistency and uniformity with the collective approach to research."
> Howie (2002, p. 120)

However, this consistency could mainly be given a reason when the scientific process was successful, which was not always the case. Also, the exact form of the prior probabilities remained ambiguous to achieve the consistency desired in Jeffreys's scientific theory, so that in total, the reception was moderate. While the idea of using a convergent series to describe the prior probabilities was tempting, it remained vague how to do this in practice. Nevertheless, one important impact was induced on Jeffreys' subsequent work: He started to think about the 'correct' form of a prior, moving from his idea of assigning hypotheses an a priori probability according to their simplicity to more sophisticated methods. Later, these efforts resulted in Jeffreys' prior (see Definition 6.10), which enjoys transformation-invariance.

7.2 Haldane's alternative Approach

John Burdon Sanderson Haldane was a British-Indian scientist mainly known for his work in evolutionary biology and physiology. Also, he made important contributions to statistics, and this is where a link between Jeffreys and Haldane is revealed.[7] Born in 1892 in Oxford, Haldane left England in 1956 because of his political dissent as a professing atheist and Marxist. His life and work in India since the 1956s is detailed in Mcouat (2017) and his work on population genetics in India was analysed by Dronamraju (2012) and Dronamraju (2015)[8]. Haldane's statistical achievements are detailed in Etz and Wagenmakers (2017), and (Howie, 2002, p. 121-126) describes Haldane's 1932 paper. Etz and Wagenmakers (2017) noted that Haldane's work on the foundation of statistics is limited to a single paper in 1932, entitled *'A note on inverse probability'*, which

[7] For details on Haldane's early work on population genetics see Edwards (1993).

[8] For details on Ronald Fisher's work on population genetics, see Thompson (1990).

was published in the *Mathematical Proceedings of the Cambridge Philosophical Society*. Nevertheless, the paper is remarkably innovative, and multiple parallels can be drawn between Jeffreys' later work on Bayes factors and Haldane's paper. As the last section has shown, Jeffreys's early work on the Bayes factor addressed the paradox induced by Broad's principle of insufficient reason. However, his treatment on Bayes factors as statistical general purpose tools for quantifying the change in belief about two competing hypotheses was not very mature. Haldane published his paper just a few months after the publication of Jeffreys's *Scientific Inference*, and just a short time before Jeffreys extended his theory on Bayes factors and introduced his hypothesis methodology in a much clearer manner.

Haldane's paper is analysed in the following to investigate this relation. In his paper, he designated Fisher's theory of maximum likelihood as using the principle of insufficient reason by inherently assuming a uniform prior, so that the posterior in a Bayesian interpretation is proportional to Fisher's likelihood function (Haldane, 1932, p. 60). Note, that while Fisher's maximum likelihood method formally can be interpreted as a Bayesian approach by combining the likelihood $f(x|\theta)$ of the data x given the parameter θ with a uniform prior $f(\theta) = 1$, Fisher sharply objected to this interpretation. For him, the likelihood was a 'measure of rational belief, and for that reason is called likelihood (...). I stress this because in spite of the emphasis that I have always laid upon the difference between probability and likelihood there is still a tendency to treat likelihood as though it were a sort of probability.' (Fisher, 1930, p. 532). Haldane (1932) started his paper by questioning the assumption of the principle of insufficient reason, which was inherently assumed in Fisher's theory (at least according to Haldane), and also in (objective) inverse probability. His goal thus was to find more appropriate a priori probabilities:

> "The problem of statistical investigation is the description of a population, or Kollektiv, of which a sample has been observed. At best we can only state the probability that certain parameters of this population lie within assigned limits, i.e. specify their probability density. It has been shown by von Mises (1), that this is only possible if we know the probability distribution of the parameter before the sample is taken. Bayes' theorem is based on the assumption that all values of the parameter in the neighbourhood of that observed are equally probable *a priori* [referring to the principle of insufficient reason]. It is the purpose of this paper to examine what more reasonable assumption may be made, and how it will affect the estimate based on the observed sample."
> Haldane (1932, p. 55)

Haldane (1932) aimed at examining whether more informed priors could be used instead of a uniform one, and how more informed priors would change the posterior estimates. He started with a population of which a proportion x possesses a character X and of which a sample of n population members was taken. n was assumed to be large enough for the Bernstein-von-Mises approximation to hold, that is, to assume that the posterior is approximately normally distributed (van der Vaart, 1998, Chapter 10), and he denoted by a the number of individuals in the sample possessing the character X. Writing $f(x)$ as the prior density of x, Haldane (1932) noted:

> "It is an important fact that in almost all scientific problems we have a rough idea of the nature of $f(x)$ derived from the past study of similar populations.

> Thus, if we are considering the proportion of females in the human population of any large area, $f(x)$ is quite small unless x lies between .4 and .6."
> Haldane (1932, p. 55)

After advocating the use of informative priors constructed from prior knowledge and previous studies, Haldane (1932) moved on and gave an example from genetics to demonstrate that obtaining the posterior distribution of the parameter of interest is possible using such informative priors, see also (Etz and Wagenmakers, 2017, p. 7-9):

> "An illustration from genetics will make this point clear. The plant *Primula sinensis* possesses twelve pairs of chromosomes of approximately equal size. A pair of genes selected at random will lie in different chromosomes in $\frac{11}{12}$ of all cases, giving a proportion $x = 0.5$ of "cross-overs". In $\frac{1}{12}$ of all cases they lie in the same chromosome, the values of the cross-over ratio x ranging from 0 to 0.5 without any very marked preference for any part of this range, except perhaps for a tendency to avoid values very close to 0.5. $f(x)$ is thus approximately $\frac{1}{6}$ for $0 \leq x < 0.5$; it has a discontinuity at $x = 0.5$, such that the probability of this value is $\frac{11}{12}$; while, for $0.5 < x \leq 1$, $f(x) = 0$."
> Haldane (1932, p. 57)

To clarify the above statement, Haldane (1932) considered the random cross-over of genes which happens in meiosis during reproduction. When such a cross-over happens, the alleles which locate a gene on a chromosome randomly switch from the mother's to the father's chromosome (or vice versa). Haldane's goal was to estimate the cross-over rate based on the available cross-bred plants. He assumed that when the two alleles are on different chromosomes, a cross-over happens with probability 0.5. The probability that the two alleles are indeed on two distinct chromosomes itself is given by 11/12. If the two alleles are on the same chromosome, both alleles can switch to another chromosome during the cross-over, which causes variation in the cross-bred plants. Therefore, Haldane (1932) assumed that when the two alleles are on the same chromosome, the probability of a cross-over (in the sense that variation is created and phenotypically observable) is uniform between 0 and 0.5. The probability that the two alleles themselves lie in the same chromosomes is given by 1/12. In total, Haldane (1932) obtained a mixture-prior for the cross-over rate θ in the form of

$$\pi(\theta) = \frac{1}{12} \cdot \mathbb{1}_{(0,0.5)} \cdot \mathcal{U}(0,0.5) + \frac{11}{12} \cdot \delta(0.5) \tag{7.1}$$

where δ is the Dirac-function, $\mathcal{U}$ the uniform distribution's density, and 1/12 and 11/12 are the prior probabilities of the two models 1) genes lie in the same chromosome and 2) genes lie in different chromosomes. $\int_0^1 \pi(\theta) d\theta = \frac{1}{12} \cdot 0.5 \cdot \frac{1}{0.5-0} + \frac{11}{12} \cdot 1 = 1$, so $\pi(\theta)$ is indeed a probability density. $\mathcal{U}(0,0.5)$ and $\delta(0.5)$ are the prior distributions for the cross-over rate parameter in the respective model, and the mixture prior results by combining both models. In summary, Haldane (1932) obtained a mixture distribution which divides the available prior probability mass into a point mass (when model 1 is true, that is, the genes lie on different chromosomes) and a continuous probability density (when model 2 is true, that is, the genes lie on the same chromosome). Using his mixture prior, Haldane (1932) argued:

> "Now if a family of 400 seedlings from the cross between a doubly heterozygous plant and a double recessive contains 160 "cross-overs" we have two

> alternatives. The probability of getting such a family from a plant in which the genes lie in different chromosomes is $\frac{11}{12}^{400}C_{160}2^{-400}$, or 1.185×10^{-5}. The probability of getting it from a plant in which they lie in the same chromosome is
>
> $$\frac{1}{6}^{400}C_{160}\int_0^{\frac{1}{2}} x^{160}(1-x)^{240}dx$$
>
> Since this integral is very nearly equal to
>
> $$\int_0^1 x^{160}(1-x)^{240}dx, \text{ or } \frac{160!240!}{401!}$$
>
> this probability is approximately $\frac{1}{6\times 401}$, or 4.156×10^{-4}."
> Haldane (1932, p. 57)

The probability $\frac{11}{12}^{400}C_{160}2^{-400}$ is derived from Equation (7.1), where all $n = 400$ mothering and fathering plants have genes on different chromosomes, and with probability 1/2 a cross-over happens, and with probability 1/2 it does not, yielding $\frac{11}{12}^{400}$ as the probability of obtaining such a sample of plants, where $2^{-160} \cdot 2^{-240} = 2^{-400}$ is the probability of getting $a = 160$ cross-overs and C_{160} is the respective binomial coefficient.

In case that all $n = 400$ mothering and fathering plants have genes on the same chromosomes, Haldane (1932) simplified the calculations a little: This leads to the factor $\frac{1}{6}^{400}$ in front of the integral instead of $\frac{11}{12}^{400}$: He used that "In $\frac{1}{12}$ of all cases they lie in the same chromosome, the values of the cross-over ratio x ranging from 0 to 0.5 without any very marked preference for any part of this range, except perhaps for a tendency to avoid values very close to 0.5. $f(x)$ is thus approximately $\frac{1}{6}$ for $0 \leq x < 0.5$." (Haldane, 1932, p. 57).

In total, Haldane (1932) thus imagined the experiment to be conducted and that $n = 400$ new plants were obtained by cross-breeding the parents, out of which $a = 160$ plants possessed the character X (e.g. a specific stem or leaf colour). What Haldane (1932) had calculated here, in modern terms, would be called the *marginal likelihood* of the data $a = 160$ given the model M_1 or M_2. If the model M_1 is assumed, that is, the genes lie on different chromosomes, this marginal likelihood becomes

$$f(a=160|M_1) = \binom{400}{160}0.5^{160}(1-0.5)^{240} = \binom{400}{160}0.5^{400}$$

In the above, the probability θ (or x, in Haldane's notation) of a cross-over is exactly 0.5, as well as the probability of no cross-over happening, and cross-overs appear independently in each offspring plant, leading to a binomial distribution. Note that in the above, the factor $\binom{400}{160}$ equals C_{160} in Haldane's notation. Under model M_2, that is when genes lie on the same chromosome, the marginal likelihood becomes

$$f(a=160|M_2) = \binom{400}{160}\int_0^{0.5} \theta^{160}(1-\theta)^{240}d\theta$$

Note that by conditioning on model M_1 (or M_2), the factor $11/12^{400}$ (or $1/6^{400}$) disappears. Using Bayes' theorem, Haldane (1932) then arrived at the posterior probabilities of both models:

$$\mathbb{P}(M_i|a=160) = \frac{\mathbb{P}(M_i)\mathbb{P}(a=160|M_i)}{\mathbb{P}(M_1)\mathbb{P}(a=160|M_1)+\mathbb{P}(M_2)\mathbb{P}(a=160|M_2)}$$

where $\mathbb{P}(M_1) = 1/12$ and $\mathbb{P}(M_2) = 11/12$. Substituting the above marginal likelihoods and the prior probabilities for both models, yields then

$$\mathbb{P}(M_1|a=160) = 0.028$$
$$\mathbb{P}(M_2|a=160) = 0.972$$

Bayes factors and posterior model probabilities are closely related – see Robert (2016) – but Haldane (1932) here did not introduce any kind of Bayes factor, and also was not satisfied with the posterior model probabilities. He went on to derive a prediction for the cross-over rate θ. The prior distributions $\delta(0.5)$ and $\mathcal{U}(0,0.5)$ were also updated, given the data, leading to the posterior distributions for each model. The Dirac function is not changed at all by the data, so the posterior stays the same, compare (**?**, Example 2.3). The uniform prior $\mathcal{U}(0,0.5)$ changes into an approximately normal $\mathcal{N}(\frac{160}{400}, 0.0245^2)$ distribution, see Haldane (1932, p. 56-57) and Etz and Wagenmakers (2017, p. 9). Using the above calculations, the marginal posterior of the cross-over rate θ then also becomes a mixture given by

$$\begin{aligned}\pi(\theta|a=160) &= \mathbb{P}(M_1|a=160)\pi_1(\theta|a=160) + \mathbb{P}(M_2|a=160)\pi_2(\theta|a=160)\\ &= 0.028\cdot\delta(0.5) + 0.972\cdot\mathcal{N}(\frac{160}{400}, 0.0245^2)\end{aligned}$$

where $\pi_1(\theta)$ and $\pi_2(\theta)$ are the respective updated posterior distributions given $a = 160$ under model M_1 and M_2. Haldane (1932) finally arrived at the prediction for the cross-over rate θ as a model-averaged expectation[9]

$$\mathbb{E}[\theta|a=160] = 0.028\cdot 0.5 + 0.972\cdot\frac{160}{400} = 0.4028 \tag{7.2}$$

which matches the results obtained by Haldane (1932):

> "Thus the probability that the family is derived from a plant where the genes lie in different chromosomes and $x = .5$ is .028. Otherwise the mean value of x is .4, with standard error .0245. The overall mean value, or mathematical expectation, of x is .4028, and the graph of the probability density $\frac{dp}{dx}$ is an approximately normal error curve centered at $x = 0.4$ with standard deviation .0245, together with an infinity at $x = 0.5$."
> Haldane (1932, p. 57)

In the above quote, the probability density $\frac{dp}{dx}$ equals the posterior probability density of the cross-over rate θ. In total, the 1932 paper of Haldane is remarkable in multiple ways: First, Haldane (1932) introduced a mixture prior comprising a point mass component and a smoothly distributed component over the remaining parameter values. This approach was novel in itself and opened the door to Bayesian model-averaging to obtain the posterior expectation $\mathbb{E}[\theta|X]$. Second, Haldane (1932) derived the posterior

[9] Thus, Haldane made explicit use of Bayesian model-averaging to estimate the cross-over rate, compare Claeskens and Hjort (2008).

distribution of the cross-over rate using Bayes' theorem and used it to quantify the evidence based on the posterior mean and the standard deviation. What is more, his focus seems to be estimation under uncertainty, instead of hypothesis testing. This orientation is in sharp contrast to the emerging trend of hypothesis testing formalism at that time, compare Chapter 3 and Chapter 4, as well as Jeffreys' earlier work with Dorothy Wrinch (Wrinch and Jeffreys, 1921).

7.3 Jeffrey's Work after Haldane

According to Howie (2002), Jeffreys' work was influenced by Haldane's pragmatism regarding prior probabilities. Etz and Wagenmakers (2017) are more reluctant, and while they state that Jeffreys may have known Haldane's paper, it remains unclear how much influence Haldane's paper exerted on Jeffreys. Nevertheless, one major innovation of Haldane (1932) was to use a mixture prior, which assigned a fraction of the prior probability mass to a single parameter value via a Dirac-measure, and distributed the rest of the probability mass across the remaining parameter values by means of a different distribution. This idea could easily be translated to the problem of the principle of insufficient reason. Employing a mixture prior as used by Haldane (1932), one only needed to assign the extremes $\theta = 0$ and $\theta = 1$ finite prior probabilities and share the rest of the probability mass uniformly in between. This idea was adopted by Jeffreys to

> "...answer finally Broad's problem of the black crows. Jeffreys recommended packing some finite value of probability, k, into each of the extreme values, 0 and 1, and distributing the rest evenly. He showed that with k independent of the size of the class (and non-zero), the posterior distribution, following repeated viewing of black crows, peaks at a value also independent of the size of the class."
> Howie (2002, p. 125)

Mathematical details can be found in the appendix of Zabell (1989b). Zabell (1989b) also notes that "Within a year of Broad's 1918 paper, Jeffreys and Wrinch (1919) noted that the difficulty could be averted by using priors which place point masses at the endpoints of the unit interval)", refering to an earlier paper of Wrinch and Jeffreys in 1919. In it, Wrinch and Jeffreys (1919) wrote:

> "Here Mr. Broad's argument is valid, and no such law can derive a reasonable probability from experience alone; some further datum is required. One way of arriving at such laws may be suggested here. Suppose we have an a priori belief that, either every x has the property ϕ or every c has the property Ψ. If then a single x, say c, is found to satisfy ϕ but not Ψ, we can infer deductively the universal proposition that all x's satisfy ϕ. Such cases are fairly frequent: if for instance we consider that either Einstein's or Silberstein's form of the principle of general relativity is true, a single fact contradictory to one would amount to a proof of the other in every case."
> Wrinch and Jeffreys (1919, p. 729)

Thus, while the idea of sharing the probability mass into the two point-masses $\mathbb{P}(0) = \mathbb{P}(1) = 0.5$ shines through, the explicit formulation as a Dirac-mixture-prior $\pi(\theta) := 0.5 \cdot \delta(0) + 0.5 \cdot \delta(1)$ in the sense of Haldane (1932) is missing. Haldane (1932) made this notion much more explicit. Denoting the parameter as x, he explicitly assigned a point mass to this value which corresponded to a general law:

> "Let us suppose then, that k is the a priori probability that $x = 0$, and that the a priori probability that it has a positive value is expressed $f(x)$, where $\lim\limits_{\varepsilon \to 0} \int_\varepsilon^1 f(x)dx = 1 - k$."
> Haldane (1932, p. 59)

Thus, his mixture prior in modern notation could be written as $\pi(\theta) := k \cdot \delta_0(\theta) + f(\theta)$, where $\delta_0(\theta) := 1$ if $\theta = 0$ and else $\delta_0(\theta) = 0$, and $f(\theta) > 0$ for $\theta > 0$ and $f(\theta) = 0$ for $\theta \leq 0$.[10] Haldane then argued that if $\theta = 0$ (in his notation, $x = 0$), the probability to observe a sample $s := (0, 0, ..., 0)$ of size n which only consists of observations not having the property of interest is one, that is $\mathbb{P}(s|\theta = 0) = 1$. Using Bayes' theorem, Haldane (1932) then obtains the posterior of the parameter:

> Hence the probability, after observing the sample, that $x = 0$ is
>
> $$\frac{k}{k + \int_0^1 (1-x)^n f(x)dx}$$
>
> Haldane (1932, p. 59)

In modern notation, this can be expressed as

$$\begin{aligned}\mathbb{P}(\theta = 0|s) &= \frac{f(s|\theta = 0)\mathbb{P}(\theta = 0)}{f(s|\theta = 0)\mathbb{P}(\theta = 0) + f(s|\theta \neq 0)\mathbb{P}(\theta \neq 0)} \\ &= \frac{1 \cdot k}{1 \cdot k + \int_0^1 f(\theta)(1-\theta)^n d\theta} \end{aligned} \tag{7.3}$$

Assuming a flat prior $f(x) = 1$, Haldane (1932) calculated the integral

$$\int_0^1 f(\theta)(1-\theta)^n d\theta = \int_0^1 (1-\theta)^n d\theta = \frac{1}{n+1}$$

In total, the posterior probability of the parameter therefore becomes

$$\mathbb{P}(\theta = 0|s) = \frac{k}{k + \frac{1}{n+1}} = \frac{kn}{kn + k + 1} + \frac{k}{kn + k + 1}$$

Importantly, Haldane's calculations show that if a point mass $k > 0$ is assumed for the prior probability of $\theta = 0$ (or $x = 0$, in his notation), the posterior probability of a hypothesis corresponding to the general law $\theta = 0$ converges to 1 for increasing sample size n, as $\frac{kn}{kn+k+1} \xrightarrow[n\to\infty]{} 1$ and $\frac{k}{kn+k+1} \xrightarrow[n\to\infty]{} 0$. Importantly, for a large point mass $\mathbb{P}(\theta = 0) = k \approx 1$, the speed of this convergence is much faster compared to the situation in which only a small point mass $\mathbb{P}(\theta = 0) = k \approx 0$ is assigned to the general law $\theta = 0$. Also, his mixture-prior revealed that the paradox of Broad (1918)

[10] This prior is proper, as can be seen by calculating $\int_0^1 p(\theta)d\theta = \int_0^1 \delta_0(\theta) + f(\theta)d\theta = \int_0^1 \delta_0(\theta)d\theta + \int_0^1 f(\theta)d\theta$ and using $\int_0^1 \delta_0(\theta)d\theta = \mathbb{P}(\theta = 0) \cdot 1 + \mathbb{P}(\theta \neq 0) \cdot 0 = \mathbb{P}(\theta = 0) = k$ and $\int_0^1 f(\theta)d\theta = \lim\limits_{\varepsilon \to 0} \int_\varepsilon^1 f(\theta)d\theta = 1 - k$.

disappears, as the probability of the general law can become close to 1 even if only a small sample size n is observed, when the point mass k is selected ≈ 1 (or close to 1).

The explicit formulation of a mixture-prior by Haldane led to the desirable situation in which a general law approached unit probability even before the whole of the population had been sampled.[11] Still, while Jeffreys' may have been influenced by the paper of Haldane (1932), the more interesting aspect with a perspective on hypothesis testing is if Jeffreys' development of the Bayes factor was influenced by Haldane's paper, too.

It is worthwhile to take a look at a paper Jeffreys published two years after Haldane to analyse the influence Haldane may have exerted on Jeffreys. Jeffreys (1935) entitled the paper *'Some Tests of Significance, Treated by the Theory of Probability'* and published it in the *Mathematical Proceedings of the Cambridge Philosophical Society*. In it, Jeffreys stated the goal of the paper as follows:

> "Suppose that two different large, but not infinite, populations have been sampled in respect of a certain property. One gives x specimens with the property, y without; the other gives x' and y' respectively. The question is, whether the difference between x/y and x'/y' gives any ground for inferring a difference between the corresponding ratios in the complete populations. Let us suppose that in the first population the fraction of the whole possessing the property is p, in the second p'. Then we are really being asked whether $p = p'$; and further, if $p = p'$, what is the posterior probability distribution among values of p; but, if $p \neq p'$, what is the distribution among values of p and p'."
> Jeffreys (1935, p. 203)

Jeffreys (1935) further assumed that two large populations (not hypothetically infinite populations as in Fisher's definition of probability) have been sampled with respect to the property of interest. In modern terms, Jeffreys opposed two hypotheses or models M_0 and M_1, where in the first model $p = p'$ holds and in the second model $p \neq p'$. In modern notation, this equals the models $M_0 : \theta_0 = \theta_1$ and $M_1 : \theta_0 \neq \theta_1$. Jeffreys then assigned a prior probability of 0.5 on both models and in the case of M_0 being true, so that $p = p'$ holds, he set the prior distribution $p_0(\theta_0)$ of θ_0 as uniform on $(0, 1)$, that is $\theta_0 \sim \mathcal{U}(0, 1)$. In the case of M_1 being true, Jeffreys gave θ_0 and θ_1 each their own uniform prior probability distributions $p_1(\theta_0)$ and $p_1(\theta_1)$, each distributed as $\mathcal{U}(0, 1)$. In summary, he obtained $p_0(\theta_0) = p_1(\theta_0) = p_1(\theta_1) = \mathcal{U}(0, 1)$, and by the assumption of independence (Jeffreys, 1935, p. 204) obtained $p_1(\theta_0, \theta_1) = p_1(\theta_0)p_1(\theta_1)$. Jeffreys (1935) then assumed the likelihood functions under the models M_0 and M_1 to be

$$f(d|\theta_0, M_0) = \frac{(x_0 + y_0)!}{x_0!y_0!}\frac{(x_1 + y_1)!}{x_1!y_1!}\theta_0^{x_0}(1-\theta_0)^{y_0}\theta_0^{x_1}(1-\theta_0)^{y_1} \tag{7.4}$$

and

$$f(d|\theta_0, \theta_1, M_1) = \frac{(x_0 + y_0)!}{x_0!y_0!}\frac{(x_1 + y_1)!}{x_1!y_1!}\theta_0^{x_0}(1-\theta_0)^{y_0}\theta_1^{x_1}(1-\theta_1)^{y_1} \tag{7.5}$$

which are simply binomial likelihood functions, see also Etz and Wagenmakers (2017, p. 15) and Ly et al. (2016b, Appendix D). Here, d denotes the observed data (x, y, x', y') in both groups. Using the prior model probabilities $\mathbb{P}(M_0) = \mathbb{P}(M_1) = 0.5$ on both

[11] See also (Howie, 2002, p. 125).

models M_0 and M_1, Jeffreys then derived the posterior distribution $p(\theta_0|d, M_0)$ for the free parameter θ_0 in the model M_0 and the posterior distribution $p(\theta_0, \theta_1|d, M_0)$ for the two free parameters θ_0, θ_1 in the model M_1. These posterior distributions are proportional to the corresponding model likelihoods, because by the independence $p(\theta_0, M_0) = p(\theta_0)\mathbb{P}(M_0)$ one obtains

$$\begin{aligned} p(\theta_0|d, M_0) &\propto f(d|\theta_0, M_0)p(\theta_0, M_0) = f(d|\theta_0, M_0)p(\theta_0)\mathbb{P}(M_0) \\ &= p(d|\theta_0, M_0) \cdot 1 \cdot 0.5 \propto p(d|\theta_0, M_0) \end{aligned} \tag{7.6}$$

and by the independence $p(\theta_0, \theta_1, M_1) = p(\theta_0)p(\theta_1)\mathbb{P}(M_1)$ it follows that

$$\begin{aligned} p(\theta_0, \theta_1|d, M_1) &\propto f(d|\theta_0, \theta_1, M_1)p(\theta_0, \theta_1, M_1) = f(d|\theta_0, \theta_1, M_1)p(\theta_0)p(\theta_1)\mathbb{P}(M_1) \\ &= f(d|\theta_0, \theta_1, M_1) \cdot 1 \cdot 1 \cdot 0.5 \propto f(d|\theta_0, \theta_1, M_1) \end{aligned} \tag{7.7}$$

Therefore, the posteriors for θ_0 and θ_1 are obtained as

$$p_0(\theta_0|d) \propto \theta_0^{x_0+x_1}(1-\theta_0)^{y_0+y_1} \tag{7.8}$$

$$p_1(\theta_0, \theta_1|d) \propto \theta_0^{x_0}(1-\theta_0)^{y_0}\theta_1^{x_1}(1-\theta_1)^{y_1} \tag{7.9}$$

where $p_0(\cdot|d)$ and $p_1(\cdot, \cdot|d)$ are the posterior probability densities of θ_0 and (θ_0, θ_1) in model M_0 and model M_1, after observing the data d. By integrating

$$p(M_0, \theta_0|d) \propto f(d|M_0, \theta_0)p(M_0, \theta_0) = f(d|M_0, \theta_0)\mathbb{P}(M_0)p_0(\theta_0) \tag{7.10}$$

with respect to θ_0, Jeffreys (1935) obtained the posterior model probability $\mathbb{P}(M_0|d)$, where in Equation (7.10) the prior model probability $\mathbb{P}(M_0) = 0.5$ and $p_0(\theta_0) = \mathcal{U}(0, 1)$ as detailed above, and $f(d|M_0, \theta_0)$ is given in Equation (7.6), which leads to Equation (7.8). The same procedure leads to the posterior probability $\mathbb{P}(M_1|d)$ by integrating

$$p(M_1, \theta_0, \theta_1|d) \propto f(d|M_1, \theta_0, \theta_1)\mathbb{P}(M_1)p_1(\theta_0)p_1(\theta_1) \tag{7.11}$$

where the independence assumption $p_1(\theta_0, \theta_1) = p_1(\theta_0)p_1(\theta_1)$ is used. Here, $\mathbb{P}(M_1) = 0.5$, and $p_1(\theta_0) = p_1(\theta_1) = \mathcal{U}(0, 1)$. Therefore (as the density of a $\mathcal{U}(0, 1)$ distribution and the factor 0.5 can be omitted when removing proportionality constants), the posterior model probabilities are proportional to the likelihoods given in Equation (7.4) and Equation (7.5). Making use of the identity

$$\int_0^1 \theta_0^{x_0}(1-\theta_0)^{y_0}d\theta_0 = \frac{x_0!y_0!}{(x_0+y_0+1)!} \tag{7.12}$$

the integration of $p(M_0, \theta_0|d)$ given in eq. (7.10) yields

$$\mathbb{P}(M_0|d) \propto \frac{(x_0+x_1)!(y_0+y_1)!}{(x_0+x_1+y_0+y_1+1)!} \tag{7.13}$$

and the integration of $p(M_1, \theta_0, \theta_1|d)$ yields

$$\mathbb{P}(M_1|d) \propto \frac{x_0!y_0!}{(x_0+y_0+1)!} \cdot \frac{x_1!y_1!}{(x_1+y_1+1)!} \tag{7.14}$$

as $p(M_1, \theta_0, \theta_1|d) \overset{\text{Equation (7.11)}}{\propto} f(d|M_1, \theta_0, \theta_1) \overset{\text{Equation (7.5)}}{\propto} \theta_0^{x_0}(1-\theta_0)^{y_0}\theta_1^{x_1}(1-\theta_1)^{y_1}$ and therefore the integration with respect to θ_0, θ_1 can be calculated as

$$\begin{aligned}\mathbb{P}(M_1|d) &\propto \int_0^1 \int_0^1 \theta_0^{x_0}(1-\theta_0)^{y_0}\theta_1^{x_1}(1-\theta_1)^{y_1} d\theta_0 d\theta_1 \\ &= \int_0^1 \theta_0^{x_0}(1-\theta_0)^{y_0} d\theta_0 \int_0^1 \theta_1^{x_1}(1-\theta_1)^{y_1} d\theta_1 \\ &\overset{\text{Equation (7.12)}}{=} \frac{x_0!y_0!}{(x_0+y_0+1)!} \cdot \frac{x_1!y_1!}{(x_1+y_1+1)!} \end{aligned} \tag{7.15}$$

Employing the above derivations, Jeffreys (1935) finally arrived at his goal: The ratio of the posterior model probabilities $\mathbb{P}(M_0|d)$ and $\mathbb{P}(M_1|d)$ as given in Equation (7.13) and Equation (7.14) is the ratio of the posterior odds (compare with Definition 6.11). Therefore, when using the prior probabilities of $\mathbb{P}(M_0) = \mathbb{P}(M_1) = 0.5$ for both models (or prior odds $\mathbb{P}(M_0)/\mathbb{P}(M_1) = 1$, not favouring one of both models a priori), the ratio of the posterior model probabilities equals the ratio which today is called the Bayes factor, which Jeffrey aimed at in his derivations. That the Bayes factor can be calculated from the posterior odds and prior odds, can be seen directly from the following equation:

$$\underbrace{\frac{f(d|M_0)}{f(d|M_1)}}_{\text{Bayes factor}} = \underbrace{\frac{\mathbb{P}(M_0|d)}{\mathbb{P}(M_1|d)}}_{\text{posterior odds}} \Big/ \underbrace{\frac{\mathbb{P}(M_0)}{\mathbb{P}(M_1)}}_{\text{prior odds}} \tag{7.16}$$

Jeffreys (1935) summarised his derivation as follows at this point:

> "We have in each case considered the existence and the non-existence of a real difference between the two quantities estimated as two equivalent alternatives, each with prior probability 1/2. This is a common case, but not general. If however the prior probabilities are unequal the only difference is that the expression obtained for $P(q|\theta h)/P(\sim q|\theta h)$ now represents $\frac{P(q|\theta h)}{P(\sim q|\theta h)} / \frac{P(q|h)}{P(\sim q|h)}$. Thus if the estimated ratio exceeds 1, the proposition q is rendered more probable by the observations, and if it is less than 1, q is less probable than before. It still remains true that there is a critical value of the observed difference, such that smaller values reduce the probability of a real difference. The usual practice is to say that a difference becomes significant at some rather arbitrary multiple of the standard error; the present method enables us to say what that value should be. If however, the difference examined is one that previous considerations make unlikely to exist, then we are entitled to ask for a greater increase of the probability before we accept it, and therefore for a larger ratio of the difference to its standard error."
> Jeffreys (1935, p. 221)

In the above quote, $\frac{P(q|h)}{P(\sim q|h)}$ are the prior odds and $\frac{P(q|\theta h)}{P(\sim q|\theta h)}$ the posterior odds and q and $\sim q$ the null and alternative hypothesis under consideration. In modern terms, this yields exactly the Bayes factor as given in Equation (7.16).

Etz and Wagenmakers (2017) noted, that Jeffreys (1935) probably alluded to Fisher's significance testing when advertising the independence of newly introduced Bayes factor from the usual practice to say that a difference is significant at an arbitrary multiple of the standard error. The Bayes factor as a dimensionless quantity gets calibrated by the sample size and prior probabilities according to Jeffreys (1935). To make his argument, Jeffreys (1935, p. 205) approximated the posterior odds via a normal approximation to the binomial distribution for large sample size as

$$\frac{P(q|\theta,h)}{P(\sim q|\theta,h)} \approx \left(\frac{(x+x'+y+y')(x+y)(x'+y')}{2\pi(x+x')(y+y')}\right)^{\frac{1}{2}} \cdot e^{\left(-\frac{1}{2}\frac{(x+x'+y+y')(xy'-x'y)^2}{(x+x')(y+y')(x+y)(x'+y')}\right)} \tag{7.17}$$

and noted that when the difference of sampling ratios $xy' - x'y$ is small, the quadratic term $(xy' - x'y)^2$ in the exponential function becomes close to zero and the exponential function becomes approximately one. Then, when the sample is sufficiently large, the first term is larger than one as can be seen from the numerator (remember that x, x', y and y' were the number of specimens in the sample with and without the property) and "q approaches certainty" (Jeffreys, 1935, p. 205)[12]. On the other hand, if $xy' - x'y$ is large, the exponential factor becomes extremely small and "q approaches impossibility" (Jeffreys, 1935, p. 205). Jeffreys (1935) summarised

> "The theory therefore shows that a small difference between the sampling ratios may establish a high probability that the ratios in the main populations are equal, while a large one may show that they are different."
> Jeffreys (1935, p. 205)

Jeffreys (1935) explained further:

> "...agreement between the two populations becomes more probable if the samples are large and the difference of the sampling ratios small; when the ratio is large at $xy' - x'y = 0$, a larger value of the exponent is obviously needed to reduce the product to unity."
> Jeffreys (1935, p. 205)

Rephrasing Jeffreys idea, if the sampling ratio $xy' - x'y$ is large, then the exponential function in Equation (7.17) becomes very small. Nevertheless, if the Bayes factor (of which Equation (7.17) is an approximation only if equal prior probabilities of 1/2 are used for the hypotheses under comparison) is large, an even larger first factor in Equation (7.17) (whose size itself depends on the sample size) is needed compared to the situation when the sampling ratio $xy' - x'y$ is close to zero.

If the difference $xy' - x'y$ is approximately zero, the exponential function is close to one and if the sample is large the first factor in Equation (7.17) also becomes large, indicating support for q, that is an agreement between the two populations, or $p = p'$, or $\theta_0 = \theta_1$.

So, increasing the sample size yields a larger first factor in Equation (7.17), and therefore obtaining a small Bayes factor which indicates rejection of the null hypothesis becomes more difficult for increasing sample size when $xy' - x'y$ is small.

Later Jeffreys (1939) published critical values of the Bayes factor in ToP (see Table 6.1) which are also arbitrary like the p-values introduced by Fisher or the fixed test level advertised by Neyman and Pearson. However, the interpretation is much more natural, because a Bayes factor BF_{01} passing the threshold one indicates that the predictive ability of the hypothesis H_0 under consideration is larger than the predictive

[12] Note, that proposition q was defined by Jeffreys (1935) as $p = p'$, so in modern words this equals the null hypothesis of no difference between both groups.

ability of the alternative H_1, where the marginal likelihood quantifies the predictive ability. Thus, a Bayes factor $\mathrm{BF}_{01} = 2$ can be interpreted as the data being generated under H_0 is two times as likely as the data being produced by H_1.

7.4 Comparison of Jeffreys' and Haldane's Approach

In sum, the last sections showed that both Wrinch and Jeffreys (1921), Haldane (1932) and Jeffreys (1935) had their impact on the evolution of the Bayes factor as it is known today. Etz and Wagenmakers (2017) discussed the personal relationship between Jeffreys and Haldane with the conclusion that more credit should be given to Haldane. As was shown in the preceding section, the more important aspect of these developments lies in the difference between Jeffreys' and Haldane's goals: While Jeffreys (1935) focussed on the comparison of two competing hypotheses under consideration, Haldane was interested primarily in the posterior distribution of the parameter θ. Also, Haldane wanted to obtain point and interval estimates like $\mathbb{E}[\theta|X]$ and quantities to estimate the uncertainty of the point estimates like the posterior distribution's parameters μ and σ^2. It would have been straightforward for Haldane (1932) to calculate the Bayes factor as it is known today from his derivations, but he probably saw no use in that.

On the other hand, Jeffreys was neither interested in posterior distributions, nor posterior point or interval estimates. His derivations crystalize the Bayes factor as the penultimate quantity to quantify the change in belief towards one of two hypotheses under consideration. Both share the idea of assigning a finite probability to the point null hypothesis and distributing the rest of the available probability mass evenly among the rest of the parameters' support. This idea was very clearly articulated by Haldane, and played a substantial role in solving the paradox of the principle of insufficient reason. Maybe this enabled Jeffreys to proceed with his derivations of the Bayes factor, but Jeffreys (1935) himself was well aware of the possibility to obtain the posterior distribution as a mixture of the null and alternative hypothesis, see Jeffreys (1935, p. 222). However, he regarded it as superfluous since already the Bayes factor could quantify the necessary change in belief. Nevertheless, two important objections to his Bayes factor did Jeffreys (1935) already mention in his own 1935 paper:

> "To raise the probability of a proposition from 0.01 to 0.1 does not make it the most likely alternative. The increase in such cases, however, depends wholly on the prior probability, and this investigation therefore separates into two parts the ratio of the observed difference to its standard error needed to make the existence of a real difference more likely than not; the first can be definitely evaluated from the observational material, while the second depends wholly on the prior probability."
> Jeffreys (1935, p. 221)

First, this means that Jeffreys did notice the strong dependence of the Bayes factor on the prior distributions on the parameters in each model (not on the prior *model* probabilities)[13]. Second, this indicates that Jeffreys also was aware of his Bayes factor only

[13] Note that the Bayes factor can also be obtained by manually calculating the ratio of marginal likelihoods $f(x|M_0)/f(x|M_1)$ for two models M_0 and M_1. The marginal likelihood $f(x|M_i) = \int f(x|M_i,\theta_i)p(\theta_i|M_i)d\theta_i$ is obtained by integrating the likelihood $f(x|\theta_i, M_i)$ with respect to the model parameters θ_i in the model M_i under consideration, $i = 1, 2$. To perform this step, prior probabilities $p(\theta_i|M_i)$ are needed for each model M_i and these exert influence on the resulting Bayes factor.

measuring the evidence of the null hypothesis *relative* to the alternative. Measuring a relative quantity, a researcher therefore never can be sure if any of two hypotheses is a good description of the experimental situation. Termed differently: "All models are wrong, but some are useful." (Box, 1976). Picking two bad models in the form of competing hypotheses, therefore, may yield a high Bayes factor for one of them. However, even the favoured hypothesis may be an imprecise description of the underlying scientific situation. As can be seen from the quote above, even a large Bayes factor leveraging the probability of a given hypothesis from 0.01 to 0.1 does not indicate that the hypothesis resembles a good model of the scientific situation at hand. It only indicates the necessity of a change in belief towards the hypothesis, and after all, the prior model probabilities are substantial for the resulting conclusion.

7.5 Fisher's Dissent

It is well-known that Fisher completely rejected any Bayesian methodology.[14] Fisher's lifelong rejection of the Bayesian approach is described in Aldrich (2008). Interestingly, Fisher (1936) himself noted concerning inverse probability – which he learned in the standard curriculum at school, see also Howie (2002, p. 61) – that he "for some years found no reason to question its validity." (Fisher, 1936, p. 248) As Howie (2002) noted, Fisher's interpretation of probability changed when he accepted the position at Rothamsted experimental station in 1919. Due to Pearson's sharp criticism of Fisher's first paper (Fisher, 1912), Fisher probably reconsidered his probability definition at Rothamsted. His early work was an attempt to reconcile the two rivalling parties of biometricians and Mendelian geneticists. The former relied on the normal distribution to describe many traits, while the latter expressed the combination and permutation of genes with combinatorial methods. Fisher's frequency definition, which involved the hypothetical infinite population as given in Chapter 3 probably took form in Rothamsted. Howie (2002) noted:

> "...Mendelism was unique in involving a chance mechanism that generated with exact and fixed probability one of a set of clearly-defined outcomes. Genetic probabilities could thus be treated as *inherent* to the world rather than reflecting incomplete knowledge."
> Howie (2002, p. 61)

This inherently fixed quantity is a characteristic description of frequentist probability, in which the true parameter of interest is regarded as a fixed quantity. Fisher's definition involving a hypothetical infinite population and the example with a die has strong analogies to a probability concept which is rooted in Mendelism:

> "Mendelism, like throws of a die or tosses of a coin, calls for a frequency definition of probability. By definition, gametes distribute by chance, and the long-run frequency of a given genotype in a large offspring generation can be predicted exactly from the genetic make-up of the parents and the rules of combinatorial analysis."
> Howie (2002, p. 62)

[14] See also Aldrich (1997) and Stigler (2005).

Another reason for Fisher's changing probability definition may be seen in his college education. John Venn, who was a fervent proponent of the frequentist probability definition, was President of Caius College at Fisher's undergraduate time. Fisher's probability concept probably settled at the late 1910s, and at that time his lifelong objection to inverse probability manifested itself. This objection was also because of the incompatibility of the Bayesian approach with the probability definition induced by Mendelism: Frequencies of genes in a population can naturally be described as a frequency ratio. The fundamental idea of Bayesian statistics to update the prior with the likelihood into the posterior made no sense from a Mendelian perspective, see also Howie (2002, p. 70). The frequency of genes in a population does not change when conducting a study. It is a fixed parameter of the population. From a Bayesian perspective, of course, this frequency is a random variable which changes due to birth and death processes which happen in real-time, so that the process of Bayesian inference leads to a 'less-delayed' estimate of the continuous, ever-changing frequency of a gene in the population. Already in the introduction chapter of the first edition of *Statistical Methods for Research Workers*, Fisher (1925a) wrote:

> "For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences are usually distinguished under the heading of **Inverse Probability**, and have at times gained wide acceptance. This is not the place to enter into the subtleties of a prolonged controversy; it will be sufficient in this general outline of the scope of Statistical Science to express my personal conviction, which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected. Inferences respecting populations, from which known samples have been drawn, cannot be expressed in terms of probability ..."
> Fisher (1925a, p. 10)

Therefore it was a vexing trend for Fisher that pioneers like Harold Jeffreys or J.B.S. Haldane introduced a kind of Bayesian renaissance, and championed the use of inverse probability despite the presence of Fisher's theory of maximum likelihood. While Fisher also disagreed with the Neyman-Pearson theory as analysed in Chapter 5, Neyman's and Pearson's theory at least was solely based on frequentist grounds.[15] After he read Haldane's 1932 publication, Fisher (1932) sharply criticised Haldane for trying to anchor his maximum likelihood method within the theory of inverse probability. The objection was targeted to Haldane's practice of combining any prior information, which could be connected with inverse probability, with Fisher's non-probabilistic likelihood function:

> "Fisher defines the likelihood of x as a quantity proportional to $e^{L(x)}$. This is a convenience of statement, but the introduction of the a priori probability density $f(x)$ allows the deduction of Fisher's results without introducing concepts other than those found in the theory of direct probability."
> Haldane (1932, p. 60)

[15] However, as shown in Chapter 4 Neyman mentioned that the Bayesian approach often is "the more logical of the two" (Neyman and Pearson, 1928, p. 176), but the Neyman-Pearson-theory somehow ended up being a frequentist approach, probably out of the general rejection of Bayesian inference due to Fisher's influential opinion at that time.

In modern notation, combining a flat prior $p(\theta) \propto 1$ with the model likelihood $f(x|\theta)$ leads to a posterior distribution $p(\theta|x)$ which is proportional to Fisher's likelihood:

$$p(\theta|x) \propto f(x|\theta)p(\theta) \propto f(x|\theta)$$

Therefore, it is possible to interpret Fisher's likelihood as a Bayesian posterior under the uniform prior $p(\theta) \propto 1$ in *any* statistical model. Fisher was well aware that with increasing sample size the influence of any prior vanishes, and tried to use this as an argument against using priors at all. Concerning Haldane's calculation of the posterior expectation of the cross-over rate of cross-bred plants given in Equation (7.2), Fisher (1932) commented:

> "Knowing the frequency distribution of x [the cross-over rate] we could, of course, calculate its mean value, its median – that value which would be exceeded in 50 trials out of 100 – or any other characteristic that might be required, and the fact with which we are here concerned is that, of the two factors of which our frequency element is composed, that which is contributed by, and may be calculated from, our observations, becomes, as the sample is increased, more and more influential, while the factor $f(x)dx$, contributed by our *a priori* knowledge, becomes less and less influential in determining these quantities; so that (...) we may say that our conclusions tend to be the same, as the abundance of our data is increased without limit, whatever the particular form of our a priori information."
> Fisher (1932, p. 258)

After underlining that the influence of any prior vanishes for increasing sample size (no matter how that prior is exactly chosen) Fisher (1932) pointed out the danger when using erroneous priors:

> "We have of course no such assurance of the harmlessness of erroneous a priori assumptions, when our observations are finite in number, as is invariably the case in practice."
> Fisher (1932, p. 258)

Fisher then went on and criticised Haldane's arbitrary selection of a uniform prior $f(x) = 1$, and after repeating his arguments about the expendability of any a priori information, he defended his likelihood approach. He struggled with Haldane's calculations, in which the prior is combined with the likelihood function to obtain the posterior. For Fisher, no priors were needed at all and combining them with his likelihood function could not lead to any probability statement[16]:

> "It [the likelihood] is not a probability and does not obey the laws of probability. It can, however, be shown to provide, not only in the estimation of a probability, but in the whole field of statistical estimation, as satisfactory a measure of "degree of rational belief" as a probability could do. For this reason I have termed it, or some arbitrary multiple of it, the likelihood, based on the information supplied by the sample, of any particular value of x."
> Fisher (1932, p. 259)

[16] According to Howie (2002), "Jeffreys objected to Fisher's point that priors were irrelevant for induction. On the contrary, since the likelihood function merely summarized the sample, any inference concerning the whole class *required* some additional information." (Howie, 2002, p. 124). Note that this attitude also was stressed by Wrinch and Jeffreys (1921) already, who emphasized that prior knowledge has to be incorporated into any statistical analysis to learn from the observed data.

The similarity to a Bayesian posterior when reading the above quote is striking. Bayesian posteriors are also derived neglecting proportionality constants, yielding the same arbitrary multiples of a posterior distribution as Fisher does with his likelihood. However, Bayes' theorem then ensures that the resulting quantity obeys the rules of probability theory and is a probability measure.

In addition to the objections to Haldane (1932), Fisher also had an argument with Jeffreys about the definition of probability. It is important to note, that while Fisher was a proponent of rigorous experimental design who obtained his data mainly from well-planned agricultural experiments, Jeffreys's work led to an entirely different probability concept. Jeffreys's work on seismology prohibited the minute design of experiments which Fisher required. Randomization, agricultural designs like the Latin Square, and independent repetition of an experiment as advertised by Fisher were simply not possible regarding Jeffreys' work on earthquakes. In contrast, Jeffreys had to observe nature and update his current belief in a hypothesis according to new observational data at hand combined with a reasonable prior. Therefore, "the probability calculus thus became for Jeffreys a model of the fundamental process of learning. (...) This cohered with the operational philosophy he had developed from Pearson, and the associated idea of scientific laws as ever-improving probability distributions." (Howie, 2002, p. 127). The clash of both men has already been described in detail in (Howie, 2002, Chapter 5) and (Aldrich, 2006). Also, Lane (1980) and Bartlett (1933) give a good account of the argument. While the initial problem appeared as a solely mathematical dispute in which it was discussed if some calculations were allowed or not, the exchange was, in fact, more profound in that it led both men to recognise that the probability concept of the other was entirely different and not in line with one's own interpretation. Jeffreys's concept of probability was a more subjective, psychological interpretation of probability which measured the degree of belief in a proposition relative to a given body of data. Fisher's frequency concept was based on the hypothetical infinite population, which included the possibility of repeating an experiment under the same circumstances again and again like it is the case in genetics or agriculture. For Fisher, in Jeffreys's concept

> "...the idea that a probability can have an objective value, independent of the state of our information, in the sense that the weight of an object, and the resistance of a conductor have objective values, is here completely abandoned."
> Fisher (1934a, p. 3-4)

For Jeffreys, Fisher's theory was built upon an error. He objected that

> "...the hypothetical infinite population does not exist, that if it did its properties would have to be inferred from the finite facts of experience and not conversely, and that all statements with respect to ratios in it are meaningless."
> Jeffreys (1933, p. 533)

Similar to the Fisher-Neyman-Pearson dissent, the difference between Fisher and Jeffreys can be attributed to the different background of both men.[17] In the case of the

[17] Fisher's position strongly influenced Karl Popper later in developing his theory of falsification and rational empirism as a theory of science, compare (Popper, 1959) and Chapter 10.

Fisher-Neyman-Pearson dissent, careful consideration of each test under incorporation of the experimental design and professional knowledge for Fisher was opposed to practical guidance for the behaviour of the researcher to control the long-term error probabilities for Neyman and Pearson. In the dispute with Jeffreys, the objection to the other's concept of probability stemmed from Fisher's roots in Mendelism and genetics where frequency concepts seem reasonable, and Jeffreys's work in seismology and astronomy, where anything like a hypothetical infinite population or even the repetition of an experiment seemed absurd. The exchange still had one beneficial effect: It clearly showed the differing frequentist versus Bayesian probabilistic concept of both men and made it difficult for each of them to discredit the other's concept. Jeffrey's objection to a frequentist interpretation of probability is expressed nicely in the following quote, which shows that for him, any parameter of interest was not regarded as fixed:

> "...you [Fisher] are regarding a probability as a statement about the composition of the world as a whole, which it is not and on a scientific procedure could not be until there was nothing more to do."
> Jeffreys in a letter to Fisher, dated 10th April 1934 (Bennett, 1990, p. 160)

Both scientists exchanged multiple letters about their issues with each others probability concept. For Fisher, Jeffreys's concept of an epistemic probability remained ambiguous, being prone to subjectivity and lacking the necessary objectivity his frequency definition offered. For Jeffreys, objectivity did not exist at all with respect to probability. For him, not the hypothetical infinite population – which does not exist – was important, but only the actual data observed during an experiment.[18] Jeffreys therefore also objected to Fisher's significance testing:

> "An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure."
> Jeffreys (1939, p. 316)

This famous quote from Jeffreys challenges the questionable practice of using p-values – see Definition C.83 – which are defined as probabilities of sets including outcomes of the experiment or study, which have *not* occurred. These are the outcomes which are more extreme than the ones observed in the actual experiment or study, and basing inference on outcomes which were not observed is highly suspicious according to Jeffreys. The null hypothesis H_0 is rejected in Fisher's significance testing theory when it has failed to predict observable results (the 'more extreme' part of the p-value's definition) which have not occurred. Fisher later accepted this criticism as valid, one of the very few situations in which he admitted problems with his own work:

> "Objection has sometimes been made that the method of calculating Confidence Limits by setting an assigned value such as 1% on the frequency of observing 3 or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly

[18] Note the analogy to the conditionality principle of Cox (1958). As already mentioned in Chapter 3, Fisher was a fervent proponent of conditional inference, which was in line with the conditionality principle that was later established by Cox (1958) and a substantial requirement for the likelihood principle to hold, see Birnbaum (1962). On the other hand, Fisher violated the conditionality principle by using p-values in his significance tests, which incorporate data which was not observed during an experiment. This inconsistency in Fisher's thinking can be seen as another cause of the dissent with Jeffreys.

> the same manner as the value 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation." (Fisher, 1956b, p. 56)

Of course, the argument of the quantity being an approximation seems a weak attempt of saving the practice. Jeffreys (1933) also sharply criticised Fisher's routine derivation of sampling statistics for his tests, where the sampling statistic is the statistic of the quantity of interest averaged over all possible samples. Jeffreys (1933) commented on Fisher's routine derivation of such distributions, which were derived as the average over all samples, that his derivation "...is an absolutely meaningless process. Yet in Fisher's constructive, as well as in his destructive work, this process is carried out again and again." (Jeffreys, 1933, p. 532) In Fisher's defense, at the time Fisher and Gosset developed a variety of precise test statistics in the form of sampling distributions few if anything else was offered as an alternative approach in the contemporary statistical literature. The sampling statistic was a way Fisher could incorporate objectivity into the analysis and overcome the inverse probability approach he objected strongly, but Jeffreys is right in saying that for the situation at hand, using a statistic being averaged over all samples is meaningless. Using a posterior distribution which quantifies the uncertainty of the statistic of interest only based on the actually observed data is more reasonable. Still, as obtaining such a posterior distribution was only possible via the use of prior probabilities, Fisher rejected this option.

Also, as the Haldane-Fisher dispute had already shown, Fisher objected strongly to any priors, and Jeffreys saw them as necessary ingredients to get from sampling alone to a statement about the whole class by the inclusion of the prior information[19]: "I simply state my previous ignorance of the composition, and proceed to consider the consequences of observational data in modifying this ignorance." (Jeffreys, 1934, p. 12)

Interestingly, Egon Pearson and Jerzy Neyman also had a much more eased perspective on priors and inverse probability than Fisher. Pearson (1966) noted, that in line with Jeffreys, for Neyman and him priors were "of no great importance, except in very small samples, where the final conclusions will be drawn in any case with some hesitation." (Pearson, 1966, p. 463).

In summary, the debate between Fisher and Jeffreys took place already before Jeffreys (1935) introduced his Bayesian hypothesis tests as a competitor to Fisher's significance tests. The cause of the dissent was primarily an entirely different concept of probability as well as a different understanding of the problem at hand. Fisher interpreted Jeffreys' Bayesian approach in frequentist terms, leading to a non-compatible solution. Jeffreys argued vice versa, but he hit a weak spot of Fisher when he criticised that significance tests and p-values violated conditional inference. Nevertheless, the clash happened only short after Jeffreys (1931) published his book *Scientific inference*, in which the Bayes factor was already presented.[20] There, Jeffreys (1931) reproduced

[19] Jeffreys even argued that in seismology there are physical reasons which make fitting polynomials of a degree higher than a small value to seismologic data absurd, leading to a prior in favour of small values and approximately zero density at higher values, see Bennett (1990, p. 156).

[20] After showing that Laplace's principle of insufficient reason leads to the paradox that general laws never achieve a large-enough posterior probability, Jeffreys detailed the previous ideas of Wrinch in it. In essence, "The number of possible laws is certainly infinite. How can an infinite number of mutually inconsistent laws all have finite probabilities? The answer to this question is provided by mathematics. Consider the series $\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + ...$ The number of terms in this series is infinite, but every term is finite, the sum of any number of terms is less than unity, and the sum tends to unity as we take an

the earlier introduction of the Bayes factor in Wrinch and Jeffreys (1921) in nearly the same words (underbraces were added for improved readability):

> "The question of the probability to be attached to a quantitative inference can now be dealt with. If p is the most probable law on the data at any stage, and q an additional experimental fact, we have
>
> $$\underbrace{\frac{P(p:q.h)}{P(\sim p:q.h)}}_{\text{posterior odds}} = \underbrace{\frac{P(q:p.h)}{P(q:\sim p.h)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{P(p:h)}{P(\sim p:h)}}_{\text{prior odds}}$$
>
> By the hypothesis we have just made about the prior probabilities of laws, $P(p:h)/P(\sim p:h)$ is not very small."
> Jeffreys (1931, p. 49)

Fisher therefore could have taken notice of Jeffreys's new approach to test hypotheses. Still, as the attention of both men was clearly directed at the probability concept of the other one, Fisher probably did not even read Jeffreys book due to his general objections. After the dissent about the proper concept of probability with Jeffreys, there was no chance that the approaches of both scientists could be reconciled without one of them sacrificing his reputation. Note also that as described by Bennett (1990), the impact of *Scientific Inference* was moderate and only few copies were sold after its publication.

After all, the dispute remained unresolved, and while Jeffrey seemed to have recognised the reason for it, Fisher's writings were sometimes obscure. He went back to business as usual after the dispute, sticking to his complete rejection of inverse probability.

Had the debate happened just a few years later, maybe Fisher's position to the Bayes factor introduced in (Jeffreys, 1935, 1936) would have been available today. From Fisher's objection to Haldane (1932), nothing regarding his opinion on the Bayes factor can be inferred. As described in Section 7.2, Haldane's primary goal was estimation under uncertainty, not hypothesis testing. Therefore, no explicit introduction of the Bayes factor was given by Haldane (1932). One may argue that because of Fisher's strong dissent to combining *any* prior with his likelihood (and connecting his theory with inverse probability, which Haldane (1932) had done), he would have had strong objections to interpreting the likelihood ratio as a Bayes factor.[21] Additionally, the likelihood ratio also was the essential quantity in the Neyman-Pearson tests, which recovered the majority of Fisher's impressive battery of significance tests as special cases under their likelihood ratio criterion λ. It can safely be assumed that Fisher would not have accepted the presence of a Bayes factor in his significance tests, as he once stressed that Jeffreys makes

> "...a logical mistake at the first page which invalidates all the 395 formulae in his book."
> R.A. Fisher in (Box, 1978, p. 441)

increasingly large number of terms from the start. The assumption we need is therefore that the prior probabilities of possible general laws are the terms of a convergent series whose sum to infinity is unity. We have been led to it purely from the assumption that it is possible to construct a theory of quantitative inference; if this can be done such an assumption about the prior probabilities of laws must be made." (Jeffreys, 1931, p. 43)

[21] The Bayes factor is in fact equal to a likelihood ratio whenever the hypotheses under consideration are simple hypotheses, that is, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, compare (Robert, 2007, p. 227).

As described in Chapter 5, the Fisher-Neyman-Pearson dissent was based on the fact that both theories led to different conclusions in some cases. Concerning the dissent with Jeffreys, the objection of Fisher was of much more generality and more profound, since, in contrast to the competing Neyman-Pearson theory, this time inverse probability was involved. The dispute was about the proper probability concept, and not about different solutions to the same problem. Concerning Jeffreys's Bayes factor, Ly (2017) stated that "for many cases the Bayesian and Fisherian analyses disagree qualitatively as well as quantitatively" (Ly, 2017, p. 22). Ly (2017) proposed to use Bayes factors instead of frequentist hypothesis tests based on p-values because of the well-known problems with p-values which were observed during the scientific replication crisis, compare Chapter 1.

Nevertheless, it is a matter of the fact that Jeffreys himself stressed that Fisher had a genuine talent to derive solutions intuitively which were later justified by more rigorous proofs and that the differences between Fisher's approach and the one of Jeffreys rarely differed: "I have in fact been struck repeatedly in my own work, after being led on general principles to a solution of a problem, to find that Fisher had already grasped the essentials by some brilliant piece of common sense, and that his results would either be identical to mine or would differ only in cases where we should both be very doubtful." (Jeffreys, 1939, p. 364-365). Regarding hypothesis testing, Jeffreys in 1939 already used his Bayes factor. Based on the dissent with Fisher about the correct probability concept his statement is astonishing. On the other hand, it is clear that Jeffreys noticed these coincidences, as Fisher's maximum likelihood solutions can be interpreted as the posterior mode in the Bayesian approach under quite general conditions (Held and Sabanés Bové, 2014). Jeffreys probably alluded to these situations in his statement. Concerning hypothesis tests, the differences between Bayesian tests via the Bayes factor and Fisher's significance tests were profound. For example, the latter could only reject a hypothesis, while the former could also confirm it.

Interestingly, Fisher himself in the dispute with Jeffreys once wrote that the correct procedure of scientific inference is that "...we are provided with a definite hypothesis, involving one or more unknown parameters, the values of which we wish to estimate from the data." (Fisher, 1934a, p. 7). It is remarkable in that Fisher states the main goal of scientific inference as estimation under uncertainty, and the hypothesis seems just like a stylistic apparatus needed to formalize the procedure. Despite Fisher's criticism of Haldane (1932), this strongly resembles Haldane's goal of estimation under uncertainty via the posterior distribution of the parameter. Also, this quote shows that probably Fisher would not have thought of any Bayes factor as useful, mainly because all Bayes factors are highly dependent on the priors selected in both models. Also, Bayes factors are not designed with estimation in mind, so that Fisher's primary goal of estimation could not be targeted at all via Bayesian hypothesis tests which employed Bayes factors.

In summary, the Fisher-Jeffreys-debate ended without a resolution of the differences. As McGrayne (2011) noted:

> "Practically speaking, however, Jeffreys lost. For the next decade and for a variety of reasons frequentism almost totally eclipsed Bayes and the inverse probability of causes. First, Fisher was persuasive in public, while the mild-mannered Jeffreys was not: people joked that Fisher could win an argument even when Jeffreys was right. Another factor was that social scientists and statisticians needed objective methods in order to establish themselves as

> academically credible in the 1930s. More particularly, physicists developing quantum mechanics were using frequencies in their experimental data to determine the most probable locations of electron clouds in nuclei. Quantum mechanics was new and chic, and Bayes was not. In addition, Fisher's techniques, written in a popular style with minimal mathematics, were easier to apply than those of Jeffreys. A biologist or psychologist could easily use Fisher's manual to determine whether results were statistically significant."
> McGrayne (2011, p. 57)

The two most severe problems that remained for Bayesian inference to be accepted even after the discourse of Jeffreys and Fisher were thus mostly computational hurdles and the fact that Bayesian statistics was still an undervalued concept which was seen with suspicion by most researchers. How this situation changed is detailed in the next Part III. From an objective perspective however, the various problems of Fisher's significance tests and the Neyman-Pearson theory were outlined in Chapter 5, and these are among the most important causes of the recent scientific replication crisis. Jeffreys noticed the majority of these problems more than half a century ago, and the most important aspect that separated his approach from Fisher's and Neyman and Pearson's as shown in this chapter is given by the fact that

> "Jeffreys was interested in making inferences from scientific evidence, not in using statistics to guide future actions."
> McGrayne (2011, p. 57)

Although in the 1930s Jeffreys' Bayesian approach to statistical hypothesis testing remained widely unheard by practitioners, the main obstacle to employing his theory was eventually overcome by the advent of modern Markov-Chain-Monte-Carlo algorithms, which caused a Bayesian renaissance and are discussed in the following Part III.

INTERMEDIATE CONSIDERATIONS

Part II analysed the evolution of Bayesian approaches to hypothesis testing with a focus on the Bayes factor. It was shown that although the Bayes factor approach did not succeed in the decades that followed, the primary reasons were mostly computational hurdles which prevented a more widespread use of Bayesian methods in scientific research. The core differences between the frequentist and Bayesian approach were analyzed and it was shown that the more appropriate approach for hypothesis testing in scientific contexts is a Bayesian one.

The following Part III discusses the development of modern Markov-Chain-Monte-Carlo algorithms and their impact on Bayesian hypothesis testing. Chapter 8 provides the basics of Markov-Chain-Monte-Carlo (MCMC), and Chapter 9 outlines the Markov-Chain-Monte-Carlo revolution which introduced a Bayesian renaissance from a statistical perspective. It is shown that the development of modern MCMC algorithms has tremendously simplified Bayesian hypothesis testing in practice, and that the largest hurdle in employing Bayesian hypothesis tests has been removed through the advent of modern Markov-Chain-Monte-Carlo methods. Also, it is shown why the burden of manual calibration of MCMC algorithms which presented another obstacle in employing them in practice, has been taken from researchers through the introduction of modern Hamiltonian-Monte-Carlo algorithms. Thus, Part III shows that the computational obstacles which prevented a more widespread use of Bayesian hypothesis tests at the time Jeffreys invented the Bayes factor have been removed through the development of modern MCMC methods.

Part III

The Evolution of Markov-Chain-Monte-Carlo and its Impact on Bayesian Hypothesis Testing

Chapter 8

Markov-Chain-Monte-Carlo

> It was at that time that I suggested an obvious name for the statistical method – a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo."
>
> Nicholas Metropolis
> The Beginning of the Monte Carlo Method

While the Bayesian approach appeals in its simplicity of interpretation and decision-theoretic considerations[1], the computation of the exact posterior troubles application in realistic settings. Therefore, simulation methods based on Markov chains were invented. The goal in Markov-Chain-Monte-Carlo (MCMC) algorithms is to obtain a sample $X_1, ..., X_n$ approximately distributed from the density f because direct simulation from f is not possible. Robert and Casella (2004, p. 268) defines such a method as follows:

Definition 8.1 (Markov-Chain-Monte-Carlo method)**.** A Markov-Chain-Monte-Carlo method for simulation of a distribution f is any method producing an ergodic Markov chain $(X^{(t)})$ whose stationary distribution is f.

All of the algorithms discussed in this chapter rely on the theory of Markov chains, especially ergodicity and stationarity as described in Meyn and Tweedie (2009, Chapter 3, Section 13) and (Robert and Casella, 2004, Chapter 6). In theory, for arbitrary starting values $x^{(0)}$, the chain $(X^{(t)})$ is constructed using a transition kernel with stationary distribution f, which ensures convergence in distribution of $(X^{(t)})$ to f. If the chain is also ergodic, the influence of $x^{(0)}$ vanishes for $t \to \infty$. The invention of Markov-Chain-Monte-Carlo methods had a profound impact on the rediscovery of Bayesian statistics, especially for Bayesian inference in hierarchical models Richey (2010). Some historical remarks about the development of MCMC methods are provided by Diaconis (2009) and Robert and Casella (2008), and for examples of MCMC methods in the context of

[1] Furthermore, as will be shown in Chapter 10 and Chapter 11, the Bayesian approach has a sound basis in philosophy of science and also is strongly mandated by a rigorous axiomatic analysis of the principles of statistical inference.

applied Bayesian hypothesis testing see Kruschke (2015), McElreath (2020) and Kelter (2020c).

8.1 The Metropolis-Hastings-Algorithm

The oldest MCMC algorithm is the Metropolis-Hastings algorithm, which goes back to Metropolis et al. (1953) and which

> "...radically changed our perception of simulation and opened countless new avenues of research and applications."
> Robert and Casella (2004, p. 267)

Moreover, it can be regarded as the most universal MCMC algorithm. It was refined into the slice sampler and Gibbs sampler later. It starts with a target density f, which should be simulated. Therefore, a conditional density $q(y|x)$ with respect to the dominating measure for the model is selected. The only restriction for q is that simulation of q should be (relatively) easy, and it must be explicitly available up to a multiplicative constant independent of x or symmetric, that is $q(x|y) = q(y|x)$. An important requirement for the Metropolis-Hastings algorithm is that the target density f must be available up to a proportionality constant. More specifically, the ratio

$$\frac{f(y)}{q(y|x)} \tag{8.1}$$

must be known up to a constant independent of x. If f therefore is a Bayesian posterior, by Bayes' theorem this is always the case.

8.1.1 The general Metropolis-Hastings algorithm

Algorithm 1 (Metropolis-Hastings). *Given* $x^{(t)}$,

1. *Generate* $Y_t \sim q(y|x^{(t)})$

2. *Take* $X^{(t+1)} := \begin{cases} Y_t, \text{ with probability } p(x^{(t)}, Y_t) \\ x^{(t)}, \text{ with probability } 1 - p(x^{(t)}, Y_t) \end{cases}$

   where

   $p(x,y) := min \left\{ \frac{f(y) \cdot q(x|y)}{f(x) \cdot q(y|x)}, 1 \right\}$

At first sight one might wonder why this innocuous algorithm has caused a Bayesian revival and allowed for the simulation from nearly arbitrary posterior distributions. In Chapter 6 it was detailed that the proportionality constant $1/f(x)$ in the denominator of the posterior distribution in Equation (6.3) can be costly to compute numerically, when no analytic solutions are available. In the Metropolis-Hastings acceptance probability $p(x,y)$, however, this constant cancels out. More specific, in the Bayesian approach when the target density f is the posterior

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{8.2}$$

as in Definition 6.2, the Metropolis-Hastings acceptance probability $p(\theta_1, \theta_2)$ for the parameter θ of interest given the data $X = x$ for the values θ_1 and θ_2 reduces to

$$\begin{aligned} p(\theta_1, \theta_2) &= \min\left\{ \frac{f(\theta_2|x) \cdot q(\theta_1|\theta_2)}{f(\theta_1|x) \cdot q(\theta_2|\theta_1)}, 1 \right\} \\ &= \min\left\{ \frac{\frac{f(x|\theta_2)f(\theta_2)}{\int f(x|\theta)f(\theta)d\theta} \cdot q(\theta_1|\theta_2)}{\frac{f(x|\theta_1)f(\theta_1)}{\int f(x|\theta)f(\theta)d\theta} \cdot q(\theta_2|\theta_1)}, 1 \right\} \\ &= \min\left\{ \frac{f(x|\theta_2)f(\theta_2) \cdot q(\theta_1|\theta_2)}{\underbrace{f(x|\theta_1)}_{\text{likelihood}} \underbrace{f(\theta_1)}_{\text{prior}} \cdot \underbrace{q(\theta_2|\theta_1)}_{\text{proposal dist.}}}, 1 \right\} \end{aligned} \tag{8.3}$$

so that the normalizing constant $\int f(x|\theta)f(\theta)d\theta$ of the posteriors in $p(\theta_1, \theta_2)$ cancels out. As the prior and likelihood are available in a (parametric) Bayesian analysis, the acceptance probability can be computed without the need to calculate the normalizing constant. Thus, steps 1. and 2. in the Metropolis-Hastings algorithm both avoid this computational burden. Also, whenever the *proposal distribution* $q(\theta_2|\theta_1)$ is symmetric, it also cancels out. As $q(\theta_2|\theta_1)$ is explicitly available up to a multiplicative constant $\mathcal{C}$ independent of θ_1, this normalizing constant $\mathcal{C}$ appears in both the numerator and denominator and also cancels out even when q is not symmetric. In both cases therefore, the Metropolis-Hastings algorithm involves no computation of the normalizing constant $\int f(x|\theta)f(\theta)d\theta$ (of the Bayesian posterior) and $\mathcal{C}$ (of the proposal distribution q) in the calculation of the acceptance probability $p(x, y)$. A few points are notable about the Metropolis-Hastings algorithm: First, if the proposal density q is symmetric, the acceptance probability is affected only by the ratio $f(y)/f(x)$ of step 2. Second, $p(x, y)$ is only defined when $f(x^{(t)}) > 0$. If the chain starts in such a value, all subsequent values have also positive mass. Third, by convention $p(x, y) = 0 \Leftrightarrow f(x) \wedge f(y) = 0$. Fourth, the sample generated by Metropolis-Hastings is not i.i.d., and also, there are constraints on the support $\mathcal{E}$ of f and q. When the support of f is connected, the Metropolis-Hastings algorithm works as expected. If it is not, it needs to proceed on one connected component of the support and the different connected components of the support $\mathcal{E}$ must be linked by the kernel of the Metropolis-Hastings algorithm. Also, there are minimal conditions which are necessary for the support for f to be the stationary distribution of the Metropolis-Hastings Markov chain, as detailed in Robert and Casella (2004, p. 272). Most importantly, the following has to hold:

$$\operatorname{supp} f \subset \bigcup_{x \in \operatorname{supp} f} \operatorname{supp} q(\cdot|x) \tag{8.4}$$

Otherwise, stepping into x in a given iteration can result in $q(\cdot|x)$ being not defined there, resulting in a Metropolis-Hastings algorithm which gets stuck in x.

The theoretical justification of the Metropolis-Hastings kernel stems from the fact that it satisfies the detailed balance condition (Robert and Casella, 2004, Def. 6.45) and therefore yields f as the stationary distribution (Robert and Casella, 2004, p. 272):

Theorem 8.2. Let $(X^{(t)})$ be the chain produced by Algorithm 1. For every conditional distribution q whose support includes the support $\mathcal{E}$ of f,

(a) the kernel of the chain satisfies the detailed balance condition with f

(b) f is a stationary distribution of the chain

By use of the Ergodic Theorem – see Robert and Casella (2004, Theorem 6.63) – and under some non-restrictive assumptions about the proposal distribution q – that is, positivity in the form $q(y|x) > 0$ for all $(x, y) \in \mathcal{E} \times \mathcal{E}$ – and allowing for the events $\{X^{(t+1)} = X^{(t)}\}$, irreducibility and aperiodicity of the Metropolis-Hastings Markov chain follow. Irreducibility in turn leads to the chain being recurrent, even Harris recurrent:

Lemma 8.3. If the Metropolis-Hastings chain $(X^{(t)})$ is f-irreducible, it is Harris recurrent.

Irreducibility then justifies the use of the posterior mean as a Bayesian point estimate from posterior distributions obtained by a Metropolis-Hastings chain (Robert and Casella, 2004, p. 274):

Theorem 8.4. Suppose that the Metropolis-Hastings Markov chain $(X^{(t)})$ is f-irreducible.

(i) If $h \in \mathcal{L}^1(f)$, then

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) = \int h(x) f(x) dx \qquad \text{f-a.e.} \tag{8.5}$$

(ii) If in addition $(X^{(t)})$ is aperiodic, then

$$\lim_{n\to\infty} \left|\left| \int K^n(x, \cdot)\mu(dx) - f \right|\right|_{TV} = 0 \tag{8.6}$$

In the above, K^n denotes the Metropolis-Hastings transition kernel. Therefore, under relatively mild assumptions the posterior mean obtained from a Metropolis-Hastings chain is a valid approximation of the mean of the true posterior distribution which is analytically not available, when the number of simulation steps T is large. Then, the total variation norm $||\cdot||_{TV}$ of the difference of the Metropolis-Hastings kernel $K^n(x, \cdot)$ and the stationary distribution f approaches 0 for $n \to \infty$. A somewhat less restrictive condition on f goes back to Roberts and Tweedie (1996):

Lemma 8.5. Assume f is bounded and positive on every compact set of its support $\mathcal{E}$. If there exist positive numbers ε and δ such that

$$q(y|x) > \varepsilon \text{ if } |x - y| < \delta \tag{8.7}$$

then the Metropolis-Hastings Markov chain $(X^{(t)})$ is f-irreducible and aperiodic. Furthermore, every nonempty compact set is a small set.

8.1.2 The independent Metropolis-Hastings algorithm

A modification of the original Metropolis-Hastings algorithm is provided by the independent Metropolis-Hastings algorithm. The difference to the original version is that now the proposal distribution q does not depend on $X^{(t)}$ anymore. For notational convenience, q is now renamed g, and the resulting algorithm is given as follows (Robert and Casella, 2004, p. 276):

Algorithm 2 (Independent Metropolis-Hastings). *Given $x^{(t)}$,*

1. *Generate $Y_t \sim g(y)$*
2. *Take*

$$X^{(t+1)} = \begin{cases} Y_t, \textit{ with probability min} \left(\frac{f(Y_t)g(x^{(t)})}{f(x^{(t)})g(Y_t)}, 1 \right) \\ x^{(t)}, \textit{otherwise} \end{cases} \tag{8.8}$$

The convergence properties of the chain $X^{(t)}$ in Algorithm 2 follow immediately, as $X^{(t)}$ is irreducible and aperiodic (and therefore ergodic), if and only if g is almost everywhere > 0 on the support of f. Stronger results for geometric or uniform convergence can also be established (Robert and Casella, 2004, p. 277):

Theorem 8.6. Algorithm 2 produces a uniformly ergodic chain if there exists a constant M such that

$$f(x) \leq M \cdot g(x), \forall x \in \operatorname{supp} f \tag{8.9}$$

In this case,

$$|| K^n(x, \cdot) - f ||_{TV} \leq 2 \left(1 - \frac{1}{M} \right)^n \tag{8.10}$$

where $|| \cdot ||_{TV}$ denotes the total variation norm. On the other hand, if for every M, there exists a set of positive measure where eq. (8.9) does not hold, $(X^{(t)})$ is not even geometrically ergodic.

In the light of Equation (8.9), it is natural to compare algorithm 2 with a classic Accept-Reject algorithm, as the pair (f, g) could also be used for a simulation via Accept-Reject simulation. Indeed, algorithm 2 dominates the Accept-Reject algorithm in that its expected acceptance probability is at least as high as in a classic Accept-Reject algorithm (Robert and Casella, 2004, p. 278):

Lemma 8.7. If Equation (8.9) holds, the expected acceptance probability associated with Algorithm 2 is at least $\frac{1}{M}$ when the chain is stationary.

This result is only one aspect which visualizes the advantages obtained by MCMC methods over more traditional Monte Carlo methods. Nevertheless, it should be stressed that using MCMC algorithms makes only sense if no direct simulation methods are available. If direct simulation methods are available, these always dominate MCMC algorithms in terms of computational efficiency.

8.1.3 The random walk Metropolis-Hastings algorithm

The original introduction of the Metropolis-Hastings algorithm by Metropolis et al. (1953) used a symmetric *random walk* as a proposal distribution g. The convergence results naturally apply in this case, because by Lemma 8.5, if $g > 0$ in a neighbourhood of 0, the chain becomes irreducible and aperiodic and therefore ergodic. The original Metropolis-Hastings algorithm was formulated this way as follows (Robert and Casella, 2004, p. 288):

Algorithm 3 (Random walk Metropolis-Hastings). *Given $x^{(t)}$,*

1. *Generate $Y_t \sim g(|y - x^{(t)}|)$.*
2. *Take*

$$X^{(t+1)} = \begin{cases} Y_t, \textit{ with probability min} \left\{1, \frac{f(Y_t)}{f(x^{(t)})}\right\} \\ x^{(t)}, \textit{otherwise} \end{cases} \tag{8.11}$$

From a theoretical point of view, the different Metropolis-Hastings algorithms detailed above produce ergodic Markov chains under relatively non-restrictive conditions. Still, they also do rarely enjoy strong ergodicity properties like geometric of uniform ergodicity, for which some results are provided by Robert and Casella (2004, Chapter 7) and Mengersen K. L. and Tweedie (2012). While there are multiple approaches for optimization, including conditioning and tuning the acceptance rate, the universality of the class of Metropolis-Hastings algorithms lies in the nearly non-existent restrictions for the proposal density q. However, this universality comes at the price of computational efficiency, and incorporating more restrictions on the proposal density q can lead to an improved algorithm. This approach leads to the slice sampler.

8.2 The Slice-Sampler

The slice sampler is the first MCMC algorithm which is based on the Metropolis-Hastings algorithm and which exploits the local conditional features of the density f. It can be seen as a generalization of the fundamental theorem of simulation (Robert and Casella, 2004, Theorem 2.15), which is recited below:

Theorem 8.8 (Fundamental Theorem of Simulation). Simulating

$$X \sim f(x) \tag{8.12}$$

is equivalent to simulating

$$(X, U) \sim U\{(x, u) : 0 < u < f(x)\} \tag{8.13}$$

Uniform generation on the subgraph $\varphi(f)$ of f,

$$\varphi(f) := \{(x, u) : 0 \leq u \leq f(x)\} \tag{8.14}$$

no matter what dimension f has, suffices therefore and f needs also only be known up to a normalizing constant.

8.2.1 The 2D Slice sampler

The idea behind the slice sampler is to use a Markov chain with stationary distribution equal to this uniform distribution on $\varphi(f)$. A natural solution is to use a random walk on $\varphi(f)$, which moves iteratively along the coordinate axes. This procedure was proposed by Neal (1997) in a technical report, and published six years later in (Neal, 2003):

Algorithm 4 (2D Slice sampler). *At iteration t, simulate*

1. $u^{(t+1)} \sim \mathcal{U}_{[0,f(x^{(t)})]}$;

2. $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ *with*

$$A^{(t+1)} = \{x : f(x) \geq u^{(t+1)}\} \tag{8.15}$$

An extremely helpful fact in the context of Bayesian hypothesis testing is that the algorithm remains valid if $f = c \cdot f_1(x)$, and f_1 is used instead of f. That is, unnormalized posteriors can be handled in straightforward manner by the slice sampler.

The validity of Algorithm 4 is due to the fact that both steps preserve the uniform distribution on $\varphi(f)$, which is shown by Robert and Casella (2004, Chapter 8). The only major issue with Algorithm 4 is that the simulation of the uniform distribution on $\mathcal{U}_{A^{(t+1)}}$ can be difficult, as the determination of the set of y's such that $f_1(y) \geq \omega$ can be intractable for complex f_1 and given $\omega \in \mathbb{R}$. Therefore, the general slice sampler extends the 2D slice sampler.

8.2.2 The general Slice Sampler

The general slice sampler builds upon the fundamental theorem of simulation, too. It relies upon a decomposition of the density $f(x)$ into components $f_i(x)$ as

$$f(x) \propto \prod_{i=1}^{k} f_i(x) \tag{8.16}$$

where, $f_i(x) > 0$ for all x. In a Bayesian context, these may be individual likelihoods building the complete-sample likelihood. As in the 2D slice sampler, where a single auxiliary variable is used, the general slice sampler uses k auxiliary variables ω_i to write each $f_i(x)$ as

$$f_i(x) = \int \mathbb{1}_{[0,f_i(x)]}(\omega_i) d\omega_i \tag{8.17}$$

and f can be written as the marginal distribution of the joint distribution

$$(x, \omega_1, ..., \omega_k) \sim p(x, \omega_1, ..., \omega_k) \propto \prod_{i=1}^{k} \mathbb{1}_{[0,f_i(x)]}(\omega_i) \tag{8.18}$$

By introducing a larger dimensionality – similar to the 2D slice sampler which uses only one auxiliary variable – the general slice sampler generalizes Algorithm 4 as follows (Robert and Casella, 2004, p. 326):

Algorithm 5 (Slice Sampler). *At iteration $t+1$, simulate*
1. $w_1^{(t+1)} \sim \mathcal{U}_{[0,f_1(x^{(t)})]}$
...
k. $w_k^{(t+1)} \sim \mathcal{U}_{[0,f_k(x^{(t)})]}$
$k+1$. $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$, *with*

$$A^{(t+1)} = \{x : f_i(x) \geq \omega_i^{(t+1)}, i = 1, ..., k\} \tag{8.19}$$

Tierney and Mira (1999) and Roberts and Rosenthal (1997) investigated the convergence properties of the slice sampler, and Tierney and Mira (1999) showed, that if f_1 is bounded and supp f_1 is also bounded, the slice sampler as given in Algorithm 5 is uniformly ergodic. Indeed, the convergence properties of the slice sampler follow from the convergence properties of the *Gibbs sampler*, of which the slice sampler itself is a special case.

8.3 The Gibbs Sampler

The slice sampler is a special case of the class of algorithms called Gibbs samplers, which rely on using the conditional distributions associated with the target distribution f. There are two Gibbs samplers: The two-stage Gibbs sampler, and the general Gibbs sampler. The two-stage Gibbs sampler enjoys even stronger convergence properties than the general Gibbs sampler and applies in a wide variety of settings. The idea behind Gibbs sampling is the same as for slice sampling: Instead of a density $f_X(x)$, a joint density $f(x, y)$ on an arbitrary product-space $\mathcal{X} \times \mathcal{Y}$ is considered. By Theorem 8.8, it suffices to simulate a uniform distribution on

$$\varphi(f) = \{(x, y, u) : 0 \leq u \leq f(x, y)\} \tag{8.20}$$

and use a random walk which moves uniformly in one component at each time step. Starting at a point (x, y, u) in the support of f, generating

1. X along the x-axis on $\mathcal{U}_{\{x:u \leq f(x,y)\}}$
2. Y along the y-axis on $\mathcal{U}_{\{y:u \leq f(x',y)\}}$, where x' is the result of Step 1.
3. U along the u-axis on $\mathcal{U}_{[u:u \leq f(x',y')]}$,, where x' is the result of Step 1 and y' the result of Step 2.

suffices to simulate f. As the sequence of uniform generations does not matter and in the limiting case, simulations along the x and u axes can be repeated several times before moving along the y-axis. In the limiting case of this scenario, where x and u simulations are repeated an infinite number of times before simulating along the y-axis, this is equal to a simulation of $X \sim f_{X|Y}(x|y)$. In the same way, in the limiting case, the simulation of Y and U correspond to a simulation of $Y \sim f_{Y|X}(y|x)$. Simulation of U gets superfluous then, as one is interested in the simulation of $f(x, y)$ rather than the uniform distribution on $\varphi(f)$. If both conditionals $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ can be simulated, the three steps above can be decomposed into the limiting case of two slice samplers, for which stationarity is maintained. Still, the two-stage Gibbs sampler needs knowledge of the conditional distributions in contrast to the 2D slice sampler.

8.3.1 The two-stage Gibbs sampler

While this is not the way the two-stage Gibbs sampler was derived, it is clear that it generates a Markov chain (X_y, Y_t) as follows (Robert and Casella, 2004, p. 339):

Algorithm 6 (Two-stage Gibbs Sampler). *Take $X_0 = x_0$. For $t = 1, 2, ...$ generate*

1. $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$
2. $X_t \sim f_{X|Y}(\cdot|y_t)$

In terms of Algorithm 4, the 2D slice sampler in Algorithm 4 can be interpreted as a special case of the two-stage Gibbs sampler, in which $f(x, y)$ is the uniform distribution on the subgraph $\varphi(f)$. Indeed, the slice sampler starts with $f_x(x)$ and by the introduction of an auxiliary variable creates the joint density $f(x, u) = \mathbb{1}_{(0<u<f_X(x))}$, which is generated artificially in the setting of the slice sampler. For simulation, the slice sampler then uses the conditional densities $f_{X|U}$ and $f_{U|X}$, which are exactly the ones a two-stage Gibbs Sampler as in Algorithm 6 would use. Therefore, the convergence properties of the slice sampler follow by those of the two-stage Gibbs sampler. While there is a multitude of special properties only holding for the two-stage Gibbs sampler, the most important convergence properties follow from the properties of the general Gibbs sampler, of which the two-stage Gibbs sampler itself is again a special case.[2]

8.3.2 The multi-stage Gibbs Sampler

The multi-stage Gibbs sampler generalizes the two-stage Gibbs sampler in the same way the general slice sampler generalizes the 2D slice sampler. While some properties like the interleaving property, Rao-Blackwellization and the Duality principle do not hold for the multi-stage Gibbs sampler, there are still enough optimality properties so that the multi-stage Gibbs sampler can be called the 'workhorse' of MCMC, next to the Metropolis-Hastings algorithm (Robert and Casella, 2004). The derivation is similar to the two-stage case: for $p > 1$, a random variable $X \in \mathcal{X}$ is decomposed as $X = (X_1, ..., X_p)$, where $X_i \in \mathbb{R}$ or $\mathbb{R}^d$. If simulation from the univariate full conditionals $f_1, ..., f_p$, given by

$$X_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_p \sim f_i(x_i|x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_p) \tag{8.21}$$

is possible for $i = 1, ..., p$, the associated multi-stage Gibbs sampler is given by as follows (Robert and Casella, 2004, p. 372):

Algorithm 7 (The Gibbs Sampler). *Given $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, ..., x_p^{(t)})$, generate*

1. $X_1^{(t+1)} \sim f_1(x_1|x_2^{(t)}, ..., x_p^{(t)})$

2. $X_2^{(t+1)} \sim f_2(x_2|x_1^{(t+1)}, x_3^{(t)} ..., x_p^{(t)})$

...

p. $X_p^{(t+1)} \sim f_p(x_p|x_1^{(t+1)}, x_2^{(t+1)} ..., x_{p-1}^{(t+1)})$

[2] Liu et al. (1994) first proved some remarkable structural properties of the two-stage Gibbs sampler, that is, that the marginal chains $(X^{(t)})$ and $(Y^{(t)})$ are reversible and satisfy the *interleaving property*. If reversibility of the subchains matters, Algorithm 6 can easily be adapted to the reversible two-stage Gibbs sampler, as detailed in (Robert and Casella, 2004, Chapter 9). Also, Diebolt and Robert (1994) built upon the work of Liu et al. (1994) and introduced the *duality principle* for interleaving chains. Later, Gelfand and Smith (1990) proposed a technique called *Rao-Blackwellization*, which builds upon Appendix C, Theorem C.55 and the early work of Liu et al. (1994) and Diebolt and Robert (1994).

Like in the two-stage case, only the conditional densities need to be known. Via a completion density, the completion Gibbs sampler is derived, for which most convergence results can be shown. A completion density g of a density f has to satisfy $\int_Z g(x,z)dz = f(x)$. For $p > 1$, rewriting y as $y = (x,z)$ and denoting the conditional densities of $g(y) = g(y_1, ..., y_p)$ as

$$Y_1|y_2, ...y_p \sim g_1(y_1|y_2, ..., y_p) \tag{8.22}$$

$$Y_2|y_1, y_3, ...y_p \sim g_2(y_2|y_1, y_3, ...y_p) \tag{8.23}$$

$$...$$

$$Y_p|y_1, y_3, ...y_{p-1} \sim g_p(y_p|y_1, y_3, ...y_{p-1}) \tag{8.24}$$

the completion Gibbs sampler is given as follows (Robert and Casella, 2004, Chapter 10):

Algorithm 8 (Completion Gibbs Sampler). *Given* $(y_1^{(t)}, ..., y_p^{(t)})$, *simulate*
1. $Y_1^{(t+1)} \sim g_1(y_1|y_2^{(t)}, ..., y_p^{(t)})$
2. $Y_2^{(t+1)} \sim g_2(y_2|y_1^{(t+1)}, y_3^{(t)} ..., y_p^{(t)})$
...
p. $Y_p^{(t+1)} \sim g_p(y_p|y_1^{(t+1)}, ..., y_{p-1}^{(t+1)})$

The two-stage Gibbs sampler in Algorithm 6 is therefore a special case of Algorithm 8, where f is completed in g with x completed as $y = (y_1, y_2)$ where y_1 corresponds to X and y_2 to Y in Algorithm 6, and both conditionals $g_1(y_1|y_2)$ and $g_2(y_2|y_1)$ are available for simulation. The convergence properties of the Gibbs sampler then follow by the following result (Robert and Casella, 2004, Section 10.2.1):

Theorem 8.9. For the Gibbs sampler in Algorithm 8, if $(Y^{(t)})$ is ergodic, then the distribution of g is a stationary distribution for the chain $(Y^{(t)})$ and f is the limiting distribution of the subchain $(X^{(t)})$.

The following Lemma, first proved by Tierney (1994), shows the condition on the Gibbs transition kernel:

Lemma 8.10. If the transition kernel associated with Algorithm 8 is absolutely continuous with respect to the dominating measure, the resulting chain is Harris recurrent.

In total, the condition of absolute continuity with respect to the dominating measure on the Gibbs transition kernel implies by Lemma 8.10 the irreducibility and Harris recurrence of $(Y^{(t)})$. If $(Y^{(t)})$ is also aperiodic, by Robert and Casella (2004, Definition 6.47, Theorem 6.51, Theorem 10.10) the ergodicity of $(Y^{(t)})$ with stationary distribution g follows.
Therefore, ergodicity for the two-stage Gibbs sampler as well as the 2D- and general slice sampler follow immediately. The multi-stage and two-stage Gibbs samplers inherit ergodicity and convergence to the stationary distribution f because they are special cases of the completion Gibbs sampler. The general slice sampler and the 2D slice sampler inherit their convergence behaviour from the multi-stage Gibbs sampler. The multi-stage Gibbs sampler can also be interpreted as a composition of p Metropolis-Hastings kernels, also leading to the convergence properties of the multi-stage Gibbs sampler (Robert and Casella, 2004, p. 381):

Theorem 8.11. The Gibbs sampling method of Algorithm 8 is equivalent to the composition of p Metropolis-Hastings algorithms, with acceptance probabilities uniformly equal to 1.

From a probability theory perspective, the Hammersley-Clifford theorem provides another theoretical justification why the Gibbs sampler works next to the formal results presented above. It highlights that the full conditional distributions suffice to recover the complete information of the joint distribution $f(x, y)$, while the marginal distributions fail to do so:

Theorem 8.12 (Hammersley and Clifford (1971))**.** Under the positivity condition (see Robert and Casella (2004, Definition 9.1), the joint distribution g satisfies

$$g(y_1, ..., y_p) \propto \prod_{j=1}^{p} \frac{g_{l_j}(y_{l_j} | y_{l_1}, ..., y_{l-1}, y'_{l+1}, ..., y'_{l_p})}{g_{l_j}(y'_{l_j} | y_{l_1}, ..., y_{l-1}, y'_{l+1}, ..., y'_{l_p})}$$

for every permutation l on $\{1, 2, ..., p\}$ and every $y \in \mathcal{Y}$.

Chapter 9

The Evolution of Markov-Chain-Monte-Carlo

Even the simplest of numerical methods may yield spurious results if insufficient care is taken in their use, and how difficult it often is to assess the magnitude of the errors. The discussion above indicates that the situation is certainly no better for the Markov chain methods and that they should be used with appropriate caution.

Wilfred Keith Hastings
Monte Carlo Sampling Methods Using Markov Chains and Their Applications

Chapter 3, Chapter 4 and Chapter 5 described the evolution of frequentist hypothesis testing by following the early beginnings of Fisher's Significance Testing to the alternative Neyman-Pearson-Theory and the hybrid approach which is commonly used today and evolved out of both theories. On the other hand, Chapter 6 and Chapter 7 introduced Bayesian statistics and the evolution of the Bayes factor as an alternative to the frequentist theories of hypothesis testing. The last chapter outlined Markov-Chain-Monte-Carlo methods, which have historically provided a significant simplification of Bayesian inference in practice, as they allowed to obtain a posterior distribution numerically instead of analytically. While there is a spectrum of statistical models for which analytic inference via Bayes' theorem – see Held and Sabanés Bové (2014) or Marin and Robert (2014) for an overview – is possible, the majority of complex and hierarchical statistical models in realistic applications escapes an analytical treatment. As a consequence, no closed-form expressions can be derived for the posterior distribution of the parameters of interest. As Chapter 8 already detailed, this is where Markov-Chain-Monte-Carlo methods are needed and have shown to be a highly efficient way to obtain posterior distributions of previously not tractable statistical models. In fact, MCMC algorithms also offered solutions to some substantial problems in frequentist inference and hypothesis testing:

"When we leave the exponential family setup, we face increasingly challeng-

> ing difficulties in using maximum likelihood techniques. One reason for this is the lack of a sufficient statistic of fixed dimension outside exponential families[1], barring the exception of a few families such as uniform or Pareto distributions whose support depends on θ (Robert 2001, Section 3.2). This result, known as the Pitman-Koopman-Lemma (see Lehmann and Casella 1998, Theorem 1.6.18), implies that, outside exponential families, the complexity of the likelihood increases quite rapidly with the number of observations, n and thus, that its maximization is delicate, even in the simplest case."
> Robert and Casella (2004, p. 10)

MCMC methods could easily be applied to such problems and did not suffer from similar problems when a suitable prior distribution was chosen on the model parameters.

> "Similar computational problems arise in the determination of the power of a testing procedure in the Neyman-Pearson approach (see Lehmann 1986, Casella and Berger 2001, Robert 2001). For example, inference based on a likelihood ratio statistic requires the computation of quantities such as
>
> $$P_\theta(L(\theta|X)/L(\theta_0|X) \leq k), \tag{9.1}$$
>
> with fixed θ_0 and k, where $L(\theta|x)$ represents the likelihood based on observing $X = x$. Outside of the more standard (simple) settings, this probability cannot be explicitly computed because dealing with the distribution of test statistics under the alternative hypothesis may be quite difficult."
> Robert and Casella (2004, p. 12)

In the above cases, the posterior can be obtained numerically via MCMC with nearly arbitrary (instead of conjugate) priors. Maximisation is then achieved by computing the posterior mode from distribution of the Markov chain samples.

Hypothesis testing from a Bayesian perspective traditionally relied strongly on the calculation of the Bayes factor as detailed in Chapter 7. However, with the advent of Markov-Chain-Monte-Carlo techniques, next to the derivation of posteriors in previously untractable statistical models, new possibilities for testing hypotheses in a Bayesian manner emerged. The introduction of MCMC, therefore, elevated Bayesian inference onto the same level of applicability as frequentist estimation and hypothesis testing. Recent developments – especially the availability of highly capable computing resources and probabilistic programming languages and MCMC samplers like JAGS (Plummer, 2003) and STAN (Carpenter et al., 2017) – have made the Bayesian approach much more popular and accessible as detailed in Chapter 1. This chapter reconstructs the milestones of the evolution of Markov-Chain-Monte-Carlo methods and shows that these methods have opened the door to obtaining a Bayesian posterior in an algorithmic fashion. Furthermore, they have allowed to perform Bayesian hypothesis testing in the same manner for nearly arbitrary statistical models[2] In cases where no analytic derivation of

[1] See (Rüschendorf, 2014, Chapter 4) for a simple proof.

[2] We do not discuss this point in detail in this thesis, as this is outside of the scope of the main text, but notice that examples where MCMC methods have considerably eased the application of Bayesian hypothesis testing are bridge sampling (Gronau et al., 2017, 2019) and the Savage-Dickey density ratio (Dickey and Lientz, 1970; Verdinelli and Wasserman, 1995; Wagenmakers et al., 2010), which is used to obtain the density under the alternative H_1 via MCMC samples and subsequently calculate the Bayes factor. For other approaches to Bayesian hypothesis testing based on MCMC see Kelter (2020a,e), Pereira and Stern (2020) and Makowski et al. (2019a) as well as Chapter 14.

the Bayes factor was possible before the availability of MCMC methods, this presents a definite plus for the applicability of Bayesian methods to hypothesis testing.

9.1 An Overview of the Evolution of Markov-Chain-Monte-Carlo

The MCMC algorithms introduced in Chapter 8 root back to the invention of the original MCMC method, introduced in 1953 by Metropolis et al. (1953). The invention of MCMC starts with this single publication, and there are no previous papers on which the ideas introduced in Metropolis et al. (1953) are built. While in frequentist statistical inference, a few people, namely Ronald Fisher, Jerzy Neyman and Egon Pearson can be attributed as the driving forces, the development of MCMC methods was achieved by various people. The cornerstone was laid in Los Alamos after World War II by Metropolis et al. (1953), and it took a long time until the ideas developed in the 1950s were rediscovered and refined. There are five milestones, which can be seen as a chain of succeeding developments leading to the modern theory of MCMC as it is available today:

1. The birth hour of MCMC: *Equations of State Calculations by Fast Computing Machines*, published in 1953 by Metropolis et al. (1953), where the famous Metropolis-algorithm (see Algorithm 3) was introduced.

2. The formal justification of the original Metropolis algorithm: Hastings (1970) publication of the paper *Monte Carlo Sampling Methods Using Markov Chains and Their Applications* in 1970.

3. Thirty years after the publication of the original Metropolis algorithm, Kirkpatrick et al. (1983) published their paper *Optimization by Simulated Annealing*, which introduced simulated annealing.

4. One year later, in 1984 the brothers Geman and Geman (1984) published the paper *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, which introduced the Gibbs-sampler (see Algorithm 7).

5. Finally, in 1990, Gelfand and Smith (1990) wrote the paper called *Sampling-Based Approaches to Calculating Marginal Densities*, in which the Gibbs sampler was presented as a general-purpose inference tool in a purely statistical context.

While there are multiple other papers which also attributed to the development of MCMC – for example the later introduction of the slice sampler as given in Algorithm 5 by Neal (1997) and Neal (2003) – the foundations were developed in the above five papers, and their impact can hardly be overstated. According to Google Scholar[3], the paper of Metropolis et al. (1953) was cited 38827 times, Hastings (1970) paper was cited 13168 times, the paper of (Kirkpatrick et al., 1983) 43035 times, the paper of (Geman and Geman, 1984) 22277 times and the paper of Gelfand and Smith (1990) 7803 times, so that in total the four papers above were cited 86380 times, which indicates the huge impact they had. Also, the Metropolis-Hastings-Algorithm was titled one of the ten most important algorithms invented in the whole century (Dongarra and Sullivan,

[3] Data obtained at the first of April, 2019.

2000).

Much has been written about the early work at Los Alamos, especially about the first Monte Carlo approaches by Anderson (1986), Hitchcock (2003), Gubernatis (2005) and Metropolis (1987) himself. Also, the emerging of the Monte Carlo method from the need for particle physics simulations for the development of the nuclear bomb has been detailed by Harlow and Metropolis (1983). Robert and Casella (2008); Robert (2015) provides a rough overview about the purpose and historical aspects of Markov-Chain-Monte-Carlo, leaving out the modern developments like Hamiltonian Monte Carlo, and also leaving out a discussion of the possibilities of Markov-Chain-Monte-Carlo methods with regards to Bayesian hypothesis testing. Richey (2010) gives an overview about the development of MCMC methods in general and a detailed account of the introduction of simulated annealing into the statistical community, but also does not address MCMC as a door-opener to Bayesian hypothesis testing.

This chapter, therefore, focusses on the milestones described above. In the following sections, these main developments are reconstructed and, in particular, the resulting consequences for Bayesian hypothesis testing are discussed.

9.2 The Introduction of the original Metropolis Algorithm

The Metropolis-algorithm was born out of the Monte Carlo methods developed and utilised in Los Alamos as described by Metropolis (1987). The development of nuclear weapons was one of the central goals of the applied physical research in Los Alamos, and one problem of interest constituted itself in the estimation of the behaviour of large particle collections inside nuclear weapons. As the physical laws describing the behaviour of such particles are often probabilistic, like the probability of an electron being located in a particular molecular orbital or suborbital, traditional analytical methods did not suffice to derive any useable results. Metropolis (1987) noted, that at that time, the idea of simulating the state of a complex system like that of a large collection of particles emerged in the form of the Monte Carlo method, but that even the simulation was a difficult task, as computing power was rarely available and slow. Nevertheless, the situation was considerably better than before World War II, when no computing power at all was available, so that simulation of complex systems was at least conductible without having to perform handwritten calculations. The idea of using simulation to solve such problems can be attributed to Ulam Stan, who wrote about that time years later (Ulam, 1991). Also, Eckhardt (1987) gave a recollection of the early Monte Carlo ideas in Los Alamos. In 1946, Stanislaw Ulam had to stay at a hospital for a few days and often played Solitaire for his entertainment with his visitors (Ulam, 1991). In the version Ulam played, once the cards were dealt the outcome of the game was completely determined, and he wondered about the probability of winning. The combinatorial problem was of a much too large scale because of the number of cards, and he came up with the idea of programming the ENIAC computer in Los Alamos to simulate the outcome of a game. Therefore, first, the shuffling of cards needed to be simulated, and subsequently, the rules of the solitaire version needed to be applied[4]. While Stanislaw

[4] ENIAC was an acronym for *Electronic Numerical Integrator and Computer*. The ENIAC was the first electronic general-purpose computer which was Turing-complete. Its original purpose was to calculate ballistic tables for the military and not combinatorial problems in card games.

Ulam never pursued this idea for the solitaire game, he proposed his approach for the simulation of particles at his work for the nuclear weapon research in Los Alamos to his good friend John von Neumann. One year later, in 1947, von Neumann and his colleagues were indeed working on the estimation of neutron diffusion and multiplication rates in nuclear fission with a particular interest in nuclear weapons (Eckhardt, 1987). Together with Ulam, von Neumann proposed to follow the idea of randomly simulating a large number of neutrons and their evolution over time at nuclear fission. After stopping the simulation, they counted the number of electrons which remained to estimate the rates of interest. Richey (2010) noted that the speed of simulation was extremely slow. In essence, Stanislaw Ulam and John von Neumann were able to simulate 100 neutrons with 100 collisions which took about five hours for the simulation to complete on the ENIAC. Nevertheless, the approach was a methodologic breakthrough. It combined the slowly emerging computing power with the century-old strong law of large numbers to obtain posterior expectations of quantities of interest. From this point on, randomized simulations became an important new technique. Metropolis (1987) recalled that the name Monte Carlo method was his idea:

> "It was at that time that I suggested an obvious name for the statistical method – a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo."' (Metropolis, 1987, p. 127)

Two years later, in 1949, Metropolis and Ulam published the ideas in a joint paper and noted, that computing machines are "extremely well suited to perform the procedures described." (Metropolis and Ulam, 1949, p. 339). They outlined their ideas in the 1949 paper, but only four years later, the seminal paper introducing the original Metropolis-Hastings algorithm was published (Metropolis et al., 1953).

The introduction of the algorithm by Metropolis et al. (1953) was motivated partially by the goal to infer properties of the well-known Boltzmann distribution used in statistical mechanics, where, in particular, the average behaviour of large particle systems was of interest. A detailed account of the ideas behind the Boltzmann distribution can be found in (Richey, 2010), and here only the most important points are detailed. In the 1950s, the scientific branch of statistical mechanics tried to describe a collection of particles by a *configuration* ω, with $\omega \in \Omega$, the configuration space. In the usual scenario, a finite set of N particles is given, and every particle is modelled using its position and velocity, each in three-dimensional space. The dimension of the configuration space Ω follows as $\dim(\Omega) = 6 \cdot N$ because Ω is a subset of $\mathbb{R}^{6 \cdot N}$. Another common approach is to model the particle system by assigning ± 1 to every grid point of the integer lattice in the plane, resulting in a bounded subset of Ω. The state $+1$ could indicate the presence of a particle at the grid point of the lattice, and -1 the absence, thus describing the motion of the whole particle system dependent on time. The dimension of the configuration space then reduces to $\dim(\Omega) = 2^N$. In statistical mechanics, one uses an energy function $E : \Omega \longrightarrow \mathbb{R}_+$ to model quantities like potential energy in the continuous case. The particle system is modelled in its equilibrium state via the relative frequency of a configuration ω in form of its so-called Boltzmann weight $e^{-E(\omega)/k \cdot T}$, with the *temperature* T and the *Boltzmann's constant* k. The *Boltzmann distribution* then uses the following *Boltzmann probability* to model the probability of the particle system being in a specific configuration:

$$B(\omega) = \frac{e^{-E(\omega)/k\cdot T}}{\sum_{\tilde{\omega}\in\Omega} e^{-E(\tilde{\omega})/k\cdot T}} \tag{9.2}$$

Richey (2010) noted that for any realistic setting, the denominator in Equation (9.2) is analytically intractable and no closed-form expression is available. In statistical physics, quantities like the total energy of the particle system

$$\langle E \rangle := \sum_{\omega\in\Omega} E(\omega)\cdot B(\omega) = \frac{\sum_{\omega\in\Omega} e^{-E(\omega)/k\cdot T} E(\omega)}{\sum_{\tilde{\omega}\in\Omega} e^{-E(\tilde{\omega})/k\cdot T}} \tag{9.3}$$

are of interest, which in a statistical interpretation are often ordinary an expectation. Standard Monte Carlo techniques which were just invented at 1947, could be utilised to randomly generate $\omega_1, ..., \omega_K$ with $K \in \mathbb{N}$ uniformly on Ω (that is, $\omega_i \sim \mathcal{U}(\Omega)$), and approximate Equation (9.3) via the use of the empirical mean. The validity follows from the strong law of large numbers. One particular drawback of this approach was that the random sampling procedure included also configuration states $\tilde{\omega}$ with a very small Boltzmann probability $B(\tilde{\omega})$ in the approximation of the expectation in Equation (9.3). This is what Metropolis et al. (1953) addressed in the paper introducing the Metropolis-Hastings algorithm. They considered a square of N particles, where each cell's value in the square either indicated the presence or absence of a particle:

> "Thus the most naive method of carrying out the integration would be to put each of the N particles at a random position in the square (this defines a random point in the 2^N-dimensional configuration space), then calculate the energy of the system according to Eq. (1), and give this configuration a weight $exp(-E/kT)$. This method, however, is not practical for close-packed configurations, since with high probability we choose a configuration where $exp(-E/kT)$ is very small; hence a configuration of very low weight. So the method we employ is actually a modified Monte Carlo scheme, where, instead of choosing configurations randomly, then weighting them with $exp(-E/kT)$, we choose configurations with a probability $exp(-E/kT)$ and weight them evenly."
> (Metropolis et al., 1953, p. 1088)

To follow this strategy, Metropolis et al. (1953) needed to be able to simulate random numbers ω_i from the Boltzmann distribution $B(\omega)$, instead of the uniform distribution on a compact set. Mathematically, instead of $\omega_i \sim \mathcal{U}(\Omega)$, Metropolis et al. (1953) needed to simulate $\omega_i \sim B(\omega_i)$. As the denominator of Equation (9.2) – also called the *partition function* – is not available in realistic settings, direct simulation of random numbers from the Boltzman distribution was not possible and this constituted the main challenge of the entire procedure. The trick of the Metropolis-Hastings algorithm was to construct a Markov chain which has the Boltzmann distribution as stationary distribution. The crucial property of the algorithm is that it does only need the Boltzmann weights, that is, the numerator of Equation (9.2). After stating their idea, Metropolis et al. (1953) introduced the algorithm as the solution to obtaining random numbers from the Boltzmann distribution as follows: They started with a finite configuration space Ω and an energy function E with temperature T, where T was fixed. Augmenting the configuration space Ω then leads to a (possibly larger) space $\hat{\Omega}$, where $\hat{\Omega}$ is a sample of configurations $\omega \in \Omega$ selected with replacement. Metropolis et al. (1953) then followed the strategy to add and remove configurations ω from $\hat{\Omega}$ until the sample $\hat{\Omega}$ approximately became a sample from the Boltzmann distribution. By doing so, Metropolis et al. (1953) then derived the now well-known detailed balance condition (see Robert and Casella (2004, Def. 6.45)), which needs to be fulfilled for Markov chains to converge to the correct stationary distribution. Suppose that $|\hat{\Omega}| = \hat{N}$ and the number of occurrences of $\omega \in \hat{\Omega}$ is denoted as $\hat{N}_\omega$. If the sample can be interpreted as being one from the Boltzmann distribution, Metropolis et al. (1953) argued, that then

$$\frac{\hat{N}_\omega}{\hat{N}} \propto \exp\left(-E(\omega)/k \cdot T\right) \tag{9.4}$$

which holds if and only if

$$\frac{\hat{N}_{\omega'}}{\hat{N}_\omega} = \frac{\exp\left(-E(\omega')/k \cdot T\right)}{\exp\left(-E(\omega)/k \cdot T\right)} = \exp\left(-\Delta E/k \cdot T\right) \tag{9.5}$$

for two configurations ω, ω' and $\Delta E := E(\omega') - E(\omega)$. Metropolis et al. (1953) then introduced an irreducible, aperiodic Markov chain on Ω with a symmetric kernel $P_{\omega,\omega'}$ to model the moves between two configurations ω, ω'. Because of the symmetry of the kernel, $P_{\omega,\omega'} = P_{\omega',\omega}$ holds for all $\omega, \omega' \in \Omega$. This Markov chain was called the proposal transition. Metropolis et al. (1953) reasoned, that for configurations ω, ω' with $E(\omega) < E(\omega')$ transitions $P_{\omega',\omega}$ from ω' to ω should always be allowed, as the particle system then moves to more balanced states in terms of energy. That means, the probability $\mathbb{P}(\omega', \omega)$ of moving (or transitioning from state ω' to state ω) is one in this case: $\mathbb{P}(\omega', \omega) = 1$. They denoted the number of occurrences of such transitions as $P_{\omega',\omega} \cdot \hat{N}_{\omega'} \mathbb{P}(\omega', \omega) = P_{\omega',\omega} \cdot \hat{N}_{\omega'}$. Here, $\hat{N}_{\omega'}$ is the number of occurrences of $\omega' \in \hat{\Omega}$. $P_{\omega',\omega}$ is the probability that the Markov kernel proposes to move from ω' to ω, so that the product $\hat{N}_{\omega'} \cdot P_{\omega',\omega}$ is the number of transition proposals from ω' to ω. Lastly, $\mathbb{P}(\omega', \omega)$ is the acceptance probability of such proposals. Metropolis et al. (1953) argued further, that to fulfill the condition in Equation (9.4), also moves from ω to ω' with $E(\omega) < E(\omega')$ need to be allowed with a specific probability $\mathbb{P}(\omega, \omega')$. The number of those moves being allowed is calculated analogue to the above as $P_{\omega,\omega'} \hat{N}_\omega \mathbb{P}(\omega, \omega')$. Then, they calculated the difference in the total number of transitions from ω to ω' and ω' to ω as

$$\begin{align}
P_{\omega',\omega}\hat{N}_{\omega'} - P_{\omega,\omega'}\hat{N}_\omega \mathbb{P}(\omega,\omega') &\overset{(1)}{=} P_{\omega,\omega'}\hat{N}_{\omega'} - P_{\omega,\omega'}\hat{N}_\omega \mathbb{P}(\omega,\omega') \tag{9.6}\\
&= P_{\omega,\omega'}(\hat{N}_{\omega'} - \hat{N}_\omega \mathbb{P}(\omega,\omega')) \tag{9.7}\\
&\overset{(2)}{=} P_{\omega,\omega'}(\hat{N}_\omega \cdot \exp(-\Delta E/k \cdot T) - \hat{N}_\omega \mathbb{P}(\omega,\omega')) \tag{9.8}
\end{align}$$

wherein (1) the symmetry of the kernel P and in (2) the fact that $\hat{N}_{\omega'} = \hat{N}_\omega \cdot \exp(-\Delta E/k \cdot T)$ resulting from Equation (9.5) was used. If Equation (9.5) holds, the distribution of energy matches the Boltzmann distribution of energy, and then the flow of energy in Equation (9.6) should be zero. This immediately implies that due to Equation (9.8) $\mathbb{P}(\omega, \omega') = \exp(-\Delta E/k \cdot T)$ needs to hold. Metropolis et al. (1953) argued that this probability of occasional moves to configurations with higher energy in total guarantees that the Markov chain converges to the stationary distribution, the Boltzmann distribution. They did not deliver a mathematical proof, but their argument is convincing and went as follows: They assumed there were too many configurations with high energy $E(\omega')$ compared to configurations with low energy $E(\omega)$, that is $\hat{N}_{\omega'}/\hat{N}_{\omega} > \exp(-\Delta E/k \cdot T)$. The total number of transitions between configurations ω with energy $E(\omega)$ and configurations ω' with energy $E(\omega')$ of energy given in Equation (9.6) is positive then, and this means the number of incoming transitions from ω' to ω is larger than the number of outgoing transitions from ω to ω'. Therefore, there will be more transitions from configurations ω' with Energy $E(\omega')$ to states ω with energy $E(\omega)$ than reversed transitions from ω to ω' with corresponding energy values $E(\omega)$ and $E(\omega')$. This implies that the inequality $\hat{N}_{\omega'}/\hat{N}_{\omega} > \exp(-\Delta E/k \cdot T)$ will move a step towards becoming equality, and the distribution of energies in $\hat{\Omega}$ will move a step towards the Boltzmann distribution. If this step is repeated, in the long run, the whole process generates a distribution of energies which converges to the Boltzmann distribution. Starting with the assumption of too many configurations, ω with low energy $E(\omega)$ yields the same conclusion. Based on these ideas, the original Metropolis algorithm was then given as follows:

Algorithm 9 (The original Metropolis algorithm). *For $\omega \in \Omega$, the transition to a configuration ω^* is defined as follows:*

1. *From an arbitrary proposal transition, select ω'.*
2. *A) If the energy $E(\omega') < E(\omega)$, that is, if $B(\omega') \geq B(\omega)$, let $\omega^* = \omega'$.*
   B) If the energy $E(\omega') > E(\omega)$, that is, if $B(\omega') < B(\omega)$, let $\omega^* = \omega'$ with probability

   $$\frac{B(\omega')}{B(\omega)} = exp(-\Delta E/k \cdot T)$$

   and else let $\omega^* = \omega$.

While Metropolis et al. (1953) were satisfied by their intuition, they noted that the rate of convergence remained unknown. The first formal proofs of the convergence of the algorithm were given by Hammersley and Handscomb (1964) and Hastings (1970), but Metropolis et al. (1953) already took notice of the remarkably important fact that the proposal parameter – in their original notation α – needed to be chosen carefully for the algorithm to run efficiently:

> "The above argument does not, of course, specify how rapidly the canonical distribution is approached. It may be mentioned in this connection that the maximum displacement α must be chosen with some care; if too large, most moves will be forbidden, and if too small, the configuration will not change enough. In either case it will then take longer to come to equilibrium."
> Metropolis et al. (1953, p. 1089)

To tune the Metropolis algorithm with just the right proposal parameter values became one of the major challenges of MCMC later, in particular in high-dimensional models (Robert, 2015).

The original Metropolis algorithm as given in Algorithm 9 has some major advantages, which are the reason it was possible to employ it for simulation of random numbers from the previously untractable Boltzmann distribution. First, it does not need the denominator of Equation (9.2), which is not available in most realistic situations (compare Chapter 6). Second, it defines the irreducible, aperiodic Markov chain on the state space without specifying the whole transition kernel. This allowed for straightforward application of the algorithm even in large-dimensional spaces where specifying the transition kernel would quickly become effortful. Third, the computational steps are mostly elementary. The first application of the algorithm in the original 1953 paper of Metropolis et al. (1953) was an analysis of the *hard spheres model*, a physical model of molecules which do not overlap (for example, a gas):

> "We set up the calculation on a system composed of N = 224 particles (i=0, 1, .. 223) placed inside a square of unit side and unit area. The particles were arranged initially in a trigonal lattice of fourteen particles per row by sixteen particles per column, alternate rows being displaced relative to each other ..."
> Metropolis et al. (1953, p. 1090)

After running the algorithm in this 224-dimensional space, the simulated results obtained by Metropolis et al. (1953) matched closely the analytical results which were available by employing more traditional methods. Also, the calculation time was moderate, where moderate means that a single calculation, of which there were hundreds to conduct, took about four to five hours on Los Alamos' MANIAC computer.

9.3 The Introduction of the Metropolis-Hastings Algorithm

After the introduction of the original Metropolis algorithm in 1953, mathematicians and statisticians took little notice of the method. While the application in the area of statistical mechanics was promising, few if any researchers had access to computing resources at that time, so that application was simply not possible for the majority of scientists. Also, the context was quite specific and masked the universality the algorithm provided for solving statistical problems. In particular, the algorithm had no apparent relationship to Bayesian statistics. Furthermore, the 1950s were a decade in which different statistical areas were topics of interest, see for example Cox (1958); Fisher (1950, 1955). This situation did not change until the 1980s, even though in 1970, Hastings (1970) published a generalisation of the original Metropolis algorithm. The only interest in Monte Carlo methods from statisticians can be found in Hammersley and Handscomb (1964), who included a short section about MCMC algorithms in their monograph on Monte Carlo methods.

At the same time, physicists started to use the Metropolis algorithm for applications to the *Ising model* and different other spin models. A prominent example is found in (Glauber, 1963), who used the Metropolis algorithm to simulate the sequential movement through lattice sites, where at the ith site, the spin ω_i is set according to the (local) Boltzmann weight $\mathbb{P}(\omega_i = s) = \exp(-s \sum_{\langle i,j \rangle} \omega_j)/k \cdot T$ with the summation being over the nearest neighbor sites $\langle i, j \rangle$ of i. Other application in the branch of statistical mechanics can be found in (Barker, 1969) and (Flinn, 1974). Interestingly, Barker (1969) constructed a different Markov chain than the original one proposed by Metropolis et al. (1953), so that the question arose, how many Metropolis-like algorithms do exist, and if there is any *best* Metropolis algorithm. The first answer to this problem came up with the seminal paper of Hastings (1970), who introduced not only a formal proof of convergence of the original Metropolis algorithm but also a generalisation. Luckily, there is an interview with Hastings himself available in (Rosenthal, 2005). In it, Hastings recalled:

> "When I returned to the University of Toronto, after my time at Bell Labs, I focused on Monte Carlo methods and at first on methods of sampling from

> probability distributions with no particular area of application in mind. [University of Toronto Chemistry professor] John Valleau and his associates consulted me concerning their work. They were using Metropolis's method to estimate the mean energy of a system of particles in a defined potential field. With 6 coordinates per particle, a system of just 100 particles involved a dimension of 600. When I learned how easy it was to generate samples from high dimensional distributions using Markov chains, I realised how important this was for Statistics, and I devoted all my time to this method and its variants which resulted in the 1970 paper."
> Hastings (2005), in (Rosenthal, 2005)

Hastings paper was a quantum leap as it allowed for the simulation of probability distributions which were untractable previously. In it, he uncoupled the original Metropolis algorithm from its context of statistical mechanics and applied it to a variety of standard distributions to simulate from. He provided a generalisation of the original algorithm which included both the algorithm of Metropolis et al. (1953) and Barker (1969) as special cases. Therefore, the original Metropolis algorithm is often also called the Metropolis-Hastings algorithm. Hastings (1970) started with the goal to sample from a distribution π. After selecting a proposal transition $Q = (q_{ij})$ on the state space Ω as in (Metropolis et al., 1953), Hastings (1970, p. 100) defined the transition matrix $P = (p_{ij})$ of the Markov chain as

$$p_{ij} = \begin{cases} q_{ij} \cdot \alpha_{ij}, \text{ if } i \neq j \\ 1 - \sum_{k \neq i} p_{ik}, \text{ if } i = j \end{cases} \qquad \text{with } \alpha_{ij} = \frac{s_{ij}}{1 + \frac{\pi_i}{\pi_j}\frac{q_{ij}}{q_{ji}}} \tag{9.9}$$

In contrast to the original Metropolis algorithm, the proposal transition Q did not need to be symmetric, and the values s_{ij} needed only to fulfill the conditions $s_{ij} = s_{ji}$ for all i, j and $\alpha_{ij} \in [0, 1]$. Hastings (1970) then noted, that

> "Two simple choices for s_{ij} are given for all i and j by
>
> $$s_{ij}^M = \begin{cases} 1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}} (\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1), \\ 1 + \frac{\pi_j q_{ji}}{\pi_i q_{ij}} (\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \leq 1) \end{cases}$$
>
> $$s_{ij}^B = 1$$
>
> With $q_{ij} = q_{ji}$ and $s_{ij} = s_{ij}^M$ we have the method devised by Metropolis et al. (1953) (...)"
> Hastings (1970, p. 100)

In Hasting's original notation, $1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}} (\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1)$ can be read as $1 + \frac{\pi_i q_{ij}}{\pi_j q_{ji}}$, if $(\frac{\pi_j q_{ji}}{\pi_i q_{ij}} \geq 1)$. To show that indeed the original Metropolis algorithm (and the one of Barker (1969)) are recovered, Hastings (1970) noted that α_{ij}^M corresponding to s_{ij}^M is given as follows:

> "(...) when $q_{ij} = q_{ji}$, we have
>
> $$\alpha_{ij}^M = \begin{cases} 1, (\pi_j / \pi_i \geq 1), \\ \pi_j / \pi_i (\pi_j / \pi_i < 1), \end{cases}$$
>
> "
> Hastings (1970, p. 100)

which follows directly from the previous definition of s_{ij}^M and Equation (9.9). After these derivations, Hastings (1970) stated what is known today better as the *Metropolis-Hastings acceptance probability*:

> "More generally, we may choose
>
> $$s_{ij} = g[\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\}]$$
>
> where the function $g(x)$ is chosen so that $0 \leq g(x) \leq 1 + x$ for $0 \leq x \leq 1$ and $g(x)$ may itself be symmetric in i and j."
> Hastings (1970, p. 100)

It is straightforward to use the last definition of s_{ij} together with Equation (9.9) to derive the acceptance probability in Algorithm 2. Therefore, Hastings (1970) noted that

> "For example, we may choose $g(x) = 1 + 2(\frac{1}{2}x)^\gamma$ with the constant $\gamma \geq 1$, obtaining $s_{ij}^{(M)}$ with $\gamma = 1$ and $s_{ij}^{(B)}$ with $\gamma = \infty$."
> (Hastings, 1970, p. 100)

For this selection of g in the original Metropolis algorithm, under the assumption (without loss of generality) of $\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\}) = (\pi_i q_{ij})/(\pi_j q_{ji})$, the quantity s_{ij} becomes

$$1 + 2(\frac{1}{2}\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\})^1 = 1 + (\pi_i q_{ij})/(\pi_j q_{ji})$$

exactly if $\frac{(\pi_j q_{ji})}{(\pi_i q_{ij})} > \frac{(\pi_i q_{ij})}{(\pi_j q_{ji})}$. This is equivalent to $\pi_j q_{ji} > \pi_i q_{ij} \Leftrightarrow \frac{\pi_j q_{ji}}{\pi_i q_{ij}} > 1$, which is the original notation in $s_{ij}^{(M)}$ of Hastings (1970).[5] Thus, the resulting acceptance probability is α_{ij}^M above, which is precisely the acceptance probability of Algorithm 2 when g is symmetric, and which is precisely the acceptance probability of the random walk Metropolis-Hastings algorithm given in Algorithm 3.

Hastings' student Peskun (1973) showed three years later that among all choices for the quantities s_{ij}, the special case leading to the original Metropolis algorithm was indeed optimal. It asymptotically leads to the smallest variance of the ergodic average obtained by the simulation. The intuition of Metropolis et al. (1953) thus was not only correct, but the convergence rate was even optimal, which is remarkable. While Hastings' achievement of generalising the Metropolis algorithm provided a theoretical justification, the result of Peskun (1973) even added that is was optimal to use the averages obtained via the algorithm. Nevertheless, in the years following the publications of Hastings (1970) and Peskun (1973), little attention was given to these results.

9.4 The Introduction of Simulated Annealing

While the paper of Hastings (1970) placed the original Metropolis algorithm on firmer theoretical grounds, the method was still far from being widely acknowledged. The first breakthrough came with the 1983 publication of Scott Kirkpatrick, C.D. Gelatt and M.P. Vecchi, named *Optimization by Simulated Annealing* (Kirkpatrick et al., 1983). In it,

[5] For the case that $\min\{(\pi_i q_{ij})/(\pi_j q_{ji}), (\pi_j q_{ji})/(\pi_i q_{ij})\}) = (\pi_j q_{ji})/(\pi_i q_{ij})$, derivations are analogue.

Kirkpatrick et al. (1983) considered combinatorial optimisation problems, in which deterministic problems were solved using the Metropolis algorithm. The algorithm suited the large combinatorial spaces extremely well. Instead of random number simulation, the primary goal of Kirkpatrick et al. (1983) was to find a global minimum value of a cost function inside of huge but discrete state spaces.

The context of Kirkpatrick's paper was how one should place electronic circuits like transistors onto computer silicon chips to optimize the signals and communication on the chip. While circuits near each other – that is, transistors on the same chip – have short signal ways, it is no easy task to position the transistors efficiently so that the total communication costs become minimal. Placing all transistors on one chip is not possible either. The problem can be formalised as a combinatorial one: Assuming that N circuits need to be positioned onto two separate silicon chips, using the statistical mechanics' terminology of Metropolis et al. (1953), a configuration ω can be modelled as an N-dimensional vector $\omega = (\omega_1, \omega_2, ..., \omega_N)$. For just two silicon chips, it suffices to use $\omega_i = \pm 1$ to indicate on which of the two chips the ith circuit is positioned. Kirkpatrick et al. (1983) reasoned

> "If we have connectivity information in a matrix whose elements a_{ij} are the number of signals passing between circuits i and j, and we indicate which chip circuit i is placed on by a two-valued variable $\mu_i = \pm 1$ then N_c, the number of signals that must cross a chip boundary is given by
>
> $$\sum_{i>j}(a_{ij}/4)(\mu_i - \mu_j)^2$$
>
> Calculating $\sum_i \mu_i$ gives the difference between the numbers of circuits on the two chips. Squaring this imbalance and introducing a coefficient, λ, to express the relative costs of imbalance and boundary crossings, we obtain an objective function, f for the partition problem:
>
> $$f = \sum_{i>j} \left(\lambda - \frac{a_{ij}}{2} \right) \mu_i \mu_j$$
>
> "
>
> Kirkpatrick et al. (1983, p. 674)

Kirkpatrick et al. (1983, p. 674) then noted, that the above objective function had the form of a Hamiltonian, or energy function which is studied in the theory of random magnets, where the common assumption is that the spins μ_i can only be oriented up or down. More precisely, they observed:

> "A typical optimization problem will contain many distinct, noninterchangeable elements, so a regular solution is unlikely. However, much research in condensed matter physics is directed at systems with quenched-in randomness, in which the atoms are not all alike. An important feature of such systems, termed "frustration," is that interactions favoring different and incompatible kinds of ordering may be simultaneously present (9). The magnetic alloys known as "spin glasses," which exhibit competition between ferromagnetic and antiferromagnetic spin ordering, are the best understood example of frustration (10). It is now believed that highly frustrated systems like spin glasses have many nearly degenerate random ground states

> rather than a single ground state with a high degree of symmetry. These systems stand in the same relation to conventional magnets as glasses do to crystals, hence the name. The physical properties of spin glasses at low temperatures provide a possible guide for understanding the possibilities of optimizing complex systems subject to conflicting (frustrating) constraints." Kirkpatrick et al. (1983, p. 673)

Furthermore, spin glasses resemble the Ising model with a modified energy function

$$E(\omega) = \sum_{i>j} (U - U_{ij})\omega_i\omega_j \tag{9.10}$$

and the resemblance to the objective function f for the circuits above is striking. In spin glass models, the quantities U_{ij} model ferromagnetic forces between neighbouring states or particles and these compete with repulsive anti-ferromagnetic forces modelled by U, therefore the phrasing *frustrated*, because both requirements cannot be satisfied simultaneously. Kirkpatrick et al. (1983) then came up with the idea, that because of this problem the original Metropolis algorithm as given in Algorithm 9 needed to be tuned carefully to identify low-temperature states in the state space due to the inherent repulsiveness of the influencing forces in the model. The trick was to slowly decrease the temperature T so that the system could converge to a low-energy state while retaining enough energy to move away from local minima into partitions of the state space where the global minimum exists. In particular, this procedure guaranteed that the simulation was not getting attracted too much by local minima. The temperature schedule, of course, needed to be tuned accordingly to allow such broader exploration in the beginning and after sufficient exploration, the decreasing temperature prevented larger moves more and more.[6] Kirkpatrick et al. (1983) summarised:

> "Using the cost function in place of the energy and defining configurations by a set of parameters $\{x_i\}$, it is straightforward with the Metropolis procedure to generate a population of configurations of a given optimization problem at some effective temperature. This temperature is simply a control parameter in the same units as the cost function. The simulated annealing process consists of first "melting" the system being optimized at a high effective temperature, then lowering the temperature by slow stages until the system "freezes" and no further changes occur. At each temperature, the simulation must proceed long enough for the system to reach a steady state. The sequence of temperatures and the number of rearrangements of the $\{x_i\}$ attempted to reach equilibrium at each temperature can be considered an annealing schedule.
>
> Annealing, as implemented by the Metropolis procedure, differs from iterative improvement in that the procedure need not get stuck since transitions out of a local optimum are always possible at nonzero temperature. A second and more important feature is that a sort of adaptive divide-and-conquer occurs. Gross features of the eventual state of the system appear at

[6] The technical details are omitted here due to space reasons. However, excellent theoretical introductions to simulated annealing from a statistical perspective (e.g. for finding the maxima or minima in posterior distributions) are given in Robert and Casella (2004). Good introductions to practical implementations of simulated annealing can be found in Robert and Casella (2010). Conceptually, all these implementations follow the early ideas of Kirkpatrick et al. (1983).

> higher temperatures; fine details develop at lower temperatures.” (Kirkpatrick et al., 1983, p. 672/673)

Kirkpatrick et al. (1983) applied their simulated annealing procedure to several diverse problems, showing promising results. Richey (2010) noted that it took some time until simulated annealing found widespread application. The hesitation ended with the development of a sound mathematical basis in the publications of Johnson et al. (1989, 1991), Laarhoven and Aarts (1987) and van Laarhoven (1988). Especially the work of Laarhoven introduced not only applications of simulated annealing, but also compared it to existing solutions and analysed its theoretical properties.

9.5 The Invention of the Gibbs-sampler

After the introduction of simulated annealing by Kirkpatrick et al. (1983), the next major step into the modern era of MCMC methods was achieved by the two brothers Donald and Stuart Geman. In 1984, just one year after simulated annealing was presented, they published their joint paper called *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images* (Geman and Geman, 1984) in which they applied a Metropolis algorithm to the problem of reconstructing a blurred image. Their paper, from today’s perspective, marks the birth hour of the general-purpose statistical algorithm which is now called *Gibbs sampling*.

Geman and Geman (1984) investigated the problem of a Bayesian analysis of images, which can be modelled as a grid of N pixels. Each pixel takes on colour values from a set $S = \{1, ..., M\}$, so that in terms of statistical mechanics an image can be represented by a configuration $\omega \in \Omega$, where $\omega = (\omega_1, ..., \omega_N)$ is defined by the allocation of values $s_i \in S$ to each ω_i for all i. This quickly leads to huge configuration spaces: An image with eight possible colours per pixel (that is, $S = \{1, ..., 8\}$) of 1000×1000 pixels leads to a configuration space with $|\Omega| = 8^{1000000}$ elements. The task Geman and Geman (1984) set themselves was then to reconstruct blurred images after transformations and noise have been applied to them. During this process, the original configuration ω is transformed into a blurred configuration $\tilde{\omega}$. After that, one wants to find out what configuration ω most probably may have been the original unblurred image. One easy example of blurring an image is given by adding additive noise in the form of $\tilde{\omega} = \omega + K$ with $K = (k_1, ..., k_N)$ and $k_i \sim \mathcal{N}(0, \sigma^2)$, for a fixed $\sigma^2 > 0$. After the raw blurring, the resulting values are rounded to values in S to get a valid image in the configuration space Ω. Geman and Geman (1984) then considered to Bayes’ theorem to reconstruct an original image ω out of a blurred one $\tilde{\omega}$:

$$p(\omega|\tilde{\omega}) = \frac{p(\tilde{\omega}|\omega)p(\omega)}{p(\tilde{\omega})} \tag{9.11}$$

Following Bayes’ theorem, Geman and Geman (1984) needed to derive which configuration ω has the highest probability of being the original image, given the blurred version. Due to the extremely large spaces in image analysis, this task was difficult. Geman and Geman (1984) therefore came up with the idea to adapt the original Metropolis algorithm to their needs. They first noted that the denominator in Equation (9.11) did not depend on ω and therefore was not necessary for the application of a Metropolis algorithm.[7] The likelihood $p(\tilde{\omega}|\omega)$ could be computed as

$$p(\tilde{\omega}|\omega) \propto \prod_{j=1}^{N} \exp\left(-\frac{k_j^2}{2\sigma^2}\right) = \prod_{j=1}^{N} \exp\left(-\frac{(\tilde{\omega}_j-\omega_j)^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}(\tilde{\omega}_j-\omega_j)^2\right) \quad (9.12)$$

because $\tilde{\omega}_j = \omega_j + k_j$ with $k_j \sim \mathcal{N}(0,\sigma^2)$ and therefore the likelihood simply models $k_j = \tilde{\omega}_j - \omega_j$. The real issue then was to select a suitable prior distribution $p(\omega)$, and this is where Geman and Geman (1984) saw a parallel between statistical mechanics and image reconstruction. They characterised an image by having patterns of some kind in it, because if there are no neighbored regions of similar pixel values, the image degenerates to white noise. This notion resembles the differences between an Ising lattice in balance or equilibrium and an Ising lattice which is unbalanced or not in equilibrium. Inspired by the analogy, Geman and Geman (1984) selected the Boltzmann probability in Equation (9.2) with the energy function of the Ising model E_I as a possible prior:

> "A Gibbs distribution relative to $\{S, \mathcal{G}\}$ is a probability measure π on Ω with the following representation:
>
> $$\pi(\omega) = \frac{1}{Z}e^{-U(\omega)/T}$$
>
> where Z and T are constants and U (...) the energy function ..."
> Geman and Geman (1984, p. 725)

Omitting the constant $\frac{1}{Z}$ and choosing the energy function $E_I(\omega) = -J\sum_{\langle i,j\rangle}\omega_i\omega_j - H\sum_{i=1}^{N}\omega_i$ of the Ising model[8], one obtains

$$p(\omega) \propto \exp(-E_I(\omega)/k \cdot T) \quad (9.13)$$

The final trick then was to set $k \cdot T < T_{crit}$ to obtain correlated pixel values, where T_{crit} is the critical threshold below which transitions occur in the Ising model. Following these ideas, the posterior in Equation (9.11) then could be written as

$$p(\omega|\tilde{\omega}) \propto \exp\left(-\frac{\sum_{j=1}^{N}(\tilde{\omega}_j-\omega_j)^2}{2\sigma^2}\right) \cdot \exp(-E_I(\omega)) = \exp\left[-\left(\frac{\sum_{j=1}^{N}(\tilde{\omega}_j-\omega_j)^2}{2\sigma^2} + E_I(\omega)\right)\right] \quad (9.14)$$

The analogy to the statistical mechanics domain becomes apparent when considering the exponent of the exponential function in the posterior probability above to be an

[7] As mentioned in Chapter 8, this presented a major advantage of the Metropolis algorithm which was first exploited by Geman and Geman (1984) systematically in their work.

[8] Here, $J > 0$ models the nearest-neighbour affinity, $H > 0$ describes the external field, and $\langle i,j\rangle$ indicates the set of nearest neighbours i and j which share at least a horizontal or vertical bond. For simplicity of calculations, no external field was assumed, so that $H = 0$ and $J = 1$. Then, the energy function reduces to $E_I(\omega) = \sum_{\langle i,j\rangle}\omega_i\omega_j$. The Ising model has long interested statistical physicists because of phase transition. A phase transition occurs when a quantity is affected by a substantial change as a parameter passes a critical value. The most familiar example of a phase transition is water as it freezes or boils and the critical value is zero or 100 degrees Celsius. For details on the Ising model see Robert and Casella (2010).

energy function, that is

$$E(\omega|\tilde{\omega}) := \frac{\sum_{j=1}^{N}(\tilde{\omega}_j - \omega_j)^2}{2\sigma^2} + E_I(\omega) = \underbrace{\frac{\sum_{j=1}^{N}(\tilde{\omega}_j - \omega_j)^2}{2\sigma^2}}_{=:(A)} + \underbrace{\sum_{\langle i,j\rangle} \omega_i\omega_j}_{=:(B)} \tag{9.15}$$

with the definition of the energy function $E_I(\omega)$ of the Ising model. Minimising the energy function $E_I(\omega)$ in Equation (9.15) then leads to the maximisation of the posterior probability in Equation (9.14) (Geman and Geman, 1984, p. 729). Here, the first component (A) can be interpreted as a penalty for too large differences between the blurred and original image, and the second component (B) is a term which gets small if neighboring pixels cluster in patterns of equal values (otherwise the products $\omega_i\omega_j$ become large). These two terms need to be balanced to get an optimal result. For example, an image ω can be quite different from $\tilde{\omega}$ so that term (A) is large, but neighbouring pixels can cluster strongly in it so that term (B) becomes small. A balanced solution yields a slightly different image with lots of clustered pixels, balancing terms (A) and (B). Geman and Geman (1984) then proceeded by introducing their modification of the Metropolis scheme, called *Gibbs sampling*. To solve the constrained minimisation problem, they introduced the nearest neighbors for a pixel ω_i and argued that in pictures in general always local patterns exist, so that ω_i is only influenced by its neighbors, and not by all pixels in total. Statistically speaking, conditional on all other pixels $\omega_j, j \neq i$, ω_i depends only on its direct neighbors. Therefore, the dependence on the blurred image $\tilde{\omega}$ as a whole is equal to dependence on the direct neighboring pixels, and Geman and Geman (1984) therefore reasoned that

$$p(\omega_i|\omega_1, ...\omega_{i-1}, \omega_{i+1}, ...\omega_N) = p(\omega_i|\omega_j \in \langle i,j\rangle) \propto \exp(-E_I(\omega_i|\omega_j \in \langle i,j\rangle)) \tag{9.16}$$

holds, where $\langle i,j\rangle$ denotes the direct neighbours of ω_j. Geman and Geman (1984) expressed this as follows:

> “Here, the computational problem is overcome by exploiting the pivotal observation that the posterior distribution is again Gibbsian with approximately the same neighborhood system as the original image, together with a sampling method which we call the Gibbs Sampler. Indeed, our principal theoretical contribution is a general, practical, and mathematically coherent approach for investigating MRF's by sampling (Theorem A), and by computing modes (Theorem B) and expectations (Theorem C).”
> Geman and Geman (1984, p. 722)

where MRF stands for Markov random field. Accordingly, the energy function for pixel ω_i then changes to

$$E_i(\omega_i|\omega_j \in \langle i,j\rangle) = \frac{(\tilde{\omega}_i - \omega_i)^2}{2\sigma^2} + \sum_{\langle i,j\rangle} \omega_i\omega_j \tag{9.17}$$

Here, the index i instead of I indicates that the energy function is applied only to the pixel ω_i, and therefore the sum over all pixels is omitted (compare Equation (9.15)). Geman and Geman (1984) then argued that it suffices to use a method which visits all of the pixels infinitely often (e.g. by iteration over all rows and columns), and at the pixel ω_i the posterior probability of $\omega_i = k$ is then given by

$$\begin{aligned} p(\omega_i = k|\tilde{\omega}) &\overset{\text{Equation (9.14)}}{\propto} \exp\left(-\frac{(\omega_i - \tilde{\omega}_i)^2}{2\sigma^2} - \sum_{\langle i,j\rangle} \omega_i \omega_j\right) \\ &\overset{\omega_i = k}{=} \exp\left(-\frac{(k - \tilde{\omega}_i)^2}{2\sigma^2} - \sum_{\langle i,j\rangle} k \cdot \omega_j\right) \end{aligned} \tag{9.18}$$

Easing the restriction of dependence on the whole blurred image in the posterior in Equation (9.14) to dependence only on the direct neighbours in Equation (9.16) in total leads to the posterior given in Equation (9.18), which gives the probability of the pixel ω_i having a specific colour value $k \in \{1, ..., M\}$. Repeating this procedure a large number of times for all pixels and colour values then gives a distribution of the posterior image. This is the essence of *Gibbs sampling*.

While Gibbs sampling was introduced in the context of image reconstruction via Bayesian analysis by Geman and Geman (1984), the whole procedure can be obtained as a special case of the Metropolis-Hastings algorithm as derived by Metropolis et al. (1953) and Hastings (1970). To see this, it suffices to set the acceptance probabilities $\alpha_{ij} = 1$ for all i, j in Equation (9.9) and let $q_{ij} := p(\omega_i = j|\tilde{\omega})$ as defined in Equation (9.18). Then, the proposal transitions are simply the posterior probabilities for the single pixels ω_i, and the acceptance probabilities are always equal to one. Detailed proofs of the convergence to the posterior $p(\omega|\tilde{\omega})$ can be found in Hammersley and Handscomb (1964), Gelfand and Smith (1990) and Casella and George (1992). Later, it was even shown that the Gibbs sampler could be interpreted as the composition of p Markovian Metropolis-Hastings kernels, for details see Robert and Casella (2004, p. 381).[9] Geman and Geman (1984) summarised their Gibbs sampling procedure as follows:

> "(...) our work is largely inspired by the methods of statistical physics for investigating the time-evolution and equilibrium behavior of large, lattice-based systems. There are, of course, many well-known and remarkable features of these massive, homogeneous physical systems. Among these is the evolution to minimal energy states, regardless of initial conditions. In our work posterior (Gibbs) distribution represents an imaginary physical system whose lowest energy states are exactly the MAP [Maximum A Posteriori] estimates of the original image given the degraded "data."
> All that is required is that the posterior distribution have a "reasonable" neighborhood structure as a MRF [Markov Random Field], for in that case the computational load can be accommodated by appropriate variants (such as the Gibbs Sampler) of relaxation algorithms for dynamical systems."
> (Geman and Geman, 1984, p. 734)

The name Gibbs sampling goes back to the original introduction of so-called *Gibbs distributions*, which are probability distributions whose conditional probabilities depend only on neighbourhood systems. The first mentioning of these distributions is found

[9] More specific: The Gibbs sampler is a composition of p Markovian Metropolis-Hastings kernels with acceptance probabilities uniformly equal to one, compare Theorem 8.11.

in Dobruschin (1968), who chose the name in honour of the physicist Josiah Willard Gibbs (1839-1903).

9.6 Markov-Chain-Monte-Carlo as a generalised statistical Simulation Technique

After the introduction of Gibbs sampling by Geman and Geman (1984), all the necessary theory was available to make use of it in Bayesian statistics. Hastings (1970) had already shown that the Metropolis algorithm was a powerful general-purpose tool for sampling, although few took notice of that fact. In particular, the intuitive appeal of the procedure (to numerically obtain previously untractable posterior distributions) was now also justified theoretically. Geman and Geman (1984) built upon the work of Kirkpatrick et al. (1983) and it became much clearer through their work that Gibbs sampling was just a special case of the concatenation of p independent Markovian Metropolis-Hastings kernels. This justified the procedure from another perspective. The technique was nevertheless first adopted widely by the statistical community after the publication of the paper *Sampling-Based Approaches to Calculating Marginal Densities* by Gelfand and Smith (1990). In their paper, Gelfand and Smith (1990) compared three different strategies of sampling, namely Gibbs sampling, the data augmentation algorithm by Tanner and Wong (1987) and importance sampling by Rubin (1987). The paper marked the first solely statistical interpretation of Gibbs sampling as a general-purpose tool for statistical inference. Previously to that paper, Besag (1986) presented his paper *On the Statistical Analysis of Dirty Picture*s at the meeting of the Royal Statistical Society in 1986. While similar to the work of Geman and Geman (1984), the paper was more a review than an introduction of a new method, and it discussed, in particular, the Gibbs sampler proposed by the Geman brothers two years earlier. In a comment to the paper, J. Haslett from Trinity College in Dublin emphasized regarding the question how to obtain a posterior distribution in Bayesian statistics

> "(...) that all such questions can be answered (in principle) by sufficient simulations (...) under Geman and Geman's method (...). It seems, therefore, that we are being offered no alternative, as statisticians, to the route of vast raw computing power being pioneered by the Gemans."
> Comment of J. Haslett in (Besag, 1986)

While Besag (1986) had some concerns regarding the necessary computing power for Gibbs sampling to work efficiently, the paper pointed out the advantages of the procedure quite clearly. The definitive breakthrough of Gibbs sampling came in the form of the two papers *Sampling-Based Approaches to Calculating Marginal Densities* (Gelfand and Smith, 1990) and *Illustration of Bayesian inference in normal data models using Gibbs sampling* (Gelfand et al., 1990), which showed the high efficiency of the Gibbs sampler in Bayesian hierarchical models. In the first paper, Gelfand and Smith (1990) set out the goal as follows:

> "In relation to a collection of random variables, $U_1, U_2, ..., U_k$, suppose that either (a) for $i = 1, ..., k$, the conditional distributions $U_i|U_j(j \neq i)$ are available, perhaps having for some i reduced forms $U_i|U_j(j \in S_i \subset \{1, ..., k\})$,

> or (b) the functional form of the joint density of $U_1, U_2, ..., U_k$ is known, perhaps modulo the normalizing constant, and at least one $U_i|U_j(j \neq i)$ is available, where *available* means that samples of U_i can be straightforwardly and efficiently generated, given specified values of the appropriate conditioning variables.
> The problem addressed in this article is the exploitation of the kind of structural information given by either (a) or (b), to obtain numerical estimates of nonanalytically available marginal densities of some or all of the U_i (when possible) simply by means of simulated samples from available conditional distributions, and without recourse to sophisticated numerical analytic methods. (...) All that the user requires is insight into the relevant conditional probability structure and technique for the efficient generation of appropriate random variates."
> Gelfand and Smith (1990, p. 398)

The novelty of this work was that the formulated goal was entirely uncoupled from the early beginnings of MCMC algorithms in statistical mechanics and circuit design. Instead, the paper was framed in a purely statistical context. Gelfand and Smith (1990) argued, that the full conditionals $f(x_i|x_1, ...x_{i-1}, x_{i+1}, ..., x_N)$ completely specify the joint distribution $f(x_1, x_2, ..., x_N)$ as these are essentially (large) neighbourhood systems as introduced by Geman and Geman (1984). Formally, this relationship was clarified completely first by the proof of the well known Hammersley-Clifford-Theorem (compare 8.12). The theorem was already proven in 1971 by John Hammersley and Peter Clifford in an unpublished paper (Hammersley and Clifford, 1971). Robert and Casella (2004, p. 377) noted, that Hammersley and Clifford did not publish it because they were unsatisfied with it and wanted to generalize it to hold also for densities which do not have strictly positive mass. Three years later, Moussouris (1974) gave a counterexample and showed that such a generalization is not possible. Only in 1990 then, Clifford (1990) published their results, long after others gave proofs. Peter Clifford wrote in a comment to Julian Besag who gave another proof of the Hammersley-Clifford theorem in 1974:

> "My final comment concerns the paper by Hammersley and myself. Whatever the historical reasons for not publishing in 1971 the paper has clearly been superseded by the work of others and notably by the excellent exposition we have heard today."
> Commentary of Peter Clifford in (Besag, 1974, p. 228)

Gelfand and Smith (1990) therefore only needed to use the necessary theory available and they argued that because of the Hammersley-Clifford-theorem

> "(...) the full conditional distributions alone, $[X|Y, Z]$, $[Y|Z, X]$, and $[Z|X, Y]$, uniquely determine the joint distribution (and hence the marginal distributions) in the situation under study. An algorithm for extracting the marginal distributions from these full conditional distributions was formally introduced by Geman and Geman (1984) and is known as the Gibbs sampler. An earlier article by Hastings (1970) developed essentially the same idea and suggested its potential for numerical problems arising in statistics."
> (Gelfand and Smith, 1990, p. 400)

where $[X|Y,Z]$ denotes the conditional distribution $X|Y,Z$. Gelfand and Smith (1990) then stated the Gibbs algorithm in its modern form as given in Algorithm 7:

> "Given an arbitrary starting set of values $U_1^{(0)}, U_2^{(0)}, ..., U_k^{(0)}$, we draw $U_1^{(1)} \sim [U_1|U_2^{(0)}, ..., U_k^{(0)}], U_2^{(1)} \sim [U_2|U_1^{(1)}, U_3^{(0)}, ..., U_k^{(0)}], U_3^{(1)} \sim [U_3|U_1^{(1)}, U_2^{(1)}, U_4^{(0)}, ..., U_k^{(0)}]$, and so on, up to $U_k^{(1)} \sim [U_k|U_1^{(1)}, ..., U_{k-1}^{(1)}]$. Thus each variable is visited in the natural order and a cycle in this scheme required K random variate generations. After i such iterations we would arrive at $(U_1^{(i)}, ..., U_k^{(i)})$."
> Gelfand and Smith (1990, p. 400)

Via the use of this Gibbs sample, inference for the parameters is easily obtained: Choosing for example the third component, the Gibbs sample $x_3^{(1)}, x_3^{(2)}, ..., x_3^{(i)}$ can be used to approximate the marginal probability distribution of x_3, that is a random sample from

$$f(x_3) = \int_{x_1} \int_{x_2} \int_{x_4} ... \int_{x_N} f(x_1, x_2, ..., x_N) dx_N ... dx_4 dx_2 dx_1 \tag{9.19}$$

can be simulated by taking the distribution of the subsample $x_3^{(1)}, x_3^{(2)}, ..., x_3^{(i)}$ of the whole Gibbs sample. In the same way, the expectation $\mathbb{E}[X_3]$ of the third component X_3,

$$\mathbb{E}[X_3] = \int x_3 f(x_3) dx_3 \tag{9.20}$$

can be approximated by the mean of the Gibbs subsample, that is, $\frac{1}{i}\sum_{j=1}^{i} x_3^{(j)}$. Other quantities, like credible intervals, can be obtained equivalently. While the paper did not formally introduce a new algorithm, it leveraged Gibbs sampling into the statistical community by achieving the same what Hastings in 1970 had achieved regarding the original Metropolis algorithm: Generalising a procedure developed in a specific context to a purely statistical routine. Next to this, they provided good examples. In section 3.2 of their paper, Gelfand and Smith (1990) considered hierarchical models, which were mostly untractable from a Bayesian viewpoint at that time. They introduced the famous nuclear-pump model of Gaver and O'Muircheartaigh (1987), which since then has become a benchmark model for performance analysis of MCMC algorithms. Details and the Gibbs sampler for this model can be found in the paper of Gaver or in Robert and Casella (2004, Example 10.17). After showing the power of Gibbs sampling with multiple examples in their paper, they provided even more illustrations from a statistical perspective in a second paper which was published soon after the first (Gelfand et al., 1990). In it, they advertised the Gibbs sampler as a general-purpose tool for statistical inference. By doing so, Gelfand and Smith (1990) managed to bring MCMC algorithms finally into the realms of statistical science. In their first paper, Gelfand and Smith (1990) also gave a hint about what would become one of the major research areas for the next decade concerning MCMC methods:

> "There are important practical problems in tuning monitoring and stopping-rules procedures for iterative sampling in large-scale complex problems"
> (Gelfand and Smith, 1990, p. 407)

9.7 Hamiltonian Monte Carlo and further Developments

After the paper of Gelfand and Smith (1990), MCMC methods finally received deserved attention as a flexible general-purpose tool in statistical research and practice. The developments after Gelfand's publication can be structured into a few distinct branches of research.

9.7.1 Convergence assessment

Already in 1970, Hastings (1970) noted, that

> "...even the simplest of numerical methods may yield spurious results if insufficient care is taken in their use, and how difficult it often is to assess the magnitude of the errors. The discussion above indicates that the situation is certainly no better for the Markov chain methods and that they should be used with appropriate caution."
> (Hastings, 1970, p. 105)

While Gelfand and Smith (1990) just advised readers to use a trace plot of the parameters $\theta^{(t)}$ of the Markov Chain against the time t, no formal convergence assessment was introduced. Convergence assessment, as detailed in Robert and Casella (2004) is concerned with (i) monitoring the convergence to the stationary distribution, (ii) monitoring the convergence of averages and (iii) monitoring to i.i.d. sampling. Robert and Casella (2004) noted, that

> "Historically, there was a flurry of papers at the end of the 90s concerned with the development of convergence diagnoses. This flurry has now quieted down, the main reason being that no criterion is absolutely foolproof (...)."
> Robert and Casella (2004, p. 461)

MCMC convergence diagnosis was approached by those papers mainly by a few main strategies. One of the earliest of those strategies was to simply use multiple chains and the between-chain and within-chain variance to indicate convergence to the stationary distribution. If all of the chains converge to the same distribution, which means that the chains are mixing well, this indicates that the stationary distribution has been reached. One problem with this approach is that the slowest chain determines the speed of convergence when following such a criterion and more severely, this method suffers from the "you've only seen where you've been"-defect. If the chains are not dispersed well, that is, they start in similar regions – in the worst case, all near a local mode – the criterion is not helpful at all. The most popular statistic resulting from these papers is the *Gelman-Brooks-Rubin* $\hat{R}^2$ (Gelman and Rubin, 1992). Raftery and Lewis (1992) proposed another technique called *binary control* as a method for assessing convergence. Other approaches can be found in Tierney (1994). Ritter and Tanner (1992), Brooks et al. (1997) and Brooks and Roberts (1998) developed so called *distance evaluations* for convergence assessment, see also Robert and Casella (2004, Chapter 12). Philippe and Robert (2001) proposed an approach called *missing mass*, which is useful but not useable in higher dimensions. For other approaches based on *renewal theory* and methods based on normality tests like the Kolmogorov-Smirnov test, see Geyer (1992). Cowles and Carlin (1996) wrote a comparative review of the available methods, and another later review can be found in Mengersen et al. (1998). Both reviews conclude that no penultimate method can be chosen to satisfy all needs in all circumstances so that diagnosis stays a problematic task.

Even after more than three decades, convergence assessment of MCMC algorithms remains a challenging task, as profound theoretical knowledge is required to judge the simulation results and understand the different criteria for assessing convergence. However, Craiu and Rosenthal (2014) reviewed the current state of affairs concerning MCMC convergence assessment and concluded that while there is still no foolproof solution to MCMC convergence diagnosis, combining multiple of the existing criteria offers reasonable security in judging the simulation outputs. Also, Vats and Knudson (2018) recently proposed a refinement of the popular Gelman-Rubin-Brooks $\hat{R}^2$ statistic, which remains (maybe because of its simplicity) the gold standard of MCMC convergence assessment until today.

9.7.2 Adaptive MCMC

Another branch of research on MCMC methods emerged out of the problem that most MCMC algorithms need a proper scaling of parameters to work efficiently. For example, the random-walk Metropolis-Hastings algorithm as given in Algorithm 3 can be used with a normal proposal $\mathcal{N}(\mu, \sigma^2)$, where often $\mu = 0$ is a common choice, and σ^2 needs to be set to a specific value. Choosing a value of σ^2 too small will lead to a high acceptance rate, but simultaneously to slow exploration of the posterior distribution. However, choosing a value of σ^2 too large will also yield a high rate of rejections in the Metropolis scheme and result in an inefficient exploration of the posterior. As the tuning parameter σ^2 needs to be chosen *just right*, this problem inherent to most MCMC algorithms (e.g. Algorithm 1, Algorithm 2 and Algorithm 3) has been entitled *Goldilocks principle* in the statistics community, see also Craiu and Rosenthal (2014) and Rosenthal (2014, Section 6.5).[10] Out of this problem, in the 2000s statisticians came up with the idea to automatically adapt the tuning parameters in the MCMC algorithms at runtime. For example, one could try to decrease σ^2 when the last 50 iterations yielded a rejection rate which was above a specific threshold, say $\delta \in [0, 1]$. Similarly, one could increase it when the last 50 iterations yielded a rejection rate which was below another threshold $\gamma \in [0, 1]$, thus balancing the rejection rate into an optimal region. One problem when following this idea was that when modifying an MCMC algorithm, it often loses its property of convergence to the correct stationary distribution. To be more specific, the resulting process is not Markovian anymore, and asymptotic ergodicity needs to be verified for every single modified algorithm, see also Craiu and Rosenthal (2014, Section 5.2). The strategy to change the tuning parameter (e.g. σ^2) 'on the fly' when running the algorithm, by using the information of previous iterations, has been utilised by Haario et al. (2001), Andrieu et al. (2005) and Roberts and Rosenthal (2007). Andrieu and Thoms (2008) and Roberts and Rosenthal (2009) gave examples and technical introductions to adaptive MCMC, promoting the approach. Bai et al. (2011), Rosenthal (2014), as well as Yang and Rosenthal (2017) and Yang et al. (2019), showed how adaptive MCMC algorithms could be improved even further by including regional information for the adaptation scheme and by using multilevel-algorithms which include two separate phases for first tuning the parameters and second running the algorithm of in-

[10] The name Goldilocks principle is taken from the tale *Goldilock and the Three Bears*, where a little girl called Goldilock tastes bowls of porridge, which should be neither too hot, nor too cold, but *just right*.

terest. The most important theoretical advances have been made by Roberts and Rosenthal (2007), who showed that in general adaptive MCMC algorithms need to fulfil two conditions to preserve ergodicity. By denoting the transition kernel from X_n to X_{n+1} as $\mathbb{P}_{\Gamma_n}$ with each fixed kernel $\mathbb{P}_\gamma$ having the stationary distribution $\pi(\cdot)$ and Γ_n being random indices chosen based on the past algorithm steps from an index set $\mathcal{Y}$, Roberts and Rosenthal (2009) defined $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : ||\mathbb{P}_\gamma^n(x,\cdot) - \pi(\cdot)||_{\text{TV}} \leq \varepsilon\}$ for the time the kernel $\mathbb{P}_\gamma$ needs to converge to the stationary $\pi(\cdot)$ with a precision of $\varepsilon > 0$ when starting in $x \in \mathcal{X}$, where $\mathcal{X}$ is the state space and $|| \cdot ||_{\text{TV}}$ the total variation distance. They showed that under the two conditions of *diminishing adaption*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} ||\mathbb{P}_{\Gamma_{n+1}}(x,\cdot) - \mathbb{P}_{\Gamma_n}(x,\cdot)|| = 0 \text{ in probability} \tag{9.21}$$

and *bounded convergence*, that is, the set

$$\{M_\varepsilon(X_n,\Gamma_n)\}_{n=0}^\infty \tag{9.22}$$

needs to be bounded in probability for all $\varepsilon > 0$, the resulting adaptive algorithm is ergodic, see Theorem 1 in Roberts and Rosenthal (2007). Roberts and Rosenthal (2009, Section I) give an accessible introduction. While these conditions have simplified verifying the ergodicity of a given adaptive MCMC algorithm, it remains a challenge and necessity to verify the ergodicity on a case-by-case basis for each adaptive MCMC algorithm. Therefore, adaptive MCMC methods have not found widespread use until today, although they are very versatile when proven to retain their convergence properties inherited from non-adaptive MCMC algorithms.

9.7.3 Hamiltonian Monte Carlo

Another branch of research emerged after the spread of MCMC methods due to the paper of Gelfand and Smith (1990) in 1990. This branch indeed goes back to another earlier paper by Duane et al. (1987), which united approaches that emerged out of the area of molecular dynamics with the existing MCMC theory. Duane et al. (1987) called their method – like the title of their paper – *Hybrid Monte Carlo*. The original application of their method was concerned with lattice field theory simulations of quantum chromodynamics. These simulations made use of the Hamiltonian dynamics of particle physics to exploit the geometry of the posterior distribution of interest when exploring it via an MCMC algorithm. Since then, this approach has become known as *Hamiltonian Monte Carlo* (HMC).[11] The first statistical applications of HMC started with improvements of traditional MCMC methods via Hamiltonian dynamics in (Neal, 1993), and the work on neural network models by Neal (1996). Early work on HMC include applications in generalised linear models (Ishwaran, 1999), as well as the papers of Liu (2004) and Schmidt (2009).

HMC became popular among statisticians after Radford Neal (who also introduced the slice sampler as given in Algorithm 5, see (Neal, 2003)) published a chapter called *MCMC Using Hamiltonian Dynamics* (Neal, 2011) in Steve Brooks' *Handbook of Markov-Chain-Monte Carlo* (Brooks, 2011). In it, he explained the ideas underlying HMC and

[11] The idea of Hamiltonian Monte Carlo is similar to the intuition behind adaptive MCMC. Both methods aim for a dynamic exploration of the posterior distribution instead of a static exploration as achieved by Metropolis-Hastings algorithms. While Hamiltonian Monte Carlo exploits the geometry of the posterior distribution to sample from the posterior more efficiently, adaptive MCMC automatically scales the proposal distribution to yield a similar behaviour.

introduced simple examples, connecting the terminology of particle physics with that of MCMC from a Bayesian perspective. The advantage of HMC is summarized by him as follows:

> "The first step is to define a Hamiltonian function in terms of the probability distribution we wish to sample from. In addition to the variables we are interested in (the "position" variables), we must introduce auxiliary "momentum" variables, which typically have independent Gaussian distributions. The HMC method alternates simple updates for these momentum variables with Metropolis updates in which a new state is proposed by computing a trajectory according to Hamiltonian dynamics, implemented with the leapfrog method. A state proposed in this way can be distant from the current state but have a high probability of acceptance. This fact bypasses the slow exploration of the state space that occurs when Metropolis updates are done using a simple random-walk proposal distribution."
> Neal (2011, p. 113-114)

The so-called *leapfrog integrator* often used in HMC algorithms results from the need to discretize the *Hamiltonian differential equations* involved in the approach. Therefore, although the method seems to evade the inherent problems of tuning the proposal parameters in common MCMC algorithms, it has itself to be tuned properly. The *stepsize* and *number of leapfrog steps* are two parameters which determine the acceptance rate of HMC. Therefore, while improving the exploration, raw HMC does not solve the problem of tuning MCMC algorithms. Neal (2011) and Betancourt (2017) give excellent introductions to the theory behind HMC. In what follows, the most substantial foundations of HMC are outlined.

Augmentation of the parameter space

The general idea behind HMC can be summarized as follows: Hamiltonian dynamics are observed on a $2d$-dimensional space, which is an augmentation of the original parameter space. If a statistical model $\mathcal{M}$ (e.g. a posterior distribution of interest) involves d parameters $\theta_1, ..., \theta_d$ of interest, then every $\theta_i, i \in \{1, ..., d\}$ is augmented via the introduction of a corresponding variable $\gamma_i, i \in \{1, ..., d\}$. In physical interpretations, the original variables of interest θ_i are the *position* variables, each of which is associated with a corresponding *momentum* variable γ_i, augmenting the d-dimensional parameter space into a $2d$-dimensional space. Denoting $p := (\theta_1, ..., \theta_d)$ and $q := (\gamma_1, ..., \gamma_d)$, a *Hamiltonian* function $H(p, q)$ is introduced which describes the Hamiltonian dynamics of the whole system.

Describing Motion with Hamilton equations

To describe the change over time of the system, one uses partial derivatives of the position and momentum variables for $i = 1, ..., d$. Therefore, the following *Hamilton equations* are assumed:

$$\frac{\partial q_i}{dt} = \frac{\partial H}{\partial p_i} \tag{9.23}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{9.24}$$

To describe the change over time of the whole system via these Hamilton equations, one needs to specify the Hamiltonian function H in the above, which usually is written as

$$H(q,p) = U(q) + K(p) \tag{9.25}$$

Here, $K(p)$ is called the kinetic energy, and is usually defined as

$$K(p) := p^T M^{-1} \frac{p}{2} \tag{9.26}$$

with a symmetric, positive-definite matrix M which describes the mass (typically diagonal). The potential energy $U(q)$ above is usually defined as the negative of the log probability density of the distribution for the parameters $q := (\theta_1, ..., \theta_d)$ which one is interested in sampling from (Neal, 2011, Section 5.2). Choosing the Hamiltonian this way, one immediately obtains the updated Hamilton equations

$$\frac{dq_i}{dt} = [M^{-1}p]_i \tag{9.27}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i} \tag{9.28}$$

for $i = 1, ..., d$. To implement the Hamiltonian dynamics in a simulation, one needs to approximate these differential equations by discretizing time with a stepsize $\varepsilon > 0$. The discretization then can be used to calculate the state of the Hamiltonian system at different times, e.g. $\varepsilon, 2\varepsilon, 10\varepsilon$ and so on. While the choice of the Hamiltonian function as given in Equation (9.25) can be a different one, for simplicity of computation often the above choice is used and also the assumption of M being diagonal is added (Neal, 2011, Section 5.2.3). Denoting the diagonal elements $m_1, ...m_d$, one obtains[12]

$$K(p) = \sum_{i=1}^{d} \frac{p_i^2}{2m_i} \tag{9.29}$$

Approximating the Hamilton equations via Euler's Method or the Leapfrog Integrator

In principle, there are multiple methods to approximate the Hamilton differential equations, the most popular ones being Euler's Method and the Leapfrog Method. Euler's method iteratively performs the following two steps:

$$p_i(t+\varepsilon) = p_i(t) + \varepsilon \cdot \frac{dp_i(t)}{dt} \overset{(1)}{=} p_i(t) - \varepsilon \cdot \frac{\partial U}{\partial q_i}(q(t)) \tag{9.30}$$

$$q_i(t+\varepsilon) = q_t(t) + \varepsilon \cdot \frac{dq_i}{dt}(t) \overset{(2)}{=} q_i(t) + \varepsilon \cdot \frac{p_i(t)}{m_i} \tag{9.31}$$

where in (1) Equation (9.28) and Equation (9.25) was used, and in (2) Equation (9.27) and Equation (9.25) was used as well as Equation (9.29). Via use of Equation (9.30) and Equation (9.31) one can then proceed by starting at $t = 0$ with fixed values $p_i(0), q_i(0)$ and simulate for a given step size $\varepsilon > 0$ the trajectory of position and momentum values for example at the points $\varepsilon, 2\varepsilon, 3\varepsilon, ...,$ and so on. This procedure is called Euler's

[12] For a simple derivation seeNeal (2011, p. 119).

method and has the disadvantage that it diverges in some cases – see (Neal, 2011) or (Betancourt, 2017) – and via a modification of Euler's method or by using the leapfrog integrator better results can be obtained. The modification of Euler's method simply uses the new value for the momentum variables p_i, when the new value for the position variable q_i is computed. The modification of Euler's method, therefore, results in the following two steps:

$$p_i(t+\varepsilon) = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t)) \tag{9.32}$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon)}{m_i} \tag{9.33}$$

The leapfrog integrator is an even more sophisticated modification of the original Euler method and uses the following iterative update scheme (compare Neal (2011, Section 5.2.3.3)):

$$p_i(t+\varepsilon/2) = p_i(t) - (\varepsilon/2) \frac{\partial U}{\partial q_i}(q(t)) \tag{9.34}$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon/2)}{m_i} \tag{9.35}$$

$$p_i(t+\varepsilon) = p_i(t+\varepsilon/2) - (\varepsilon/2) \frac{\partial U}{\partial q_i}(q(t+\varepsilon)) \tag{9.36}$$

The scheme shows the origin of the name leapfrog integrator. The algorithm starts with half a step (or jump) for the momentum variables p_i, adding a full step (or jump) for the position variables q_i and another half step (or jump) for the momentum variables p_i resembling the movement of a jumping frog.

Theoretical considerations of HMC

As Hamiltonian dynamics expand the parameter space artificially, the geometric properties of the new space need to fulfil some conditions. There are four important properties of Hamiltonian dynamics which need to be fulfilled. First, there needs to be a one-to-one mapping from a starting point and the endpoint of a trajectory obtained via Euler's method or the leapfrog integrator. This requirement is called *reversibility*. Second, *invariance of the Hamiltonian* is required, which results in the invariance of acceptance probabilities when using Metropolis updates in HMC. Third, *volume preservation* is necessary, because otherwise, one would have to adjust Metropolis acceptance probabilities due to the change in volume in the augmented parameter space. The fourth and last requirement is *symplecticness*, which itself causes the Hamiltonian dynamics to preserve volume.[13]. It is possible to show that when approximating the Hamiltonian equations for example by using the leapfrog integrator, reversibility, preservation of volume as well as symplecticness are maintained, so HMC does not cause theoretical problems which could invalidate the procedure in practice. This is important and guarantees that HMC still converges to the stationary distribution and maintains ergodicity.[14] While the discretization of the Hamilton equations via the leapfrog integrator or Euler's method inevitably leads to the introduction of an error term, this error term

[13] For technical details about these properties, see Neal (2011) and Betancourt (2017)

[14] An accessible proof of both properties can be found in Neal (2011, Section 5.3).

behaves benignly in the way that when the stepsize ε goes to zero, the error term does, too. For a detailed account, see Leimkuhler and Reich (2004).

Applying Hamiltonian dynamics to sample from a posterior

To sample more efficiently from a posterior distribution, one needs to define the posterior in terms of a Hamiltonian function as given in Equation (9.25). The density is translated into the potential energy, and the artificial momentum variables are introduced. Therefore, one uses an energy function $E(x)$ for a state x of a physical system (compare Section 9.2 and Section 9.4), for which the *canonical distribution* over states has the probability density

$$\mathbb{P}(x) = \frac{1}{Z}\exp\left(\frac{-E(x)}{T}\right) \tag{9.37}$$

where T is the temperature of the system as already introduced in Section 9.4 and Z is the normalising constant (Neal, 2011). If the goal is to sample a probability density $f(x)$, one can rewrite it as a canonical distribution of a physical system by setting $T = 1$ and $E(x) = -\log(f(x)) - \log(Z)$ with $Z \in \mathbb{R}_+$.[15] Expanding these ideas to the augmented parameter space, the Hamiltonian as given in Equation (9.25) can be interpreted as an energy function for the augmented space, defining a joint distribution for both position and momentum via

$$\mathbb{P}(p,q) = \frac{1}{Z}\exp\left(\frac{-H(p,q)}{T}\right) \overset{(1)}{=} \frac{1}{Z}\exp\left(-\frac{U(q)}{T}\right)\exp\left(-\frac{K(p)}{T}\right) \tag{9.38}$$

where (1) follows from Equation (9.25). To sample from a posterior $P(q|x) \propto L(q|x)p(q)$ with prior $p(q)$ and likelihood $L(q|x)$ for q given the data x, it suffices to use the assumption $T = 1$ and write the posterior as a canonical distribution with a potential energy

$$U(q) := -\log\left(L(q|x)p(q)\right) \tag{9.39}$$

Hamiltonian Monte Carlo operates then by sampling from the joint canonical distribution for p and q given in Equation (9.38) after specifying the kinetic energy $K(p)$ as given in Equation (9.26). This results in p having a $\mathcal{N}(0, M)$ distribution because $\exp\left(-\frac{K(p)}{T}\right)$ in Equation (9.38) then becomes $\exp\left(-\frac{p^T M^{-1}\frac{p}{2}}{1}\right) = \exp\left(-\frac{p^2}{2M}\right)$. Thus, incorporation of the posterior as a joint canonical distribution yields a product of the log density of the posterior (omitting proportionality constants) and a multivariate normal. The simplest HMC algorithm then proceeds in two steps:

1. New values for the momentum variables p_i are drawn from their Gaussian distributions $\mathcal{N}(0, M)$, independently from the position variables q_i.
2. A Metropolis acceptance step is performed, where Hamiltonian dynamics are used to propose a new state. Therefore, the leapfrog integrator starts at (q, p) and computes a trajectory of L steps with stepsize $\varepsilon > 0$. The variables $(\tilde{q}, \tilde{p})$ at the end of the trajectory are then taken as proposal value and accepted with probability

$$\min\{1, \exp\left(-H(\tilde{q}, \tilde{p}) + H(q,p)\right)\} = \min\{(1, \exp\left(-U(\tilde{q}) + U(q) - K(\tilde{p}) + K(p)\right)\}$$

[15] Then, $\mathbb{P}(x) = \frac{1}{Z} \cdot \exp(\frac{-(-\log(f(x)) - \log(Z))}{1}) = \frac{1}{Z}\exp(\log(f(x)) \cdot \exp(\log(Z)) = f(x)$.

This approach is similar to Algorithm 3. However, here the proposal from an arbitrary proposal density is replaced with the result of the leapfrog integrator, which approximates the solutions to the Hamiltonian differential equations. In total, this leads to the endpoints $\tilde{q}, \tilde{p}$.While it may not seem obvious from this perspective, the ultimate quantity of interest, the original model parameters $q := (\theta_1, ..., \theta_d)$ are then readily available by simply repeating this process a large number of times T like in normal MCMC and using the simulated values of e.g. $\tilde{q}_1^{(i)}$, that is $\tilde{q}_1^{(1)}, \tilde{q}_1^{(2)}, ..., \tilde{q}_1^{(T)}$. Of course, $\theta_1, ..., \theta_d$ could each be multidimensional depending on the statistical model $\mathcal{M}$ for which posterior inference is required. Posterior point and interval estimates like the mode and credible intervals are then easy to obtain. While the posterior is rewritten as a canonical distribution in Equation (9.39), the joint distribution of (p, q), that is $\mathbb{P}(p, q) = \frac{1}{Z}\exp\left(\frac{-H(p,q)}{T}\right)$ as given in Equation (9.38) is used to obtain posterior estimates of both the parameters q and p of the posterior. This is necessary, as the momentum variables p are the main reason why HMC improves traditional MCMC methods. While the posterior estimates of p are available too, these are not used for inference in the model $\mathcal{M}$. Extracting the posterior chain $\tilde{q}_i^{(1)}, \tilde{q}_i^{(2)}, ..., \tilde{q}_i^{(T)}$ for parameter θ_i for a large T suffices to draw inference about the parameters of interest $q := (\theta_1, ..., \theta_d)$. Illustrations of HMC can be found in (Betancourt, 2017) and (Neal, 2011) as well as multiple others like (Hoffman and Gelman, 2014) and (Gelman et al., 2015).

The key property why HMC has performed so well in practice is that the invariance of H when using Hamiltonian dynamics implies that a trajectory computed by the leapfrog or any other integrator will (if simulated with a sufficiently small error) lie in a hyperplane of *constant* probability density value. Therefore, moves to points $(\tilde{q}, \tilde{p})$, which are far away from the original destination (q, p) still have a fairly good probability of being accepted (if the approximation error is not too large so that the trajectory has approximately constant density). This situation is in sharp contrast to normal MCMC methods like random-walk Metropolis-Hastings, where jumps far away from the current state are often rejected due to the difference in probability density, slowing down the parameter space exploration.

Introduction of the No-U-Turn-Sampler

After the initial introductions of HMC into the statistical community, interest in this method grew and one of the most substantial advances was the recent introduction of the No-U-Turn-Sampler (NUTS) of Hoffman and Gelman (2014), which automatically sets the path lengths ε and leapfrog steps L in the HMC algorithm. This property removed the challenge of manual calibration and tuning of MCMC from researchers. The idea of the No-U-Turn-Sampler can be summarised as aborting the leapfrog length whenever the curvature of the planned trajectory indicates that the next step will move back to its original position similar like a boomerang. This condition guarantees that the parameter space is explored efficiently without the sampler returning to its latest position in each successive step.

9.7.4 Probabilistic Programming Languages

Next to work on adaptive MCMC methods, a new branch of research emerged in the form of *probabilistic programming languages* to facilitate the use and spread of MCMC methods in science and practical applications. In general, any programming language can be used to implement MCMC algorithms. Due to the heavy computational load required by even low-dimensional statistical models, programming libraries which offer efficient reference implementations in fast programming languages like C or C++ were developed in the 2000s. The general idea behind probabilistic programming languages consists in defining a statistical model in the probabilistic programming language, that is, the model code. The software package then parses this model code into executable – that is, compiled – code (e.g. C or C++ code), which then is used to generate a sample from the posterior distribution of interest. Probabilistic programming languages, especially software packages to simplify the use of MCMC methods have become increasingly popular in the last years, see for example Lunn et al. (2009).

BUGS and WinBUGS

The first probabilistic programming language was BUGS. BUGS is an acronym for ***B**ayesian Inference **U**sing **G**ibbs **S**ampling* and the BUGS project started in 1989 after the huge success of the Gibbs sampler as a solution to previously untractable hierarchical models in statistics. BUGS itself was developed by researchers of the Imperial College School of Medicine in London and the MRC Biostatistics Unit at Cambridge University. To simplify the use of BUGS, the software package WinBUGS was developed for Microsoft Windows and made it possible to use the BUGS language with a graphical user interface. With BUGS, one could specify a statistical model which was then compiled into executable code. This code was run to obtain a sample from the posterior distribution of interest. While WinBUGS was known to statisticians involved in the research of MCMC methods, it did not become prevalent in other scientific domains due to the need of manual programming each model as well as the computational limitations of that time (Lunn et al., 2009). In 2007, the last version 1.4.3 of WinBUGS was released, and the development team switched to OpenBUGS, which was already started in 2005 and is an open-source version of the original BUGS software (Lunn et al., 2009).

OpenBUGS

OpenBUGS had some important advantages over WinBUGS: First, it was open source, making use and adaptation of the source code to one own's needs much easier. Second, it was platform-independent, which allowed users of the commonly used operating systems to use the program. Third, it was accessible from the statistical programming language *R*, which by then was already widespread among mathematicians, physicists, statisticians and data scientists. Another difference to WinBUGS lied in its simplicity of use: The package automatically chose the updating algorithm for the class of conditional distributions of each stochastic node in the statistical model. This property yielded higher flexibility of the algorithms OpenBUGS used for obtaining the posterior distribution (Lunn et al., 2009). As in WinBUGS, the model code was specified via the probabilistic programming language BUGS, and then parsed and compiled into executable code by OpenBUGS to obtain a sample from the posterior distribution of interest. From the beginning, OpenBUGS was specifically developed to run with the programming language S-Plus, the predecessor of the statistical programming language R. Therefore, OpenBUGS works seamlessly with S-Plus and R, and there are R packages like *R2OpenBUGS*[16] and *BRugs*[17] which catalyse the use of OpenBUGS and BUGS in R.

[16] See `https://cran.r-project.org/web/packages/R2OpenBUGS/index.html` for details

[17] See `https://cran.r-project.org/web/packages/BRugs/index.html` for details

This constituted a strong advantage of OpenBUGS.[18]

JAGS

Next to OpenBUGS, in 2007 the software package JAGS was officially released. JAGS is an acronym for ***J**ust **A**nother **G**ibbs **S**ampler* and was developed by Martyn Plummer already since 2003 (Plummer, 2003). Since its official release in 2007, JAGS has been used in a broad spectrum of disciplines, for example in biology (Semmens et al., 2009), medicine (McKeigue et al., 2010), management (Johnson and Kuhn, 2013) and psychology and the cognitive sciences (Kruschke, 2015).

One of the main reasons for the widespread use of JAGS is that it was included in a lot of Linux distributions and written in C++ (while OpenBUGS and WinBUGS were written in Component Pascal, a less widely popular programming language) (Lunn et al., 2009). Also, there are convenient packages for R like `rjags`[19] as well as command-line support for scripts and 64-bit support for modern and more capable processors.[20] JAGS uses the hierarchical BUGS models to sample from the posterior of interest. After handing a BUGS model to JAGS, in which the relationships between the variables of the statistical model are specified, JAGS identifies the likelihoods as functions defining a variable for which observations are available (Plummer and Northcott, 2017). Subsequently, the distributions in the model are analysed, and before sampling, JAGS automatically chooses an appropriate MCMC algorithm. In general, this will be a Gibbs sampler as given in Algorithm 7. If the full conditionals are not available, JAGS resorts to Metropolis-Hastings-algorithms like Algorithm 1, Algorithm 2, or Algorithm 3 or even slice sampling as introduced by Neal (2003) and given in Algorithm 5. Plummer and Northcott (2017) and (Coro, 2017) provide an overview about the details of the parsing process of the BUGS model into stochastic nodes which are subsequently transformed into a sample from the corresponding posterior. The automaticity of selecting an appropriate MCMC algorithm for sampling made the use of JAGS easier than previous programming packages for MCMC methods, and this can be seen as the main reason why JAGS has been used successfully in so many scientific branches since its publication.

STAN

In 2012, the probabilistic programming language STAN was released by researchers of the Columbia University (Hoffman et al., 2012; Stan Development Team, 2018). The software name was chosen in honour of Ulam Stan, who participated in the invention of the original Monte Carlo method, as detailed in Section 9.2. STAN is a probabilistic programming language similar to BUGS, which allows users to specify a statistical model. In a second step, multiple MCMC algorithms are available to sample from the posterior. STAN also includes the No-U-Turn-Sampler of Hoffman and Gelman (2014). Thereby, STAN makes it possible to adaptively set the path lengths in Hamiltonian Monte Carlo via the No-U-Turn-Sampler and removed the burden of manual MCMC tuning from users. Since its introduction, it has been maintained and extended

[18] A comprehensive overview about OpenBUGS written by the author of the *R2OpenBUGS* R package Neal Thomas can be found at `http://www.openbugs.net/w/Overview`.

[19] See `https://cran.r-project.org/web/packages/rjags/index.html` for details

[20] The code repository for JAGS can be found at `https://sourceforge.net/p/mcmc-jags/code-0/ci/default/tree/` and the official website is located at `http://mcmc-jags.sourceforge.net`.

(Gelman et al., 2015). Stan has caused a significant breakthrough in terms of the application of Bayesian statistics in more applied areas because the No-U-Turn-Sampler provided a nearly automatic way to obtain a posterior distribution without the need of detailed theoretical or programming knowledge (Kruschke and Liddell, 2018b). Also, STAN can be accessed from a variety of programming languages, including R, Python, Matlab, Julia, Stata or the shell, making it accessible to a wide audience. In contrast to WinBUGS, OpenBUGS or JAGS, STAN implements gradient-based MCMC algorithms, in particular HMC, which exploit the geometry of the posterior distribution of interest. This property accelerates the speed of simulations and has allowed for successful use even when the computational capacities are only moderate. Also, it supports multiple other algorithms for variational Bayesian inference as well as gradient-based optimisation (Gelman et al., 2015). The "workhorse" of STAN for Bayesian inference, however, is the *No-U-Turn sampler* as introduced by Hoffman and Gelman (2014) and outlined above[21]. The main advantage of this algorithm is given by the fact that the sampler automatically chooses the steplength L and stepsize $\varepsilon > 0$ in the leapfrog integrator to obtain optimal performance when sampling. This feature speeds up the simulations and results in better exploration of the parameter space compared to traditional MCMC methods. The No-U-Turn-Sampler can be interpreted as an algorithmic solution to the challenge of *Goldilocks principle* as detailed in Section 9.7.2.

Due to the implementation of state-of-the-art MCMC algorithms and broad support of different programming languages, STAN has already been used in a variety of fields like social sciences (Goodrich et al., 2012), medical imaging (Gordon et al., 2018) and pharmaceutical statistics (Natanegara et al., 2014) since its introduction.

9.8 The Impact of Markov-Chain-Monte-Carlo on Bayesian Hypothesis Testing

In Chapter 6, the basics of Bayesian hypothesis testing were introduced. It was shown that a key problem with Bayesian inference in practice is presented by obtaining the posterior distribution, which quickly becomes untractable analytically in complex statistical models. This chapter detailed the evolution of Markov-Chain-Monte-Carlo algorithms from the early beginnings in Los Alamos until the modern Hamiltonian Monte Carlo methods, which are now implemented in freely available software. Together, these algorithms solve the difficult task of obtaining a posterior distribution even in complex and possibly high-dimensional models. Thus, no analytical calculations are required anymore when MCMC algorithms are used for inference.

As shown in Chapter 6, the use of Bayes factors for Bayesian hypothesis testing is the most popular approach and has the longest history (Jeffreys, 1961). A substantial advantage of the Bayes factor was given by the fact that it could be calculated analytically for some standard models. Still, for increasingly complex statistical models, calculation of the Bayes factor analytically in a closed-form expression becomes challenging quickly. The advent of MCMC methods as detailed in this chapter thus had three substantial effects on Bayesian hypothesis testing.

First, the availability of MCMC algorithms led to the development of multiple new indices of significance and effect size next to the Bayes factor (Makowski et al., 2019b).

[21] See also the Stan reference manual (Stan Development Team, 2018).

The uniting approach of these indices is to use some combination of the prior distribution, likelihood and posterior distribution to quantify the evidence regarding a null hypothesis H_0 or alternative hypothesis H_1. Most often, MCMC algorithms obtain the posterior distribution numerically in practice, and thus provide a previously not existing freedom in using these different ingredients for quantifying the statistical evidence about a hypothesis. Examples include the e-value (Pereira and Stern, 1999; Pereira et al., 2008; Pereira and Stern, 2020) and the *Full Bayesian Significance Test* (*FBST*), the MAP-based p-value (Mills, 2018), the probability of direction (PD) (Makowski et al., 2019b), the region of practical equivalence (ROPE) (Kruschke, 2013; Kruschke and Liddell, 2018b; Kruschke, 2018) and the support interval (Wagenmakers et al., 2020). These indices are discussed in detail in Part IV in Chapter 14.

Second, MCMC methods made it possible to compute Bayes factors numerically after the posterior distribution was simulated via an MCMC sample. An example is given by the *Savage-Dickey density ratio method* (Dickey and Lientz, 1970; Verdinelli and Wasserman, 1995; Wagenmakers et al., 2010), which provides a closed-form expression for the Bayes factor and only requires the prior and posterior density for the calculation. Thus, even when no closed-form derivations of the Bayes factor are available, these methods now offer a numerical solution when at least the posterior distribution of the parameter of interest can be obtained. Employing probabilistic programming languages like JAGS or STAN, this latter requirement is nearly always fulfilled. Therefore, even for complex models, MCMC methods have opened the door to an algorithmic way to compute Bayes factors.

Third, as a consequence of the second point, MCMC algorithms made it possible for the first time to derive Bayesian versions of a variety of frequentist hypothesis tests. For example, Wetzels et al. (2009) used WinBUGS to derive a Bayesian equivalent of the frequentist two-sample Student's t-test. Other examples include Kruschke (2013), who presented another Bayesian alternative to the one- and two-sample Student's and Welch's t-test by using STAN and JAGS, or van Doorn et al. (2020), who derived a Bayesian version of the Wilcoxon-rank-sum test via MCMC methods. While some of these tests like the one presented by van Doorn et al. (2020) use customised MCMC techniques like data augmentation (Tanner and Wong, 1987) or Gibbs sampling after deriving the full conditional distributions, others rely only on standard MCMC samplers like STAN or JAGS (Kruschke, 2013). Another example is given in Chapter 15, where a new Bayesian solution to the Behrens-Fisher-problem based on the Hodges-Lehmann-paradigm is presented, which provides a new Bayesian two-sample t-test.

Summing up, the advent of modern MCMC algorithms has removed the necessity of analytic derivations for each statistical model from Bayesian hypothesis testing and replaced it with a simulation-based approach. In this approach, a variety of alternatives to the Bayes factor are available, and posterior distributions are routinely obtained numerically via Markov-Chain-Monte-Carlo algorithms.

INTERMEDIATE CONSIDERATIONS

The first two parts have shown that statistical hypothesis testing has been evolved out of two main schools of statistical thought: On one side, there is the frequentist school, founded by Fisher, Neyman and Pearson, which is based on a deductive argument of rejecting hypotheses to draw inferences. In the frequentist school, hypothesis testing is performed by employing the distribution of the data and treating the parameter as an unknown, fixed quantity. In contrast, there is the Bayesian school, which goes back to Laplace, Jeffreys and others and proceeds by accumulating evidence in light of the observed data. In the Bayesian school, hypothesis testing is primarily performed based on the posterior distribution, for example in form of the Bayes factor. This part showed that the problems of conducting Bayesian hypothesis tests from a computational point of view have been solved largely by the introduction of modern MCMC algorithms, because these allow to obtain a posterior distribution numerically in the majority of cases. However, both statistical philosophies have their benefits and drawbacks: The frequentist philosophy is based on falsification which from Neyman's and Pearson's perspective is aimed at long-term performance and type I error control, while from Fisher's perspective an individual case-based situation should be considered. Chapter 5 has shown that eventually, the Neyman-Pearson framework succeeded, while the methodology that is actually used today often resembles a hybrid of Fisherian significance testing and the original Neyman-Pearson theory. The long-term false-positive control inherited from the original Neyman-Pearson theory can be seen as a benefit which yields control over the type I errors in the long run (if the necessary assumptions are met), but at the same time it is guaranteed that false positive results will occur with a fixed percentage, even for huge sample sizes and otherwise perfect experimental design. While the reconstruction in Chapter 4 showed that the Neyman-Pearson theory was never intended as a theory for hypothesis testing in scientific contexts, Chapter 5 showed that it has become the standard for hypothesis testing in the biomedical, social and cognitive sciences.

The Bayesian philosophy does not incorporate explicit error control, but states relative degrees of belief in a hypothesis via the posterior distribution of parameters about which the hypothesis makes a statement. The goal can therefore be rejection, but also confirmation of a hypothesis, and it is possible to judge evidence for the study or experiment conducted, instead of relying on long-term performance guarantees. While error rates cannot be predetermined, in practice they can be estimated via Monte Carlo studies as will be demonstrated in Chapter 14.

However, by now it has not been clarified which scientific theories underpin each of the two statistical philosophies. Therefore, the following part is structured into two chapters: Chapter 10 analyses the current situation from a philosophical perspective. In particular, the recent philosophical interpretation of the replication crisis of Mayo (2018) is analysed, which leads to the traditional problems of induction. It is shown that the often stated criticism to induction and thereby Bayesian inference are not tenable, and that induction (and thereby Bayesian statistics) cannot be rejected when implemented as a version of *probabilistic affirming the consequent*. In fact, it will be argued that Bayes' Theorem itself is a direct implementation of probabilistic affirming the consequent which itself is a weaker form of enumerative induction, and that Bayes' Theorem therefore presents an appealing scientific theory for judging the statistical evidence about research hypotheses in a statistical model.

The second Chapter 11 is split into two parts and the first part builds upon Chapter 10 by analysing the replication crisis from an axiomatic perspective. It is shown that based on the arguments of Birnbaum (1962) and Berger and Wolpert (1988) only very elementary principles can be assumed – similar to Kolmogorov's axioms in probability theory – which lead to the likelihood principle, already mentioned in Part I. It is discussed why based on the likelihood principle both the classic Fisherian significance tests and Neyman-Pearson tests are not tenable in scientific contexts, and why the likelihood principle itself can hardly be rejected based on a purely axiomatic point of view. Also, various of the currently experienced problems in the scientific replication crisis are explained on the basis of these violations of axiomatic principles of statistical inference. The second part of Chapter 11 then discusses whether robust Bayesian inference is a possible replacement of null hypothesis significance tests in scientific contexts. It is discussed whether robust Bayesian inference is suitable to mitigate the problems experienced in the scientific replication crisis.

Part IV

On the Axiomatic Foundations of Statistical Inference

Chapter 10

Philosophical Considerations on Bayesian Statistical Inference

> The probability of a statement (...) simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests.
>
> Karl Popper
> The Logic Of Scientific Discovery

10.1 The traditional Problem of Induction

From a of philosophy of science perspective, both frequentism and Bayesianism can be interpreted as realisations of scientific theories, see Mayo (2018). The most widely adopted perspective is that frequentism can be seen as a statistical implementation of Karl Popper's falsificationism, which itself is rooted deeply in deductive reasoning (Popper, 1959). Fisher's significance testing strongly influenced Popper's falsificationism, as Mayo (2018) notes: "Early on, Popper (1959) bases his statistical falsifying rules on Fisher, though citations are rare." (Mayo, 2018, p. 83). The name falsificationism goes back to Lakatos and Musgrave (1970), who dubbed Popper's philosophy 'methodological falsificationism' (Lakatos and Musgrave, 1970, p. 109). Popper wrote his highly influential monograph *'The Logic of Scientific Discovery'* in 1959, and at that time frequentist statistics in the form of Fisher's significance testing or the Neyman-Pearson theory were already standard in university classes. The major achievement of Popper (1959) can be seen in framing the statistical approach of frequentism under his newly developed theory of science, that is, under falsificationism. On the other hand, Bayesian statistics with Bayes' theorem at its core can be interpreted as an implementation of inductive reasoning (Mayo, 2018). The reason is that Bayes' theorem can be interpreted as an implementation of probabilistically affirming the consequent, as will be discussed in this chapter. The debate between frequentism and Bayesianism, therefore, can be reallocated to the decision between induction and deduction. In the discussion about the appropriateness of induction or deduction for judging the relevance of new scientific findings the more relevant question, however, is whether induction is a suitable method. While deduction builds on axioms and the conclusions follow from these, induction is a method of reasoning where the premises are interpreted as supplying some amount of evidence, but not full assurance of the truth of the conclusion of interest. Thus, from a statistical point of view, induction is of immediate interest because in scientific contexts the validity (or probability) of a research hypotheses needs to be judged in light of the observed data, which can only provide some evidence but not full assurance of the hypothesis unless the whole population is sampled.

In this chapter, the focus is on the recent work of Mayo (2018) because her account is also discussing these philosophical aspects in light of the replication crisis, and her perspective is rooted in the critical rationalism of Popper (1959), which builds on the earlier logical empirism of Carnap (1950). Popper's critical rationalism thus provides the central scientific theory – falsificationism – which is implemented in the frequentist null hypothesis significance tests of Fisher and Neyman and Pearson. A widely debated argument is that probability is no good measure of corroboration for a scientific theory or hypothesis, and the original argument goes back to Popper (1959). This provides a challenge for inductive approaches, including the Bayesian approach. In this chapter, the traditional problem of induction is discussed and several arguments against enumerative induction and its probabilistic implementation in the form of Bayes theorem are analyzed. It is shown that these arguments are based on a variety of misconceptions and thus do not show that enumerative induction and Bayesian inference cannot be used as a scientific theory.

The traditional problem of induction can be described as seeking to justify a specific type of argument, which takes the form of enumerative induction (EI) (Mayo, 2018). EI tries to infer:

Enumerative induction
Premise: All observed $A_1, A_2, ..., A_n$ have been B's.
Conclusion: Therefore H : all A's are B's.

The problem thus can be translated in plain words as inferring a general rule H in the sense of Wrinch and Jeffreys (1923b) when only a subset of the whole set of all existing elements has been observed. Clearly, the argument is deductively invalid because the premise EI can be true, although the conclusion H is false. The traditional problem of induction, therefore, reduces to justifying the method of enumerative induction itself. One needs to seek an argument for the conclusion that EI is rationally justified and a reliable rule. Using any kind of inductive argument itself to justify EI is impossible because inductive enumeration itself is not justified so it cannot be used in the proof. This situation inevitably pushes the enquiry into a spiral of circular reasoning. One would be using the method one is trying to justify, to justify the very same method. As this logical fallacy does not allow to proceed this way, different options to justify EI need to be considered. A different option to justify EI would be to use a deductively valid argument. Mayo (2018, p. 62) argued that one possible premise may be

Premise 1: EI has been reliable in a set of observed cases.

The premise can be true, but clearly, EI need not be reliable in general. Mayo (2018, p. 62) thus proposed to add a second premise:

Premise 2: Methods that have worked in past cases will work in future cases.

This premise is exactly the statement of EI, so it lands the enquiry again in a circle. Mayo (2018) (and others) called this the logical problem of induction, because all attempts to justify EI collapse into assuming EI in the first place. Therefore, from the 1930s to 1960s, philosophers of science were looking for logics which represented plausible inductive reasoning. The general approach at that time – also called *logical positivism* – was to build logics which embodied EI. Mayo (2018, p. 63) called these attempts *evidential-relation* (*ER*) *logics*, and gave the following example:

Evidential-relation logic
Premise 1: If H: all A's are B's, then all observed A's $(A_1, A_2, ..., A_n)$ are B's.
*Premise 2:*All observed A's $(A_1, A_2, ..., A_n)$ are B's.
Conclusion: Therefore, H: all A's are B's.

The first added premise, of course, is true. However, the second premise can be true, but the conclusion still is false, because one cannot infer that all A's are B's when only all observed A's $(A_1, ..., A_n)$ are B's. Therefore, the above logic is not a deductively valid argument. Still, logics or arguments like this are often called *affirming the consequent* (Mayo, 2018). While any logic which is affirming the consequent is not deductively valid, analytical philosophers tried to solve the problem by weakening the logic into a probabilistic version of it. This probabilistic version was called *probabilistic affirming the consequent*. The idea behind logics which are probabilistic affirming the consequent is that the conclusion does not follow, but only becomes more probable or '*gets a boost in confirmation of probability*' (Mayo, 2018, p. 63), often called a *B-Boost*.

What is gained by lowering the strength of the argument of enumerative induction into a probabilistic version of it? The original problem was that EI could not justify EI, so there is no inductive argument available. Also, EI itself is no valid deductive argument. In total, EI can neither be justified via an inductive nor via a deductive argument. Interestingly, when switching to the probabilistic version of EI this problem disappears. The reason is that now a deductively valid argument can be found, which is exactly *Bayes' theorem*. Bayes' theorem indeed can be formalised as a valid deductive argument, when an underlying formal system of probability – that is, a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ - is assumed:

Bayes Theorem
Premise: $\mathbb{P}(H_1), ..., \mathbb{P}(H_n)$ are the prior probabilities of an exhaustive set of hypotheses on the parameter space Θ; data $\boldsymbol{x}$ are given and the likelihoods $\mathbb{P}(x|H_i)$ are defined for each $i = 1, ..., n$.
Conclusion: It follows that

$$\mathbb{P}(H_i|x) = \frac{\mathbb{P}(x|H_i)\mathbb{P}(H_i)}{\mathbb{P}(x|H_1)\mathbb{P}(H_1) + ... + \mathbb{P}(x|H_n)\mathbb{P}(H_n)}$$

The above argument shows that Bayes' Theorem is a valid deductive argument for probabilistic affirming the consequent. Therefore, enumerative induction can be justified by changing EI into its probabilistic version EI_p:

Probabilistic enumerative induction
Premise: All observed $A_1, A_2, ..., A_n$ have been B's.
Conclusion: Therefore the probability of H : all A's are B's is increased (by employing Bayes' theorem).

Note that employing Bayes' theorem is the reason for this probabilistic version to remain a deductively valid argument, which was the goal. As a consequence, probabilistic enumerative induction can be justified as a valid deductive argument by involving Bayes theorem. It should be stressed that this idea is often used to interpret Bayes' theorem as a plausible confirmation theory, because it probabilistically justifies EI by embodying probabilistically affirming the consequent.

Historically, the central question when proceeding this way was how to obtain the probabilities defined by $\mathbb{P}$ and, in particular, the probability measure $\mathbb{P}$ itself. The question, therefore, was how to define the measure $\mathbb{P}$ on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ used in Bayes' theorem. Only then, probabilistic affirming the consequent is achieved in a meaningful way.[1] One of the most ambitious programs to resolve this issue was the one of Carnap (1962), who tried to assign probabilities to hypotheses by deducing these from the "logical structure of a particular (first order) language" (Mayo, 2018, p. 63). In short, history showed that this attempt was not successful at all. Also, it was vehemently opposed especially by Salmon (1966, 1988). Therefore, it remained unclear (at least for philosophers of science) how to define the measure $\mathbb{P}$ in Bayes' theorem to provide meaningful inference.

This confusion can be attributed primarily to two aspects:

1. The first point is the lack of mathematical background of both Popper and Mayo. For example, Mayo (2018, p. 86) noted from personal correspondence: "When Popper wrote me "I regret not studying statistics", my thought was "not as much as I do".". When Bayes' theorem is used from a modern statistical perspective, the foundations of measure theory precisely define which measure $\mathbb{P}$ is used on the parameter space Θ. If, for example $\Theta := \mathbb{R}^d$ for $d \geq 1$, then the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ with the standard Lebesgue measure λ^d defines the probability space $(\Omega, \mathcal{A}, \mathbb{P}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$. Thus, in almost all realistic situations, the answer to the question of how to 'define the probabilities' assigned to sets by the measure $\mathbb{P}$ follows from modern measure-theory. In the case of the most common setting of continuous parameter spaces, only a single solution presents itself by the foundations of measure theory. In these cases, including the example above, the measure used most often is the Lebesgue measure λ, see also Rüschendorf (2014) and Bauer (2001). In discrete settings, the counting measure presents the standard solution. Situations in which the measure $\mathbb{P}$ can be defined in multiple ways do thus not exist for practical applications. The question of how to select the measure $\mathbb{P}$ is, therefore, less a relevant question, as already recognised by Wrinch and Jeffreys (1921).

2. While the first point makes the choice of measure from a mathematical point of view less debatable, the more important question for practice is how to select prior probabilities for each hypothesis, of which there may be an infinite number to be considered. Wrinch and Jeffreys (1921) answered this question from a theoretical perspective so that the selection of prior probabilities is handled both for the finite and infinite case. Their solution consisted of assigning the parameter a mixture prior distribution as detailed in Chapter 6. This made it possible to test a hypoth-

[1] Note the similarity to Wrinch and Jeffreys (1921). For them, the selection of the measure $\mathbb{P}$ was less a problem, and the more important question was how to assign an infinite number of hypotheses a prior probability which is distinguishable from zero. Their approach used trans-finite series to assign the hypotheses prior probabilities, but the resonance was only moderate.

esis even when the parameter space was continuous and a uniform distribution could not be assigned to parameter spaces like $\mathbb{R}^d$ for $d \in \mathbb{N}$.

The second point in the confusion about how to select a prior probability measure clearly is more relevant for practice. However, even when the number of hypotheses to be considered through the statistical model is infinite, the finite measurement precision in any experiment or study implies that a finite number (which still may be huge) of hypotheses suffices to be considered in practice. The problem that no uniform distribution may exist on a continuous noncompact parameter space is therefore softened, as the measurement process always is finite and therefore only discretisations of continuous parameter spaces are measured.

Nevertheless, for Popper (1959), it was a serious problem that one could use different measures $\mathbb{P}$, which may be attributed to the fact that he was not aware of the necessary measure-theoretic concepts. Therefore, he declared a lack of a clear justification of the probabilistic version of enumerative induction. Influenced by Popper's writings, Hacking (1980) postulated in 1980 that "there is no such thing as a logic of statistical inference" (Hacking, 1980, p. 145). What is more, he also added that the attempts of probabilistic affirming the consequent were "founded on a false analogy with deductive logic" (Hacking, 1980, p. 145). His reasons for the statement were that probability is not a good measure for confirmation based on an argument which in turn was based on an example Popper (1959) gave years earlier to discredit probabilistic affirming the consequent.

10.2 Popper's Criticism to Inductive Reasoning

So, the choice of the measure $\mathbb{P}$ in Bayes' Theorem is clear from a modern measure-theoretic perspective. However, Popper (1959) argued that even when putting this problem aside for a moment, there remain problems. According to him, it remains unclear how to update or boost the probability of a hypothesis when using probabilistic affirming the consequent in form of Bayes' theorem. This caused Popper (1959) to state that probability is no good measure of confirmation.[2] He reasoned:

> "By 'the problem of degree of corroboration' I mean the problem (i) of showing that there exists a measure (to be called degree of corroboration) of the severity of tests to which a theory has been subjected, and of the manner in which it has passed these tests, or failed them; and (ii) of showing that this measure cannot be a probability, or more precisely, that it does not satisfy the formal laws of the probability calculus."
> Popper (2005, p. 402)

[2] For space reasons the famous Popper-Miller argument against probability is not discussed at length in this chapter, but the interested reader is referred to the excellent monograph of Sprenger and Hartmann (2019) for more details, in particular Variation 9 in Sprenger and Hartmann (2019). Interestingly, Sprenger and Hartmann (2019) in their explication of a probabilistic measure of corroboration arrive at the Kemeny-Oppenheim measure (Kemeny and Oppenheim, 1952) which is ordinally equivalent to the weight of evidence, which is the log-Bayes factor, compare the early results of Good (1960, 1968) and Good (1985). Good (1985) credits Turing (1942) for the original explication of the weight of evidence as the appropriate measure for corroboration. Thus, the analysis of Sprenger and Hartmann (2019) essentially arrives at the same conclusion that this chapter points at: That the Bayes factor is a well-justified probabilistic measure of corroboration. For a more detailed discussion of induction from a philosophy of science perspective the reader is also referred to Sprenger (2016).

The second problem thus presented an unsurmountable obstacle for Popper to accept enumerative induction in the form of Bayes theorem. Popper thus advocated a

> "mathematical refutation of all those theories of induction which identify the degree to which a statement is supported or confirmed or corroborated by empirical tests with its degree of probability in the sense of the calculus of probability. The refutation consists in showing that if we identify degree of corroboration or confirmation with probability, we should be forced to adopt a number of highly paradoxical views"
> Popper (2005, p. 405)

Thus, his goal was to present an example which shows that the degree of confirmation of a scientific theory is not well described via probability:

> "It is often assumed that the degree of confirmation of x by y must be the same as the (relative) probability of x given y, i.e., that $Co(x,y) = P(x,y)$. My first task is to show the inadequacy of this view."
> Popper (1959, p. 396)

In the above $Co(x,y)$ denotes the confirmation of x by y and $P(x,y) = p(x\&y)/p(y)$ the definition of conditional probability. Popper's P is equal to the measure $\mathbb{P}$ in Bayes' Theorem. He presented the following example:

> Consider the (...) throw with a homogeneous die. Let x be the statement 'six will turn up'; let y be its negation, that is to say, let $y = \overline{x}$; and let z be the information 'an even number will turn up'.
> (Popper, 2005, p. 406)

Popper argued that these events have the probabilities

$$p(x) = \frac{1}{6} \quad p(y) = \frac{5}{6} \quad p(z) = \frac{1}{2} \tag{10.1}$$

where the use of p implies that the same probability measure $\mathbb{P}$ is used for these counting densities. Thus, all of these probability statements are based on the same probability mass function p with respect to the counting measure. However, Popper did not define how the data are modelled: For example, if interest lies in the number of dice rolls which turn out as a six, a binomial model would be appropriate. If interest lies in each of the dice faces, a multinomial model would be required. Popper first calculated the conditional probabilities

$$p(x|z) = \frac{1}{3} \quad p(y|z) = \frac{2}{3} \tag{10.2}$$

which are obtained as

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} = \frac{1 \cdot \frac{1}{6}}{\frac{1}{2}} = \frac{1}{3} \tag{10.3}$$

$$p(y|z) = \frac{p(z|y)p(y)}{p(z)} = \frac{\frac{2}{5} \cdot \frac{5}{6}}{\frac{1}{2}} = \frac{2}{3} \tag{10.4}$$

and thus Popper argued that "x is supported by the information z, for z raises the probability of x from $1/6$ to $2/6 = 1/3$" (Popper, 2005, p. 406). Likewise, the probability

$p(y) = \frac{5}{6}$ decreases to $p(y|z) = \frac{2}{3}$. Therefore, based on the conditional probabilities based on z the probability for x increases and the probability for y decreases. However, Popper then inspected the change in probabilities before conditioning on z and after conditioning on z, which is given as

$$\frac{p(x|z)}{p(x)} = \frac{\frac{1}{3}}{\frac{1}{6}} = 2 > 1 \tag{10.5}$$

$$\frac{p(y|z)}{p(y)} = \frac{\frac{2}{3}}{\frac{5}{6}} = \frac{4}{5} < 1 \tag{10.6}$$

Now, Popper denoted $Co(x|z)$ as the event that $p(x|z) > p(x)$ and then argued based on the above:

> “There exists statements x, y, and z which satisfy the formula
>
> $$Co(x|z) \& \sim Co(y|z) \& p(x|z) < p(y|z) \tag{10.7}$$
>
> ... we have established by our example: that x may be supported by z, and y undermined by z, and that nevertheless x, given z, may be less probable than y, given z.”
> Popper (2005, p. 406-407)

In the above, $\sim Co(y|z)$ denotes that $p(y|z) \leq p(y)$. Clearly, based on the probabilities Popper assumed for x, y and z, $Co(x|z)$ and $\sim Co(y|z)$ holds due to Equation (10.5) and Equation (10.6), and $p(x|z) < p(y|z)$ follows from Equation (10.3) and Equation (10.4). Popper thus concluded:

> “Thus we have proved that the identification of degree of corroboration or confirmation with probability (and even with likelihood) is absurd on both formal and intuitive grounds: it leads to self-contradiction.”
> Popper (2005, p. 407)

Summing up, the above reasoning led Popper (1959) to reject any kind of probabilistic affirming the consequent because it also was based on probability to quantify the uncertainty about a hypothesis, and thus he proceeded with deductive reasoning, following the paths which pioneers like Fisher or Neyman and Pearson had already paved intuitively. He also claimed that

> “the probability of a statement ... simply does not express an appraisal of the severity of the tests a theory has passed, or of the manner in which it has passed these tests.“
> Popper (1959, p. 394-395)

10.2.1 Mayo’s Interpretation of Popper’s Example

Mayo (2018) followed Popper’s approach closely in constructing her error statistical account. Therefore, she introduced the terms absolute and relative B-Boost:

Definition 10.1 (Incremental / relative B-Boost)**.** The hypothesis H is confirmed by data x if and only if $\mathbb{P}(H|x) > \mathbb{P}(H)$, H is disconfirmed by x if and only if $\mathbb{P}(H|x) < \mathbb{P}(H)$, where $\mathbb{P}(H|x) + \mathbb{P}(\sim H|x) = 1$ and $\sim H$ is the set complement of H, that is, $\sim H := \Theta \setminus H$ where Θ is the parameter space.

Definition 10.2 (Absolute B-Boost)**.** The hypothesis H is confirmed by data x if and only if $\mathbb{P}(H|x)$ is high, at least $\mathbb{P}(H|x) > \mathbb{P}(\sim H|x)$, where $\mathbb{P}(H|x) + \mathbb{P}(\sim H|x) = 1$ and $\sim H$ is the set complement of H, that is, $\sim H := \Theta \setminus H$ where Θ is the parameter space.

The absolute B-Boost thus corresponds to the situation when $\mathbb{P}(H|x) > \frac{1}{2}$. Then, she used his example as follows:

> "His example consists of a homogeneous die: The data x: an even number occurs; the hypothesis H: a 6 will occur. It's given that $P(H) = 1/6$, $Pr(x) = 1/2$. The probability of H is increased by data x, while $\sim H$ is undermined by x (its probability goes from 5/6 to 4/6). If we identify probability with degree of confirmation, x confirms H and disconfirms $\sim H$. However, $Pr(H|x) < Pr(\sim H|x)$. So H is less well confirmed given x than is $\sim H$, in the sense of (2)."
> Mayo (2018, p. 67)

In the above, (2) refers to the absolute B-Boost. Also, Mayo gave two additional arguments for not using probabilistic affirming the consequent: (1) The paradox of irrelevant conjunctions and (2) the inability of B-Boosts to update the evidence in case of 100% reliable sources.

However, Mayo's treatment conflates the probability of a hypothesis and the probability of observing data. A hypothesis makes a statement about the parameter $\theta \in \Theta$ both in the frequentist and Bayesian approach, compare Appendix C. Her hypothesis H: a 6 will occur thus is no valid hypothesis. It is the value of the probability mass function p that Popper used. Popper's original writings show that he did not even formulate a specific hypothesis, and he only intended to obtain paradoxical probability statements with regard to the sets x, y and z.

Mayo's argument then is the same as Popper: Given that $\mathbb{P}(H) = 1/6$ and $\mathbb{P}(x) = 1/2$, the probability of H is increased by x, because after observing an even number, the probability of $\mathbb{P}(H|x)$ is obtained via the definition of conditional probability as

$$\mathbb{P}(H|x) = \frac{\mathbb{P}(x|H)\mathbb{P}(H)}{\mathbb{P}(x)} = \frac{1 \cdot 1/6}{1/2} = \frac{1}{3}$$

Therefore, H is confirmed in the sense of her incremental B-Boost, see Definition 10.1. The probability of the negation of H, $\mathbb{P}(\sim H|x)$ is obtained in the same way and decreases from $5/6$ to $2/3 = 4/6$. However, if the absolute B-Boost is now taken as the degree of confirmation, x disconfirms H and confirms $\sim H$. Mayo concludes this by noting that

$$1/3 = 2/6 = \mathbb{P}(H|x) < \mathbb{P}(\sim H|x) = 4/6$$

Therefore, the hypothesis H is still less confirmed by x than its complement $\sim H$ in the sense of the absolute B-Boost. In total this leads to a situation in which the two interpretations of probabilistic affirming the consequent – the incremental and absolute B-Boost lead to contradictory inferences about the hypothesis H. However, Mayo's interpretation of Popper is clearly inadequate because she misinterprets Popper's original statements which are statements about observing data in the sample space as statements about parameters in the parameter space and then (incorrectly) assigns an explicit Bayesian prior probability to a set which is contained in the sample space.

10.2.2 Argument (1) – The paradox of irrelevant conjunctions

Next to reciting Popper's counterexample, Mayo's first argument is based on the fact that it is possible to attach irrelevant hypotheses J to a given hypothesis H, and if x confirms H then it also confirms $H\&J$. Note that

$$\frac{\mathbb{P}(H|x)}{\mathbb{P}(H)}=\frac{\mathbb{P}(x|H)\mathbb{P}(H)}{\mathbb{P}(H)\mathbb{P}(x)}=\frac{\mathbb{P}(x|H)}{\mathbb{P}(x)}$$

so an incremental B-Boost is equivalent to $\mathbb{P}(x|H) > \mathbb{P}(x)$. Then, two assumptions are made by Mayo:

1. $\mathbb{P}(x|H)/\mathbb{P}(x) > 1$, that is, x causes an incremental B-Boost to H
2. $\mathbb{P}(x|H\&J) = \mathbb{P}(x|H)$, which is interpreted as J being an irrelevant conjunction

Substituting 2. into 1. yields $\mathbb{P}(x|H\&J) > \mathbb{P}(x)$. Mayo (2018) gives the example of H:*'the general theory of relativity deflection of light effect is 1.75'* and J:*'the radioactivity of the Fukushima water being dumped in the pacific ocean is within acceptable levels'*, which shows the absurdity of the paradox: Whenever the data x cause an incremental B-Boost to H, they also cause an incremental B-Boost to the irrelevant conjunction J which is entirely unrelated to H.

10.2.3 Argument (2) – The inability of B-Boosts to update the evidence in case of reliable sources

The second argument of Mayo (2018) follows Achinstein (2001, 2010), who criticised that a B-Boost is a problematic concept because a certain event cannot be "b-boosted" anymore. His example consists of the data x :'the newspaper says Harry won' (some prize at a tombola), and the newspaper is never wrong, so the probability of H:'Harry has won' is now one, $\mathbb{P}(H) = 1$. After that, a radio also assumed to be 100% reliable announces y:'Harry has won'. According to Achinstein (2001, 2010), the latter should count as evidence for H, but the probability of 1 cannot be "B-boosted" anymore. Mayo (2018) agreed.

10.3 Reconstructing the Critiques to Inductive Reasoning

The above arguments pose a challenge for Bayesian inference interpreted as probabilistic affirming the consequent. While the search for the probability measure $\mathbb{P}$ on the parameter space Θ is solved from a modern measure-theoretic perspective, the three open critiques given are the counterexample of Popper (1959) (in its correct original form, not Mayo's interpretation) and the two additional arguments of Mayo (2018). In the following, it is shown that all three arguments against probabilistic affirming the consequent do not hold.

10.3.1 Reconstructing Popper's B-Boost fallacy from a Bayesian perspective

In Mayo's terms, Popper's counterexample shows that probability and confirmation cannot be used synonymously because the absolute and relative B-Boost can yield different conclusions. Before reconstructing Popper's example, it needs to be stressed that both the relative and absolute B-Boost are no standard concepts ever used in Bayesian inference, and neither parameter estimation nor hypothesis testing is based on the relative or absolute B-Boost in practice.

While Popper thought he had found a counterexample to reject probabilistic affirming the consequent in the form of Bayes' Theorem, his example is just exploiting the fact that the posterior probability and the ratio of posterior to prior probability do not necessarily need to coincide. It is important to make a few distinctions before discussing his example from a Bayesian perspective.

First, as Popper's original "statements" all operate only in the sample space, a full Bayesian analysis is not possible. Popper's treatment includes no specifics about the statistical model $\mathcal{P}$, the unknown parameter(s) θ and the hypothesis to be tested. For a full Bayesian analysis it is necessary to make some assumptions about the statistical model, the unknown parameter of interest and the prior distributions. Here, it is assumed that the generated data of the dice are distributed as Binomial with $n = 1$ and parameter $\theta \in [0, 1]$, where θ is the probability of obtaining a six. Two hypotheses are compared then for illustration: The hypothesis $H : p = 1$ which states that the dice always yields a six, and $\sim H : p \neq 1$ which states that the dice does not always yield a six. These two hypotheses are primarily introduced to demonstrate that a difference as observed by Popper in his example between what Mayo calls an absolute and relative B-Boost is natural and even required for a probability measure to be a desirable method for quantifying the degree of corroboration of a hypothesis.

Second, Popper's original example only uses the model distribution for the observed data which operate on the sample space. His criticism thus pertains primarily to probability measures without any reference to Bayesian inference, but his conclusion is that due to his paradoxical results all methods based on inductive reasoning and probability measures need to be rejected. As these include Bayesian methods, Popper's original example is taken by Mayo to demonstrate that enumerative probabilistic induction and also Bayesian inference must be rejected. However, the reconstruction above shows that the behaviour observed by Popper is indeed required and no contradiction to probability as a method for quantifying the degree of corroboration. As a consequence, Bayesian statistics does not suffer from the "paradoxical behaviour", too.

Now, we reconstruct Popper's example from a fully Bayesian perspective. The first option to reconstruct Popper's example in a fully Bayesian approach is the hypothesis testing stance: In it, prior probabilities $\mathbb{P}(H)$ and $\mathbb{P}(\sim H)$ are used with the constraint $\mathbb{P}(H) + \mathbb{P}(\sim H) = 1$. The hypothesis $H : p = 1$ is compared with $\sim H : p \neq 1$ for illustration. Although Popper introduces no hypothesis, the introduction of these hypotheses will demonstrate why his conclusion does not hold. The prior probabilities $\mathbb{P}(H) = \mathbb{P}(\sim H) = 1/2$ are assigned to the hypotheses, as nothing else is known about the dice. After fixing the prior probabilities, Bayesian hypothesis testing can be conducted. While Bayesian hypothesis testing is most often concerned with the Bayes factor, here the posterior probabilities are used to follow Popper's concept of confirmation, that is, Mayo's B-Boosts. Therefore, two statistical models M_1 and M_2 are compared, each incorporating one of both hypotheses $H : p = 1$ vs. $\sim H : p \neq 1$. Thus, model M_1 corresponds to the hypothesis $H : p = 1$, where the probability p of obtaining a six with the dice is set to one. Model M_2 corresponds to the complement $\sim H : p \neq 1$ of H, that is to the model where the probability p of obtaining a six with the dice takes any other value than one. The ratio of posterior model probabilities can be calculated as

$$\frac{\mathbb{P}(M_1|x)}{\mathbb{P}(M_2|x)} = \frac{\mathbb{P}(x|M_1)}{\mathbb{P}(x|M_2)} \cdot \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)}$$

In model M_2 a uniform prior $p \sim \mathcal{U}(0,1)$ is used, as nothing is known about the true parameter p. In model M_1, a Dirac-prior with density $f(p) = \mathbb{1}_{\{p=1\}}(p)$ is used, as one is certain that the probability p of the dice yielding a six is $p = 1$. Now, Popper's z amounted to observing an even number. Suppose that this number is a six, so $X = 1$. Then, the binomial likelihood in M_1 can be written as

$$\mathbb{P}(X = 1|M_1) = \binom{1}{1}(1)^1(1-1)^{1-1} = 1$$

where the observed data x are modeled as a random variable X counting the successes (a six occurs). The binomial likelihood for $\mathcal{M}_2$ can be written as

$$\mathbb{P}(X = 1|M_2) = \int_0^1 \binom{1}{1} p^1(1-p)^{1-1} dp = \int_0^1 p dp = 1/2$$

as now the parameter p is unknown. Assuming the prior probabilities $\mathbb{P}(M_1) = \mathbb{P}(M_2) = 1/2$, the ratio of posterior model probabilities can now be calculated as

$$\frac{\mathbb{P}(M_1|X=1)}{\mathbb{P}(M_2|X=1)} = \frac{\mathbb{P}(X=1|M_1)}{\mathbb{P}(X=1|M_2)} \cdot \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} = \frac{1}{1/2} \cdot \frac{1/2}{1/2} = 2$$

Due to the constraint $\mathbb{P}(M_1|x) + \mathbb{P}(M_2|x) = 1$ it follows that $\mathbb{P}(M_1|x) = 2/3$ and $\mathbb{P}(M_2|x) = 1/3$. Therefore, from the perspective of the absolute B-Boost, $H : p = 1$ – or model M_1 – is confirmed, because

$$2/3 = \mathbb{P}(M_1|x) = \mathbb{P}(H|x) > \mathbb{P}(\sim H|x) = \mathbb{P}(M_2|x) = 1/3$$

while $\sim H$ – or model M_2 – is disconfirmed.

Now, we discuss the relative B-Boost. We obtain

$$2/3 = \mathbb{P}(M_1|x) = \mathbb{P}(H|x) > \mathbb{P}(H) = \mathbb{P}(M_1) = \frac{1}{2}$$

which shows that H – or M_1 – is confirmed when using the relative B-Boost for interpretation, too. Furthermore, $\sim H$ – or M_2 – is disconfirmed from the perspective of the relative B-Boost too, because

$$1/3 = \mathbb{P}(M_2|x) = \mathbb{P}(\sim H|x) < \mathbb{P}(\sim H) = \mathbb{P}(M_2) = \frac{1}{2}$$

Summing up, when following a Bayesian interpretation, no contradiction occurs at all for the relative and absolute B-Boost in the example of Popper (1959). When the observed even number is not a success, that is, $X = 0$ instead of $X = 1$, we arrive at $\mathbb{P}(X = 0|M_1) = 0$ and $\mathbb{P}(X = 0|M_2) = \frac{1}{2}$ so that $\mathbb{P}(M_1|X = 0) = 0$ and $\mathbb{P}(M_2|X = 0) = 1$ (after observing a single throw which is not a six the probability of $H : p = 1$ reduces to zero) the absolute B-Boost shows confirmation of M_2 or $\sim H : p \neq 1$ because

$$0 = \mathbb{P}(M_1|X = 0) < \mathbb{P}(M_2|X = 0) = 1 \qquad (10.8)$$

and the relative B-Boost

$$\frac{\mathbb{P}(M_1|X=0)}{\mathbb{P}(M_1)} = \frac{0}{\frac{1}{2}} = 0 < \frac{\mathbb{P}(M_2|X=0)}{\mathbb{P}(M_2)} = \frac{1}{\frac{1}{2}} = 2 \tag{10.9}$$

signals the same confirmation of M_2. So no matter which even number we observe, the absolute and relative B-Boost behave identically.

However, it could be argued that the probability of the hypotheses H and $\sim H$ was balanced in this reconstruction, and in Popper's original example we had $p(x) = 1/6$ versus $p(y) = 5/6$, and instead of H and $\sim H$ the events x and y were compared in terms of the absolute and relative B-Boost. For completeness and to show where Popper's and Mayo's error lies, we change the prior probabilities of H and $\sim H$ to $\mathbb{P}(H) = \frac{1}{6}$ and $\mathbb{P}(\sim H) = \frac{5}{6}$ so that they align with Popper's original example. Note that although Popper's probability measure operated on the sample space, the situation is identical: A probability measure can be used to quantify the ratio of conditional probability to unconditional probability (or posterior to prior probability in a Bayesian interpretation) or to quantify only the conditional probabilities.

Suppose now the prior probabilities $\mathbb{P}(H) = \frac{1}{6}$ and $\mathbb{P}(\sim H) = \frac{5}{6}$ are used which implies prior probabilities of $1/6$ and $5/6$ for M_1 and M_2 are employed. When the observed even number is a six, that is a success and $X = 1$, this leads to posterior probabilities $\mathbb{P}(M_2|x) = 0.71428$ and $\mathbb{P}(M_1|x) = .28572$, because

$$\frac{\mathbb{P}(M_1|X=1)}{\mathbb{P}(M_2|X=1)} = \frac{\mathbb{P}(X=1|M_1)}{\mathbb{P}(X=1|M_2)} \cdot \frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)} = \frac{1}{1/2} \cdot \frac{1/6}{5/6} = \frac{2}{5} \tag{10.10}$$

Using

$$\frac{\mathbb{P}(M_1|X=1)}{\mathbb{P}(M_2|X=1)} = \frac{2}{5} \Leftrightarrow \frac{\mathbb{P}(M_1|X=1)}{1-\mathbb{P}(M_1|X=1)} = \frac{2}{5} \Leftrightarrow \mathbb{P}(M_1|x) = 0.28572 \tag{10.11}$$

the relative B-Boost shows that $\mathbb{P}(M_1|x) = 0.28572 > 1/6 = \mathbb{P}(M_1)$ and $\mathbb{P}(M_2|x) = 0.71428 < 5/6 = \mathbb{P}(M_2)$, confirming M_1 and disconfirming M_2. The absolute B-Boost now leads to Popper's 'contradiction' that $\mathbb{P}(M_2|x) = 0.71428 > 0.28572 = \mathbb{P}(M_1|x)$, confirming M_2 instead of M_1 now.

So where is the mistake Popper made? First, Popper did not clearly specify the statistical models under consideration, which may contribute to the paradoxical result he obtained. However, more importantly, when the statistical models Popper used are formally defined as shown above, it is revealed that whenever the prior probabilities for the hypotheses under consideration are balanced, no contradiction occurs between the absolute and relative B-Boost. The reason therefore is that in modern terms, the absolute B-Boost is the posterior odds, which are meaningless without normalisation by the prior odds, compare Equation (6.12). Thus, whenever the prior probabilities are balanced, the posterior odds are *equal* to the Bayes factor which precisely quantifies the relative change in beliefs towards the hypothesis H or $\sim H$.

When unbalanced prior probabilities were used, Popper's paradoxical result occurs. The reason is that when an unreasonable bulk of prior probability mass of $5/6$ is placed on model M_2, or $\sim H : p \neq 1$, a single observation will not change the posterior distribution substantially. Therefore, the relative B-Boost yielded that H gets confirmed by the data, or in Poppers terms $1/3 = \mathbb{P}(H|x) > \mathbb{P}(H) = 1/6$. The absolute B-Boost led to $1/3 = 2/6 = \mathbb{P}(H|x) < \mathbb{P}(\sim H|x) = 4/6$ and this led Popper (1959) to reason that inductive reasoning is flawed and has to be rejected, because the absolute B-Boost now confirms $\sim H$ instead of H. The reason the absolute B-Boost did confirm $\sim H$ instead of H in Popper's calculations lies in its very definition: The absolute B-Boost compares the posterior probabilities $\mathbb{P}(H|x)$ and $\mathbb{P}(\sim H|x)$, see Definition 10.2. In the Bayesian reconstruction with Popper's unbalanced probabilities above, the same prior probabilities for H and $\sim H$ were used which Popper used for his events x and y without any reference to the Bayesian approach. Thus, the situation is identical: While in the Bayesian reconstruction, the prior probabilities strongly favour the alternative $\sim H$, Popper (1959) picked an example where nearly all of the available probability mass of his probability mass function p is assigned to the event y, namely $5/6$. From a Bayesian perspective, as the posterior is produced as a product of likelihood and prior via Bayes' theorem, the posterior of $\sim H$ is barely changed by only a *single observation*. In Popper's original example, there is no Bayesian interpretation to x or y, but the probability measure behaves identical because the events x and y have such a different probability mass. Thus, the conditional probability in Popper's example gets smaller (from $5/6$ to $4/6$), but does not change substantially. In the Bayesian reconstruction, the behaviour is identical: As the prior distribution assigns $5/6$ of the probability mass to $\sim H$, the posterior is strongly influenced by the prior. Therefore, the posterior $\mathbb{P}(H|x)$ is not larger than $\mathbb{P}(\sim H|x)$, which is to be expected due to the substantial mass of prior probability put on $\sim H$. The behaviour of Popper's example is indeed to be expected: From a frequentist perspective, had Popper (1959) chosen an example where the events do not differ so drastically in regard to their probability, the paradox would have disappeared. From a Bayesian perspective, assigning reasonable prior probability like $\frac{1}{2}$ to each model, the absolute B-Boost and relative B-Boost show identical behaviour.

Summing up, to assign the bulk of available probability mass on one of both models and then being baffled by the fact that a *relative comparison* of the prior and posterior probability in the relative B-Boost does not reflect the same behaviour as an *absolute comparison* via the absolute B-Boost is no flaw in inductive reasoning, it is just exploiting the fact that the absolute B-Boost is of course sensible to the prior probabilities of each model (or hypothesis) under consideration. While Popper's original example has no Bayesian interpretation, the situation is analogue to the Bayesian reconstruction because the events x and y are selected identically as the hypotheses H and $\sim H$ in the reconstruction. From a Bayesian point of view, the absolute B-Boost in modern notation is simply the posterior model odds. These need to be normalised by the prior odds to yield the Bayes factor, which in this case then states only anecdotal evidence due to a single observation. Using the absolute B-Boost, which is the posterior model odds *without* normalisation of the prior odds, will lead to the strong influence of the prior odds. The absolute B-Boost will not reveal the correct evidence for $\sim H$ when extreme priors and tiny sample sizes are used. Still, from a Bayesian perspective, had Popper collected more and more data, the absolute B-Boost would finally have overcome the strength of the extreme prior probabilities. Then, even the absolute B-Boost would break the imbalance of the assigned prior probabilities, and eventually end up indicating evidence for $\mathcal{M}_1$ (or H) in the same way the relative B-Boost does.[3] The lesson from the phenomenon observed by Popper is that extreme priors should be avoided when the amount of collected data is very limited.

Returning to the problem of inductive reasoning, from an explicit hypothesis testing

[3] Such a resolution of the difference between absolute and relative B-Boosts is, of course, not attained when a strictly frequentist perspective is taken.

perspective (1) Popper (1959) did not show that inductive reasoning leads to contradictory conclusions. He only discovered that extreme priors require (1) enough data to be overcome, and (2) the absolute B-Boost is a controversial measure for extremely subjective inductive (or Bayesian) inference. As from a modern perspective, subjective Bayes has become a niche, and the commonly agreed on paradigm is the objective or weakly informative Bayes – see also Held and Sabanés Bové (2014); Wagenmakers et al. (2018); Ly et al. (2016a,b); Carlin and Louis (2009); Gelman et al. (2013); Kelter (2020b) – this causes few problems. Also, the relative B-Boost is proportional to the marginal likelihood, and the absolute B-Boost is the posterior model odds. When normalising the latter with the prior odds, the resulting Bayes factor yields the evidence about H and $\sim H$. Popper's example only works because he is using a single observation combined with extreme prior probabilities and both aspects can be fixed easily. Even when not fixing the extreme priors, these will eventually be overwhelmed by increasing the amount of collected data. Using a single observation for testing a hypothesis is absurd in any real research setting.

Another important point is the apocryphal notation of Mayo (2018), which shows that the connection of their 'B-Boosts' to elementary Bayesian objects which already have a definition and name were not recognized. The relative B-Boost is proportional to the marginal likelihood of the corresponding model, and the absolute B-Boost is simply the posterior model odds. To see this, note that the relative B-Boost $\mathbb{P}(H|x) > \mathbb{P}(H)$ thus is equivalent to $\mathbb{P}(H|x)/\mathbb{P}(H) > 1$ which in turn holds if and only if

$$\underbrace{f(x|H)/f(x)}_{\propto f(x|H)} > 1$$

according to Equation (6.12), when the hypothesis H is identified with a model M_1. Therefore, from a Bayesian perspective, the relative B-Boost is proportional to the marginal likelihood $f(x|M_1)$ of the corresponding model M_1 (analogue for model M_2). Values larger than one indicate the necessity of a change in belief towards H, which corresponds to M_1. The absolute B-Boost from a Bayesian perspective is simply the posterior odds without incorporation of the prior odds according to Equation (6.12): $\mathbb{P}(H|x) > \mathbb{P}(\sim H|x)$ is equivalent to

$$\underbrace{\frac{\mathbb{P}(H|x)}{\mathbb{P}(\sim H|x)}}_{\text{posterior odds}} > 1$$

What is more, incorporation of the prior odds is essential and necessary, because the prior odds strongly influence the posterior odds as the prior odds are updated by multiplication with the Bayes factor to yield the posterior odds. In Popper's example the probabilities of x and y were chosen identical to the prior odds of H and $\sim H$ in the Bayesian reconstruction, $\mathbb{P}(H)/\mathbb{P}(\sim H) = (1/6)/(5/6) = 1/5$. Therefore, normalising the posterior odds $\mathbb{P}(H|x)/\mathbb{P}(\sim H|x) = (1/3)/(4/6) = 1/2$ with the prior odds yields a Bayes factor of $BF_{01} = (1/2)/(1/5) = 2.5$, which indicates only weak evidence for the null hypothesis $H : p = 1$ which is bare worth mentioning according to Lee and Wagenmakers (2013), Held and Ott (2018), and Jeffreys (1961), see Table 6.1. The absolute B-Boost can easily be disregarded from any further discussion based only on the above explanation. Basing any inference on solely the posterior model odds without incorporation of the prior model odds misses the point completely. The relative

B-Boost $\mathbb{P}(H|x)/\mathbb{P}(H)$ as shown above equals the marginal model likelihood $f(x|H)$ which quantifies the predictive ability of the model to predict the data x. The Bayes factor compares the predictive ability of two hypotheses H and $\sim H$ (or equivalently, of the two models M_0 and M_1) exactly via the ratio of these marginal model likelihoods, and thereby expresses the necessary change in beliefs towards either of both models. What Popper's example and Mayo's interpretation in fact show is the difference between objective and subjective Bayesian inference: A subjective Bayesian statistician will incorporate subjective prior odds into the analysis and report the posterior odds after computing the Bayes factor and multiplying it with the prior odds. An objective Bayesian will prefer to solely report the Bayes factor, because the Bayes factor is, in general, independent of the prior odds (Kleijn, 2022, Lemma 2.6).

The above analysis showed that Popper's example can be shown to provide no relevant objection to probability as a method for quantifying the confirmation of a hypothesis. However, the example is still insightful as it sheds light on another important issue: It remains open how to use the posterior distribution in the Bayesian approach to test hypotheses.[4] By now, the example of Popper (1959) has been reconstructed in from the perspective of hypothesis testing. However, as there is just a single observation, one could also argue for an approach which prefers parameter estimation under uncertainty instead. Neither a frequentist nor a Bayesian would accept a single observation as strong evidence for any hypothesis, so this fact directly motivates the second perspective, which embraces uncertainty and tries only to estimate the parameter p of the dice. In what follows, Popper's example is also reconstructed from this second perspective.

In this second perspective no explicit model comparison is conducted. Therefore, the inference is concerned with estimating the parameter p and inferring, which values are most probable a posteriori. Reframing Popper's example into correct statistical terminology, one would use a prior $\theta \sim \mathcal{B}eta(\alpha, \beta)$ where p is replaced by θ for notational convenience. The prior for the probability $\theta \in [0, 1]$ of obtaining a six can be shown to be completely uninformative and equivalent to a uniform prior $p \sim \mathcal{U}(0, 1)$ when choosing $\alpha = \beta = 1$. Thus, this choice of a $\mathcal{B}eta(1, 1)$ prior distribution reflects no preference for any value of θ inside $[0, 1]$. The likelihood $f(x|\theta)$ would be modelled as binomial, where it actually reduces to a Bernoulli likelihood because $n = 1$ (single dice toss). The posterior is easily shown to be again Beta distributed with updated parameters $\mathcal{B}eta(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$, see for example Gelman et al. (2013), which here becomes $\mathcal{B}eta(1 + 1, 1 + 1 - 1) = \mathcal{B}eta(2, 1)$. Figure 10.1 now shows two reconstructions of Popper's example, where the parameter p has been replaced by θ for notational convenience. The upper row shows the uniform $\mathcal{B}eta(1, 1)$ prior, the Bernoulli likelihood $f(\theta|x)$, and the $\mathcal{B}eta(2, 1)$ posterior (regard the different y-scale) after observing Popper's data $x = 1$, that is, a single six. Note that in Popper's original formulation, only an even number was observed, so actually $x = \{2, 4, 6\}$. When using this original model, the Dirichlet model below is required. In the model considered here, we suppose that a six is observed, and then the resulting posterior is $\mathcal{B}eta(2, 1)$. If a two or a four is observed instead of a six, the resulting posterior will be $\mathcal{B}eta(1, 2)$, because then $\sum_{i=1}^n x_i = 0$, that is, we observe no success (where a six equals a success). So perspective (2) would proceed by first deriving such a posterior and after that, estimate the most probable values of θ given the data x. For example, estimation under uncertainty would become calculating a 95% highest-density-interval for θ and subsequently judg-

[4] A detailed comparison of available Bayesian evidence measures which helps answering this remaining question is provided in Chapter 14.

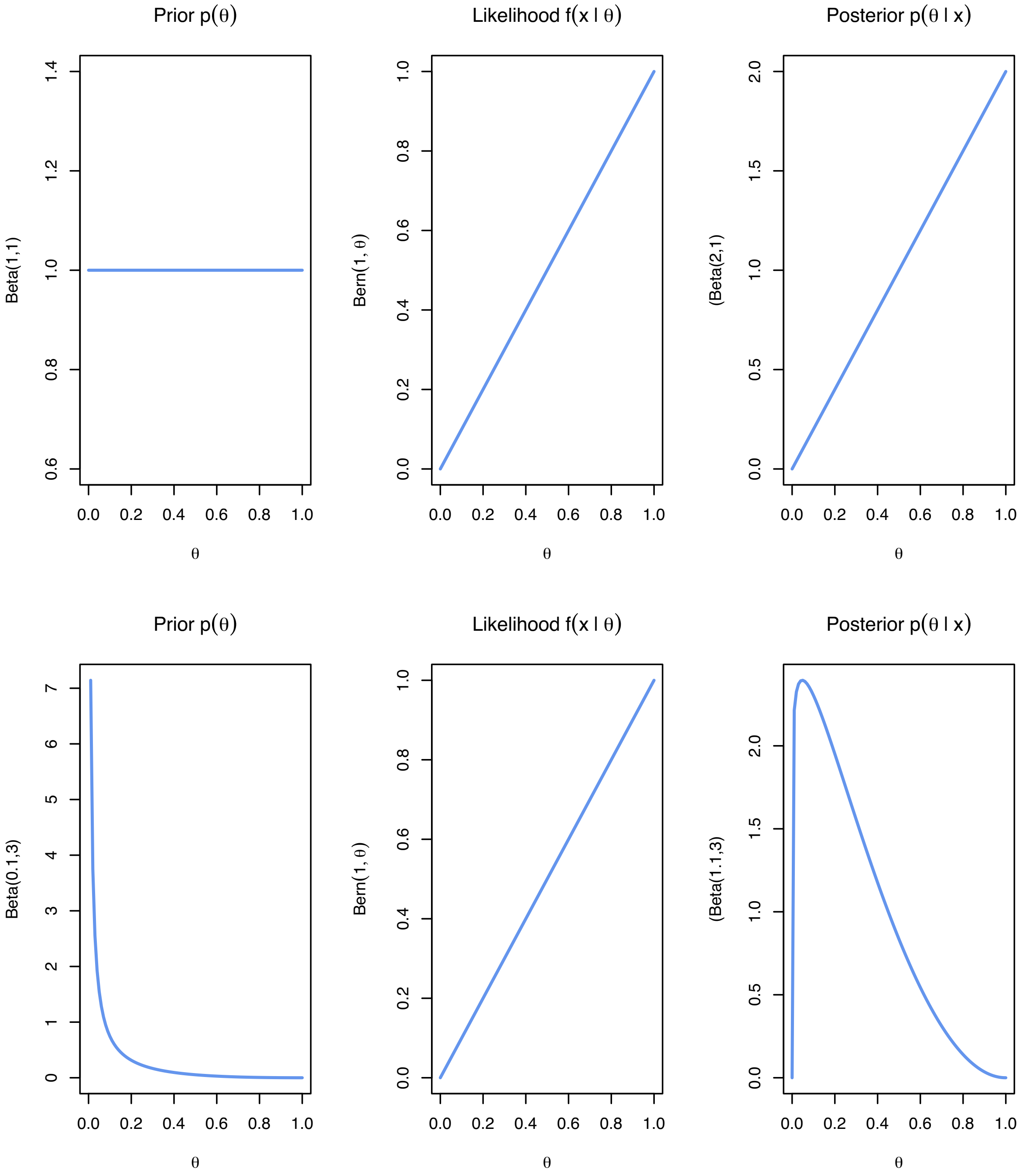

Figure 10.1: Reconstructing Popper's dice toss example against inductive inference with the beta-binomial model

ing if $\theta = 1$ is inside or outside the credible interval. Explicit testing of the hypothesis $H : \theta = 1$ against $\sim H : \theta \neq 1$ is thus avoided. In this second perspective where an explicit model comparison is avoided, neither inductivists nor Bayesians explicitly use B-Boosts (that is, marginal likelihoods or posterior odds) when conducting inference. Instead, simple estimation under uncertainty is used to draw conclusions. The lower row of Figure 10.1 shows the situation under a different prior where one assumes a priori that the dice is likely to be unfair. This results, of course in a different posterior distribution and shows that no inductivist would be baffled by the amount the beliefs change from the prior to the posterior distribution of θ. Termed differently: A single toss of a dice is poor evidence, no matter if perspective (1) favouring hypothesis testing or perspective (2) favouring estimation under uncertainty is applied.

Another option to reconstruct Popper's example is even more realistic and elementary: As the experiment uses a dice with six sides, a much more plausible statistical model for Bayesian inference would require using a vector $p = (p_1, ..., p_6)$ of probabilities for each dice face. Then, Popper's original data $x = \{2, 4, 6\}$ can be modelled even more explicitly as now each dice face is separated. The standard Bayesian model in this setting would be the Dirichlet-Multinomial-model, see Gelman et al. (2013, Chapter 3). In this model, a noninformative Dirichlet prior $p \sim \mathcal{D}ir(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6), \alpha_i > 0$ is combined with a multinomial likelihood $\mathcal{M}(p_1, ..., p_6)$ with $p_i \geq 0$ and $\sum_{i=1}^{6} p_i = 1$ for the six faces of the die. The above Dirichlet prior can be shown to be equivalent to a uniform prior on the probability vector $p = (p_1, ..., p_6)$ for each dice face (Gelman et al., 2013, Chapter 3), and is a conjugate prior to the multinomial likelihood. Therefore, it follows from standard Bayesian theory that the posterior is given as $f(p_1, ..., p_6|x) = \mathcal{D}ir(\alpha_1 + x_1, \alpha_2 + x_2, \alpha_3 + x_3, \alpha_4 + x_4, \alpha_5 + x_5, \alpha_6 + x_6)$, which is again Dirichlet distributed with updated parameters $\alpha_i + x_i$, see Gelman et al. (2013). Here $x_1, ..., x_6$ are the observed number of faces one to six in the sample. Thus, when the dice is rolled for example two times and a six and a four are observed, $x_6 = 1$ and $x_4 = 1$, and all other $x_i = 0$. Note that in the posterior $f(p_1, ..., p_6|x)$, x now is the vector of results obtained. In Popper's example, either a 2, a 4 or a 6 is observed. Thus, when a six is observed in Popper's example, $x = (x_1, x_2, x_3, x_4, x_5, x_6) = (0, 0, 0, 0, 0, 1)$, that is a single toss yields a six. Using a Dirichlet prior which is equal to the uniform prior, that is $f(p) \sim \mathcal{D}ir(1, 1, 1, 1, 1, 1)$, the posterior is given as

$$f(p_1, ..., p_6|(0, 0, 0, 0, 0, 1)) = \mathcal{D}ir(1, 1, 1, 1, 1, 2)$$

The conclusions are then identical to the previous model. When a 2 or a 4 are observed, the posterior changes accordingly.[5] Summing up, the argument of Popper (1959) collapses in both models. Why is that? First, all of his reasoning depends on minimal data and extreme priors in combination with the absolute B-Boost, which is strongly influenced by the extreme prior. In a hypothesis testing perspective (1), balanced priors are more appropriate and eliminate the "problem", and if extreme priors are reasonable, Popper only discovered that larger amounts of data are needed to overcome such prior assumptions. Then, the evidence stated by the absolute and relative B-Boost is reconciled eventually. Perspective (2) embraces estimation under uncertainty and makes use of the whole *posterior distribution*. In a second step, a credible interval, the Bayes factor or even point estimates like the posterior mean or median can be computed. Based on the tiny amount of data, estimation under uncertainty is the more realistic perspective here. Also, the absolute B-Boost is not used in practice and needs to be normalised with the prior odds. By this normalisation the Bayes factor is obtained, which states only anecdotal evidence for both hypotheses. As Jeffreys (1939) already noted, strong a priori beliefs like 1/6 to 5/6 for $H : p = 1$ need large amounts of data to be overcome. The Bayes factor highlights that a single dice throw which yields a six necessitates only an anecdotal change in beliefs about the hypotheses, given the extreme a priori beliefs. Popper's counterexample therefore only demonstrates the difference between subjective and objective Bayesians, where the former prefer to report the ratio of posterior to

[5] One could also roll the dice three times and obtain the sample $\{2, 4, 6\}$ and the posterior would then be $f(p_1, ..., p_6|(0, 1, 0, 1, 0, 1)) = \mathcal{D}ir(1, 2, 1, 2, 1, 2)$.

prior odds and the latter prefer the Bayes factor. The contradiction is therefore in both cases removed, and probabilistic affirming the consequent encompassed by Bayes' theorem must not be rejected based on Popper's counterexample to probability as a measure of corroboration.

10.3.2 Reconstructing Mayo's and Achinstein's irrelevant conjunctions paradox

The first of the other two arguments of Mayo (2018) against probabilistic affirming the consequent was based on the fact that it is possible to attach irrelevant hypotheses J to a given hypothesis H, and if x confirms H then it also confirms $H\&J$.

Formally, there is nothing wrong with the argument, but the example is inappropriate for a statistical context. No researcher with common sense would accept inferring a statement about two completely independent hypotheses, chiefly when the experimental design, which Mayo (2018) hides in her discussion, is not concerned with one of the hypotheses. For example, in her reasoning, either the deflection of light is measured or the radioactivity levels in Fukushima water being dumped in the pacific ocean. Thus, attaching irrelevant conjunctions and still stating evidence for the conjunction $H\&J$ is to treat scientists as people with no common sense, not being able to connect the statistical inferences made with the data collected or observed. Problematically, what Mayo and Achinstein define as irrelevant is an arbitrary definition: $\mathbb{P}(x|H) = \mathbb{P}(x|H\&J)$ does not imply that J is irrelevant, it only implies that x has the same probability to be observed when H holds and when both H and J hold. Whether J is *relevant* or not is implied by the definition of J: If J is a statement as $\theta \in \Theta_0$ with Θ_0 a subset of the parameter space Θ, then the probability of observing data x given $H\&J$ will be influenced and the central assumption $\mathbb{P}(x|H) = \mathbb{P}(x|H\&J)$ does not does not hold anymore. Whenever J makes no statement about the parameter θ under consideration, it is indeed irrelevant and can be excluded from any further analysis. Problematically, as any hypothesis is a subset of the parameter space both in the frequentist and Bayesian paradigm[6] an irrelevant conjunction needs to be a null-set with respect to the dominating measure of the statistical model $\mathcal{P}$. This immediately shows that J can be omitted from further analysis.

What is more, that the same argument (if taken to be valid) could be applied to deductive reasoning. Conducting a frequentist hypothesis test (in Fisher's significance framework of the Neyman-Pearson theory) which rejects the hypothesis that the deflection effect in Einstein's theory of general relativity is 1.75 allows at the same time to attach the irrelevant conjunction that the Fukushima water being dumped in the pacific ocean is within acceptable levels. The 'problem' remains the same as when using inductive reasoning, but the problem is no problem at all when the statements made by a hypothesis are restricted to be concerned with the statistical model $\mathcal{P}$ under consideration, compare Kleijn (2022).[7]

Another crucial aspect is that the definition of an irrelevant conjunction is debatable. It could well be argued that the irrelevance of a conjunction should be defined by $\mathbb{P}(H|x) = \mathbb{P}(H\&J|x)$, because then the data x predict H equally well as $H\&J$. This

[6] Compare Appendix C.

[7] This holds both in the parametric and nonparametric situation, see Kleijn (2022, Chapter 1).

invalidates the above reasoning because based on

$$\frac{\mathbb{P}(H|x)}{\mathbb{P}(H)} = \frac{\mathbb{P}(H\&J|x)}{\mathbb{P}(H)} > 1 \tag{10.12}$$

H is confirmed in the sense of a relative B-Boost, but one cannot infer

$$\frac{\mathbb{P}(H\&J|x)}{\mathbb{P}(H\&J)} > 1 \tag{10.13}$$

anymore. However, the latter inequality is precisely the relative B-Boost for H and the irrelevant conjunction J. Thus, while H is still 'B-Boosted' as shown in Equation (10.12), $H\&J$ are not 'B-Boosted' anymore as shown by Equation (10.13). The reason is that we cannot substitute the numerator in Equation (10.12) anymore based on the new definition $\mathbb{P}(J|x) = \mathbb{P}(H\&J|x)$ of an irrelevant conjunction.

10.3.3 Reconstructing the inability of B-Boosts to update the evidence for perfectly reliable sources

The second argument of Mayo (2018) followed Achinstein (2001, 2010), who criticised that a B-Boost is not possible when a 100% reliable information has been provided. This whole problem vanishes because the situation described by Achinstein (2001) requires no inference at all. Talking about B-Boosts becomes useless: If the information from either the newspaper or the radio is 100% reliable, this can be interpreted as observing the whole population instead of a sample, and the information is fully available. In this case, no testing is needed anymore, and no statistical inference, because in enumerative induction the assumption was that only a proper subset $A_1, ..., A_n$ of all A's had been observed:

Enumerative Induction
Premise: All observed $A_1, A_2, ..., A_n$ have been B's.
Conclusion: Therefore H : all A's are B's.

If Achinstein (2001) argues that the radio or newspaper is absolutely reliable, this means in terms of enumerative induction that all A's have been observed, which constitutes the whole population. In this case, looking at all A's and checking if they are B's suffices. In the newspaper example, no more evidence is needed after getting the 100% reliable statement that Harry has won by the newspaper in the first place: When the tombola results are available and the newspaper has checked all tickets, this implies the whole population (of submitted tickets) is available. The whole inferential situation collapses at this point, as now statements can be made with certainty, and no *statistical* inference – in the form of B-Boosts, hypothesis tests, Bayesian or frequentist procedures, deductive or inductive reasoning – is required anymore.

10.4 Conclusion

The previous section showed that the three main arguments against probabilistic affirming the consequent do not hold. This situation makes it possible to use Bayes' theorem as an implementation of probabilistic affirming the consequent, and solves the problem of enumerative induction by switching to its probabilistic version. Also, tacking on irrelevant conjunctions is artificial and the argument remains questionable due to the arbitrary definition of an irrelevant conjunction, while also holding for frequentist testing if taken to be valid. The inability of B-Boosts to update the posterior in case of reliable resources can be attributed to the fact that the whole population is available in such cases which questions the point in applying statistical inference, and Poppers counterexample is just a lesson in extremely subjective priors[8], which are not overcome by a single observation. Also, this inability shows that the ratio of marginal likelihoods and the posterior model odds do not yield the same conclusions, which is why an "objective" Bayesian would normalise the latter is in practice by dividing the posterior model odds through the prior model odds and then obtain the Bayes factor (which is the ratio of the marginal likelihoods). A "subjective" Bayesian would prefer to report solely the posterior model odds based on his prior model odds. Together, these problems seem therefore constructed and artificial. Even worse, they do not argument correctly and provide no substantive reason against probability as a measure of confirmation of a statistical hypothesis.

Mayo (2018) primarily used these arguments to push her error statistical account including severe testing, which is founded on Poppers deductivism and therefore biased to blame induction as false. Reviews of her recent book *'Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars'* (Mayo, 2018) can be found in Gelman et al. (2019). From a statistical perspective, her ideas of severe testing are too vague to be implemented. From a philosophical perspective, she only discredits probabilistic affirming the consequent by (incorrectly) repeating the criticism of Popper (1959) (which originally had no direct relationship to Bayesian inference but to contemporary probability theory in general) and adding the two arguments above.

Christian Robert described the ideas put forward by Mayo (2018) quite well:

> "I sort of expected a different content when taking the subtitle, How to get beyond the Statistics Wars, at face value. But on the opposite the book is actually very severely attacking anything not in the line of the Cox-Mayo severe testing line. (...) Another subtitle of the book could have been testing in Flatland given the limited scope of the models considered with one or at best two parameters and almost always a normal setting. I have no idea whatsoever how the severity principle would apply in more complex models, with e.g. numerous nuisance parameters. By sticking to the simplest possible models, the book can carry on with the optimality concepts of the early days, like sufficiency (p. 147) and monotonicity and uniformly most powerful procedures, which only make sense in a tiny universe."
> Christian Robert in (Gelman et al., 2019, p. 16).

[8] Or a lesson that subjective and objective Bayesian approaches can differ when extreme priors are chosen and limited data is observed, which was shown in the difference between the posterior model odds and the Bayes factor.

Chapter 11

Axiomatic Considerations on the Foundations of Statistical Inference

The only theories which are formally complete, and of adequate scope for treating statistical evidence and its interpretation in scientific research contexts, are Bayesian.

Alan Birnbaum
The anomalous concept of statistical evidence

Chapter 10 showed that from a philosophical perspective, there is no sound argument against using Bayesian inference as an implementation of probabilistic affirming the consequent (or probabilistic enumerative induction) by means of Bayes' theorem. Therefore, Bayesian inference can be seen as a grounded scientific theory from a philosophical perspective. Still, it is unclear whether the problems inherent in null hypothesis significance testing in the veins of Fisher, or hypothesis testing according to Neyman and Pearson are avoided by employing Bayesian inference. This chapter analyses a set of several important axioms underlying statistical inference to investigate this question. Similar to probability theory, one can start from some first principles which are assumed to be true, and successively derive further results with wider implications, in particular for practical data analysis. In this chapter it is shown that starting from basic principles of statistical inference inevitably leads to the (*relative*) *likelihood principle* (LP), and that the LP itself is in conflict with both Fisher's significance testing and the Neyman and Pearson theory of hypothesis testing. The implications of this result are profound, since based on these results, the dominating practice of statistical hypothesis testing in contemporary science stands in direct conflict with the axiomatic foundations of statistical inference. As discussed in Chapter 5, the Neyman-Pearson theory and Fisher's theory differ substantially, and as noted by Lehmann (1993), "specification of the appropriate frame of reference takes priority, because it determines the meaning of the probability statements." The conflict between unconditional Neyman-Pearson and conditional Fisher inference demonstrated that "a fundamental gap in the theory is the lack of clear principles for selecting the appropriate framework." (Lehmann, 1993, p. 1248).

The axiomatic analysis in this chapter therefore shows that many of the recent problems like failed replications of research results can be attributed to violations of the LP by both Fisher's significance tests and Neyman-Pearson tests. While Fisher's theory of significance testing will be shown to be more in line with a frame of reference that is adequate for scientific research, both frequentist theories violate principles which are themselves consequences of the LP like the *censoring principle* (CP) and *stopping rule principle* (SRP). After discussing the violation of the LP by classic frequentist hypothesis testing, it is shown that Bayesian inference can be interpreted as a natural implementation of the likelihood principle. As a consequence, Bayesian inference avoids the problems of null hypothesis significance testing in the spirit of Fisher or Neyman and Pearson (Birnbaum, 1962; Berger and Wolpert, 1988). As proposed by Lehmann (1993), this chapter thus shows that Bayesian inference is coherent with an important set of statistical principles from an axiomatic point of view, and provides an appropriate frame of reference for statistical hypothesis testing in scientific research.

11.1 Principles of Statistical Inference

Principles of statistical inference have a long tradition in statistical science, see Fisher (1955), (Popper, 1959), Jeffreys (1931), Cox (1958), and Birnbaum (1962). In mathematics, axioms are mostly structural ones in the sense that they provide the rules to work with like in group or set theory. Importantly, these structural axioms are not of normative nature in that they argue for or against how the mathematical objects should be used. In sharp contrast, principles of statistical inference can be interpreted as axioms which are strongly normative. They provide fundamental rules how to judge the evidence provided by data in a practical statistical analysis and thus are different from structural mathematical axioms. Nevertheless, these normative fundamental statistical axioms can be used to derive more farreaching results identically to non-normative structural mathematical axioms. Interestingly, in the early beginnings of modern statistics, neither Fisher, Neyman and Pearson nor Jeffreys started with clear statistical principles as shown in Part I and Part II. For example, Fisher's method of maximum likelihood and, in particular, the tests he developed with Gosset were justified mostly by mathematical intuition and the practical problems to be solved. As detailed in Section 3.2.3, Fisher's position was that his likelihood function,

> "when properly interpreted must contain the whole of the information respecting x which our sample of observations has to give."
> (Fisher, 1934c, p. 297)

While Fisher's idea that the likelihood function must give the whole of the information available in the data was appealing, there was no rigorous proof. The proof followed nearly three decades later and can be attributed to Birnbaum (1962), who showed in his landmark paper *'On the Foundations of Statistical Inference'* that the *Likelihood Principle*, intuitively assumed by Fisher, follows from the *Sufficiency Principle* and the *Weak Conditionality Principle* (compare Theorem 11.8). Birnbaum's work can be seen as the first structured attempt to clarify the axiomatic foundations of statistical inference while simultaneously discussing the concept of mathematical and statistical evidence in a broad sense. Historically, Fisher's intuition led him to the likelihood principle without formal proof, but he did not strictly adhere to it. He violated the likelihood principle himself when introducing his methodology of significance tests as described in Part I, Appendix C.6. By calculating his p-values, for example in the 2×2 tables given in

(Fisher, 1935), he based his inference on data which were not observed at all, namely on the probabilities of observing 11 or more monozygotic twins, also compare Equation (3.48). Therefore, his inference was not based solely on the likelihood function as required by the likelihood principle. It is therefore not difficult to recognize that the likelihood principle as proven by Birnbaum (1962) is not compatible with frequentist inference in the form of Fisher's significance tests.[1]

The first structured approach to principles of statistical inference can indeed be attributed to Birnbaum (1962), who showed that the likelihood principle follows from the more elementary sufficiency principle and conditionality principle in the discrete case. The sufficiency principle states that the evidence provided by a statistic which captures all information in the data without any loss is identical to the evidence provided by the original experimental data. The conditionality principle states that experiments not actually performed must be irrelevant to the conclusions drawn. The likelihood principle states that all evidence obtained from an experiment about an unknown quantity θ is contained in the likelihood function for θ for given data x. While the implications of the likelihood principle are farreaching, Berger and Wolpert (1988) stressed that the principle and its implications have "been ignored by most statisticians" (Berger and Wolpert, 1988, p. 1). Acceptance of the likelihood principle is maybe the central difference between Frequentists and Bayesians. Therefore, it is so important to clarify the foundations of statistics before arguing for or against a particular school of thought in the practical use of a method like hypothesis testing.

11.2 The Principle of Adequacy

Before the likelihood principle is discussed, this section outlines maybe the most elementary principle of statistical inference. It was introduced by Pratt (1977), who named it the *Principle of Adequacy* (AP). It is useful to quote an example to understand Pratt's reasoning[2]:

Example 11.1 (Berger and Wolpert (1988))**.** Suppose $H_0 : \theta = -1$ and $H_1 : \theta = 1$ are two competing hypothesis and the data are observed as $X \sim \mathcal{N}(\theta, .25)$. The rejection region $X \geq, 0$ of the Neyman-Pearson theory, gives a test with error probabilities (type I and II) of .0228. When $x = 0$ is observed, it is permissible to state that H_0 is rejected and that the corresponding error probability of a type I error is $\alpha = .0228$. Common sense, however, indicates that $x = 0$ fails to provide any evidence for or against one of both hypotheses, as it is located exactly between $\theta = -1$ and $\theta = 1$. On the other hand, suppose $x = 1$ is observed. Then pre-experimentally the Neyman-Pearson theory again can only state that $x = 1$ is in the rejection region $X \geq 0$ so H_0 can again be rejected at $\alpha = .0228$, but this time the evidence against $H_0 : \theta = -1$ seems overwhelming.

In particular, the type I error probability is calculated straightforward: The probability of a type I error is given as $\mathbb{P}(X \in [0, \infty)|H_0) = \int_0^\infty \varphi_{-1,.25}(x)dx = 1 - \Phi_{-1,.25}(0) = 1 - 0.9772499 = 0.0227501 \approx 0.228$. Here, $\varphi_{-1,.25}(x)$ is the probability density of the $\mathcal{N}(-1, .25)$ distribution, where $\mu = .25$ and $\sigma^2 = .25$, and $\Phi_{-1,.25}$ is the corresponding distribution function. Similarly, the type II error probability is calculated as

[1] As Fisher's exact test is an example of conditional inference, this shows that conditioning on an ancillary statistic as recommended by Fisher and Cox may be useful but can still violate the LP.

[2] For more details see (Berger and Wolpert, 1988, p. 17).

$\mathbb{P}(X \in (-\infty, 0)|H_1) = \int_{-\infty}^{0} \varphi_{1,.25}(x)dx \approx 0.228$. Clearly, the intuitive evidence obtained by observing x can be quite different from the pre-experimental evidence, which the Neyman-Pearson theory is targeted at. While the intuitive evidence of $x = 1$ against H_0 is much larger than the intuitive evidence against H_0 obtained from $x = 0$, the Neyman-Pearson test rejects H_0 in both cases at the $\alpha = .0228$ level without quantifying the strength of the evidence. The only goal is the type I error control, as indicated by the test level α. (Berger and Wolpert, 1988, p. 7) noted that this issue has "led many frequentists to prefer the use of P-values to fixed error probabilities". Interestingly, this is another reason why the hybrid of Fisher's significance testing and the Neyman-Pearson theory has evolved as detailed in Chapter 5. Many statisticians and practitioners were simply not satisfied by a measure which does not gauge the strength of the evidence, but only provides a binary decision threshold into significant and non-significant results. Therefore, the p-value was often used as a complement in addition to a conducted Neyman-Pearson test, even though this is not allowed as discussed in Chapter 5. Based on this situation, Pratt (1977) reasoned as follows:

> "Even for simple hypotheses, the question arises whether the tail probabilities (...) are to correspond to the particular data observed, or are to be fixed in advance with only 'accept' or 'reject' determined by the data. The latter seems to me clearly a very inadequate expression of the evidence. (The Principle of Adequacy: a concept of statistical evidence is (very) inadequate if it does not distinguish evidence of (very) different strengths.) This accords with the view that a F-value (critical level) is preferable to a report of 'significant' or 'not significant' in usual current practice where only tail probabilities under the null hypothesis are seriously considered in the final analysis."
> Pratt (1977, p. 62)

In the Neyman-Pearson theory it is not allowed to complement a test with a p-value, as there exists no concept of a p-value in the theory. Still, as shown in Chapter 5 researchers combined both theories and Fisher's p-value succeeded as the reported evidence measure, although Neyman-Pearson tests are widely used in practice. This shows how most scientists intuitively reject the dichotomous separation into 'significant' and 'not significant' in favour of a continuous measure of evidence that employs the size of the p-value (or test statistic). Pratt's principle, therefore, can be assumed as valid and is stated below.

Principle of Adequacy (AP). *A concept of statistical evidence is (very) inadequate if it does not distinguish evidence of (very) different strengths.*

Note that there are exceptions to this principle, like quality control, where the target is to minimise the number of defects, or the construction of medical tests, where the goal is to develop a test which will reliably indicate if a patient has a disease or not. Still, in the latter case, it can be discussed if long term type I error control as dictated by the Neyman-Pearson theory is desirable. The results of false-positive diagnosis of disease are revealed quickly in the routinely following examinations. In contrast, a false-negative result (a patient with the disease is told she is healthy) is much more severe in practice. Type II error control, therefore, may be preferable to type I error control in these situations.

11.3 The Likelihood Principle

The likelihood principle makes a statement about the evidence obtained by an experiment. More specific, it states that the evidence provided by data x for a parameter θ of the assumed statistical model depends solely on the likelihood function $\theta \mapsto L(\theta; x) = f_\theta(x)$. From a mathematical perspective, the likelihood principle (LP) makes a statement about settings in which the random variable X has density $f_\theta(x)$ with respect to some measure ν for all $\theta \in \Theta$. The likelihood function $L(\theta; x) = f_\theta(x)$ for $\theta \in \Theta$ as usual is the density evaluated at the observed data x and interpreted as a function of θ instead of x.

A preliminary note on notation: The notation and formulation of certain definitions and statistical principles differs between authors. Different influential notations have been used at least by Birnbaum (1962), Birnbaum (1972), Basu (1975), Dawid (1977), Kalbfleisch et al. (1986), Berger and Wolpert (1988) and Gandenberger (2015). For each definition or principle used in this chapter, a reference is provided immediately after its first statement to make clear which notation is used. As Berger and Wolpert (1988) provide the most coherent treatment of the likelihood principle up to date, most definitions and principles are taken from their notation.

11.3.1 Birnbaum's work on the foundations of statistical inference

The first proof of the likelihood principle was given by Birnbaum (1962), who derived it from the intuitively more plausible principles of sufficiency and conditionality. The limitation of his proof was that it does only hold for discrete densities. However, based on philosophical arguments given later this still suffices for practice. Birnbaum explained the goal of his work in the introduction of his 1962 paper as follows:

> "This paper treats a traditional and basic problem-area of statistical theory, which we shall call *informative inference*, which has been a source of continuing interest and disagreement. The subject-matter of interest here may be called *experimental evidence*: when an experimental situation is represented by an adequate mathematical statistical model, denoted by E, and when any specified outcome x of E has been observed, then (E, x) is an instance of *statistical evidence*, that is, a mathematical model of an instance of experimental evidence. Part of the specification of E is a description of the range of unknown parameter values or of statistical hypotheses under consideration, that is, the description of a parameter space Ω of parameter points θ. The remaining part of E is given by a description of the sample space of possible outcomes x of E, and of their respective probabilities of densities under respective hypotheses, typically by use of a specified probability density function $f(x, \theta)$ for each θ."
> Birnbaum (1962, p. 269-270)

Birnbaum then separated two problems: First, the task to find an appropriate mathematical characterization of statistical evidence as such, and second, the problem of evidential interpretation. The former aims solely at characterizing statistical evidence without providing any guidance how to interpret the resulting mathematical characterization, while the latter is only concerned with determining concepts and terms appropriate to interpret statistical evidence. Birnbaum set out for the first problem by introducing the symbol $Ev(E,x)$ for the evidential meaning of an instance (E,x) of statistical evidence:

> "that is, $Ev(E,x)$ stands for the essential properties (which remain to be clarified) of the statistical evidence, as such, provided by the observed outcome x of the observed specified experiment E."
> Birnbaum (1962, p. 270)

Birnbaum's main idea to mathematically characterize statistical evidence consisted in finding conditions under which one would assert that two instances (E,x) and (E',y) of statistical evidence are equivalent and he denoted such an assertion of evidential equivalence as $Ev(E,x) = Ev(E',y)$. To investigate the necessary conditions to make such an assertion, he introduced the principle of sufficiency first:

> "A first condition for such equivalence, which is proposed as an *axiom*, is related to the concept of sufficient statistic which plays a basic technical role in each approach to statistical theory. This is:
> *The principle of sufficiency* (S): If E is a specified experiment, which outcomes x; if $t = t(x)$ is any sufficient statistic; and if E' is the experiment, derived from E, in which any outcome of x of E is represented only by the corresponding value $t = t(x)$ of the sufficient statistic; then for each x, $Ev(E,x) = Ev(E',t)$ where $t = t(x)$."
> Birnbaum (1962, p. 270)

A translation into modern notation is postponed until the next section, and for now it suffices to note that a sufficient statistic captures all information in the data without any loss. Thus, the evidence $Ev(E,x)$ provided by the original data x in E is equivalent to the evidence $Ev(E',t)$ in the experiment E' where the original data x have been replaced by the sufficient statistic t and only t is reported. The second statistical principle that Birnbaum proposed as an axiom was the conditionality principle:

> "A second condition for equivalence of evidential meaning is related to concepts of conditional experimental frames of reference; such concepts have been suggested as appropriate for purposes of informative inference by writers of several theoretical standpoints, including Fisher and D.R. Cox. (...) The second proposed axiom, which many statisticians are inclined to accept for purposes of informative inference, is:
> *The principle of conditionality* (C): If E is any experiment having the form of a mixture of component experiments E_h, then for each outcome (E_h, x_h) of E we have $Ev(E,(E_h,x_h)) = Ev(E_h,x_h)$. That is, the evidential meaning of any outcome of any mixture experiment is the same as that of the corresponding outcome of the corresponding component experiment, ignoring the over-all structure of the mixture experiment."
> Birnbaum (1962, p. 271)

Clearly, as shown in Part I, Fisher was a fervent proponent of conditional inference, and Cox also argued for hypothesis testing conditional on the sub-experiment actually performed when considering a mixture experiment. Fisher made this notion explicit by arguing that conditioning on an ancillary statistic (like the margin totals in the 2 × 2-contingency tables discussed in Chapter 3) is necessary for correctly performing statistical inference. Also, this notion even seeped into his concept of probability through his idea of relevant subsets which the statistician needs to condition on.

The last statistical principle included in Birnbaum's paper was the likelihood principle, which "has been proposed and supported as self-evident principally by Fisher and G.A. Barnard, but which has not hitherto been very generally accepted." (Birnbaum, 1962, p. 271). Birnbaum therefore introduced what it means when two likelihood functions are the same, and then stated the likelihood principle:

> "This condition concerns the likelihood function, that is, the function of θ, $f(x,\theta)$, determined by an observed outcome x of a specified experiment E; two likelihood functions $f(x,\theta)$ and $g(y,\theta)$ are called the same if they are proportional, that is if there exists a positive constant c such that $f(x,\theta) = cg(y,\theta)$ for all θ. This condition is:
> *The likelihood principle* (LP): If E and E' are any two experiments with the same parameter space, represented by density functions $f(x,\theta)$ and $g(y,\theta)$; and if x and y are any respective outcomes determining the same likelihood function; then $Ev(E,x) = Ev(E',y)$. That is, the evidential meaning of any outcome x of any experiment E is characterized fully by giving the likelihood function $cf(x,\theta)$ (which need to be described only up to an arbitrary positive constant theory), without other reference to the structure of E."
> Birnbaum (1962, p. 271)

While the likelihood principle has a less immediate natural justification, Birnbaum translated it informally as the "irrelevance of outcomes not actually observed" (Birnbaum, 1962, p. 271). Birnbaum's achievement was to show based on these three simple principles, that (S) and (C) together are equivalent to (L). This proof provided a solution to a mathematical characterization of statistical evidence, because it showed that the likelihood function is the mathematical object to quantify evidence provided through data and the likelihood principle gives the conditions under which evidential equivalence is established in two experiments:

> "The fact that relatively few statisticians have accepted (L) as appropriate purposes of informative inference, while many are inclined to accept (S) and (C), lends interest and significance to the result, proved herein, that (S) *and* (C) *together are mathematically equivalent to* (L). When (S) and (C) are adopted, their consequence (L) constitutes a significant solution to the first problem of informative inference, namely that a mathematical characterization of statistical evidence as such is given by the likelihood function."
> (Birnbaum, 1962, p. 271)

Frequentist statistics was built on the shoulders of concepts like sufficiency and conditional inference as shown in Part I and thus (S) and (C) were readily accepted by most frequentist statisticians. However, there was little reason for frequentists to accept (L) from an axiomatic perspective. What is today called Birnbaum's theorem, was named by himself only Lemma 2 and was stated as follows:

> "Lemma 2. (L) implies, and is implied by, (S) and (C)."
> Birnbaum (1962, p. 284)

11.3.2 A simple proof of Birnbaum's theorem

The proof is elementary and went as follows (using Birnbaum's original notation): He denoted E and E' as two mathematical models of experiments with common parameter space Ω and probability density functions $f(x,\theta)$ and $g(y,\theta)$ on their samples spaces S and S' which are regarded to be distinct, disjoint spaces. The hypothetical mixture experiment E^* is considered whose components are E and E' with equal probabilities $\frac{1}{2}$. Birnbaum denoted z as a generic sample point of E^* and C as the set of points z so that $C = A \cup B$ with $A \subset S$ and $B \subset S'$. Then, the probability that Z is in C given θ is given as

$$\begin{aligned}\text{Prob}(Z \in C|\theta) &= \frac{1}{2}\text{Prob}(A|\theta, E) + \frac{1}{2}\text{Prob}(B|\theta, E') \\ &= \frac{1}{2}\int_A f(x,\theta)d\mu(x) + \frac{1}{2}\int_B g(y,\theta)d\nu(y)\end{aligned}$$

where A and B are measurable sets (and in modern notation, $f(x,\theta)$ and $g(y,\theta)$ are the $\mu-$ and ν-densities of the measures $\mathbb{P}_X$ and $\mathbb{P}_Y$ which operate on S and S'). The probability density function of E^* therefore can be written as

$$h(z,\theta) = \begin{cases}\frac{1}{2}f(x,\theta), \text{ if } z = x \in S \\ \frac{1}{2}g(y,\theta), \text{ if } z = y \in S'\end{cases} \tag{11.1}$$

Each outcome z of the mixture experiment E^* has a representation

$$z = \begin{cases}(E,x), \text{ if } z = x \in S \\ (E',y), \text{ if } z = y \in S'\end{cases}$$

Now, to show that (C) and (S) together imply (L), Birnbaum first used (C) and it follows that

$$Ev(E^*,(E,x)) = Ev(E,x) \text{ for each } x \in S \text{ and} \tag{11.2}$$
$$Ev(E^*,(E',y)) = Ev(E',y) \text{ for each } y \in S' \tag{11.3}$$

Suppose x' and y' are two outcomes of E and E' respectively which determine the same likelihood function, that is

$$f(x',\theta) = cg(y',\theta) \tag{11.4}$$

for all θ, where c is some positive constant. Then, $h(x',\theta) = ch(y',\theta)$ for all θ, too. Thus, the two outcomes (E,x') and (E',y') determine the same likelihood function.

Now, if two outcomes x, x' of *one* experiment E determine the same likelihood function (that is, if for some positive c one has $f(x,\theta) = cf(x',\theta)$ for all θ), then there exists a (minimal) sufficient statistic t such that $t(x) = t(x')$.[3] Thus, if two outcomes x, x' of any experiment E determine the same likelihood function, then they have the same

[3] This follows from the Neyman-Fisher factorization theorem, compare Theorem C.51: By assumption, $f(x,\theta) = cf(x',\theta)$ for all θ for $c > 0$, which can be expressed in more familiar notation as $f(x|\theta) = cf(x'|\theta)$. The Neyman-Fisher factorizations $f(x|\theta) = g(T(x)|\theta)h(x)$ and $f(x'|\theta) = g(T(x')|\theta)h(x')$ together with $\frac{f(x|\theta)}{f(x'|\theta)} = c$ imply that $\frac{g(T(x)|\theta)h(x)}{g(T(x')|\theta)h(x')} = c$, which is equivalent to $\frac{g(T(x)|\theta)}{g(T(x')|\theta)} = \frac{h(x')}{h(x)}c$. The right-hand side is constant for all θ and given x, x', and this implies that the left hand-side also is constant for all θ and given x, x'. The left-hand side, however, is only constant for all θ and given x, x' if and only if $T(x) = T(x')$ holds. Furthermore, the existence of such a (minimal) sufficient statistic is always guaranteed as long as the statistical model $\mathcal{P}$ is separable with regard to the total variation norm, compare Rüschendorf (2014, Theorem 4.2.9), so the assumption is very weak and non-restrictive.

evidential meaning $Ev(E,x) = Ev(E,x')$, because from (S) it follows that $Ev(E,x) = Ev(E',t(x)) = Ev(E',t(x')) = Ev(E,x')$.[4]

As the two outcomes (E,x') and (E',y') determine the same likelihood function, from (S) and the above it follows that

$$Ev(E^*,(E,x')) = Ev(E^*,(E',y')) \tag{11.5}$$

Using Equation (11.2), Equation (11.3) and Equation (11.5) it now follows that

$$Ev(E,x') \overset{(11.2)}{=} Ev(E^*,(E,x')) \overset{(11.5)}{=} Ev(E^*,(E,y')) \overset{(11.3)}{=} Ev(E',y') \tag{11.6}$$

for any two outcomes $x' \in S, y' \in S'$ of any two experiments E, E' with the same parameter space Ω. But Equation (11.6) states that based on the assumption Equation (11.4), (S) and (C), the evidential meaning of x' in E and y' in E' is identical, that is, (L) holds.

To show that (L) implies (C) Birnbaum noted that this follows immediately from the fact that the likelihoods in the mixture experiment $(E^*,(E_h,x_h))$ and the mixture component (E_h,x_h) are proportional (which is immediate from the expression of $h(z,\theta)$ above in Equation (11.1), and the proportionality constant is $c = \frac{1}{2}$). To show that (L) implies (S) Birnbaum supposed t to be sufficient in $E := (\Omega, S, f)$ and considered the experiment $E' = (\Omega, S', f')$ with transformation $t := t(x)$, $S' := t(S)$ and density

$$f'(t,\theta) = \sum_{x \in S : t(x) = t} f(x,\theta)$$

From the earlier considerations, a statistic $t(x)$ is sufficient in $E = (\Omega, S, f)$ only if $t(x) = t(x')$ implies that for some $c > 0$, $f(x,\theta) = cf(x',\theta)$ for all θ.[5] As the summands in $f'(t,\theta)$ all fulfill the condition that they are mapped to the same sufficient statistic value t, the likelihoods of them are proportional to each other, and proportional to $f(x,\theta)$. As a consequence, $f'(t,\theta)$ has the form $f'(t,\theta) = cf(x,\theta)$, where $t = t(x)$, for some $c > 0$. This implies that by assumption of the likelihood principle, that $Ev(E,x) = Ev(E',t)$ where $t = t(x)$, which is the statement of (S).[6]

The implication of Birnbaum's theorem that (S) and (C) together are equivalent to (L) were profound. Traditional frequentist statisticians widely accepted (S) and (C), but did not adhere to (L), because both Fisher's significance tests and the Neyman-Pearson theory explicitly violate (L), as will be shown below. In contrast, (L) implies

[4] Birnbaum formulated this as Lemma 1 in his paper, which follows from (S).

[5] Again, this follows from the Neyman-Fisher factorizations $f(x|\theta) = g(T(x)|\theta)h(x)$ and $f(x'|\theta) = g(T(x')|\theta)h(x')$, the ratio of which is equal to $\frac{f(x|\theta)}{f(x'|\theta)} = \frac{g(T(x)|\theta)h(x)}{g(T(x')|\theta)h(x')}$. As the sufficient statistics $t(x) = t(x')$ (here denoted as $T(x)$ and $T(x')$) are equal, it follows that $g(T(x)|\theta) = g(T(x')|\theta)$ and thus $\frac{f(x|\theta)}{f(x'|\theta)} = \frac{h(x)}{h(x')}$. As the right-hand side $\frac{h(x)}{h(x')}$ is constant as a function of θ for given x, x', it follows from $t(x) = t(x')$ (here denoted as $T(x) = T(x')$) that the left-hand side is also constant as a function of θ for given x, x'. Thus, $f(x|\theta) = cf(x'|\theta)$ for some $c > 0$. The fact that c is strictly positive follows from the Neyman-Fisher factorization theorem, because the function h needs to be strictly positive, compare Rüschendorf (2014, Theorem 4.1.15) and Theorem C.52.

[6] Birnbaum's original argument was much shorter in his Birnbaum (1962) paper, and consisted primarily of stating that the implication that (S) follows from (L) is due to Lemma 1 in his paper. The arguments presented above were given by Birnbaum (1972) as a separate Theorem later for clarification (Birnbaum, 1972, Theorem 2). However, as the much more important implication in the 1962 paper was that (S) and (C) imply (L), this lack of clarity was of little importance regarding the practical implications of Birnbaum's theorem.

that no data other than the observed can be used in statistical inference, and rejecting (L). Birnbaum's theorem implied that it was only possible to reject (L) when simultaneously rejecting either (S) or (C). However, both (S) and (C) were essential concepts of frequentist statistical inference was detailed in Part I, and thus rejection of either (S) or (C) questioned the core concepts of frequentism in statistical inference.

11.3.3 The reception of Birnbaum's theorem

After the original introduction by Birnbaum, his theorem was widely discussed and the result came as a shock to many frequentist statisticians. L.J. Savage noted in the discussion of Birnbaum's paper:

> "Without any intent to speak with exaggeration or rhetorically, it seems to me that this is really a historic occasion. This paper is a landmark in statistics because it seems to me improbable that many people will be able to read this paper or to have heard it tonight without coming away with considerable respect for the likelihood principle.
> I, myself, like other Bayesian statisticians, have been convinced of the truth of the likelihood principle for a long time. Its consequences for statistics are very great. A person who after an experiment like those discussed by Birnbaum proposes to use an analysis which is not in conformity with the principle, it seems to me, will have to think quite hard of his excuses for doing so."
> Savage et al. (1962b, p. 307)

Other reactions included that although the result was preposterous, earlier writers like Fisher, Neyman or Pearson were well acquainted with all of this, but noticed that reporting a sole likelihood function or conducting a Bayesian analysis was no solution. Irwin Bross fervently criticized Birnbaum's result and tried to defend frequentist statistics:

> "Finally, I would like to point out that the basic themes of this paper were well-known to Fisher, Neyman, Egon Pearson and others, well back in the 1920's. But these men realized, as the author doesn't, that the concepts cannot be used directly for scientific reporting. So, they went on to develop confidence intervals in the 1930's, and these proved to be very useful. The author here proposes to push the clock back 45 years, but at least this puts him ahead of the Bayesians, who would like to turn the clock back 150 years."
> Irwin Bross in Savage et al. (1962b, p. 310)

However, for the majority of discussants Birnbaum's theorem came as a shock. Jerome Cornfield noted in his comment that

> "I haven't quite recovered from the shock of seeing that two principles I had thought reasonable and one which I had thought doubtful imply each other. It is clear that I must either believe all three of disbelieve at least one of the two reasonable ones."
> Jerome Cornfield in Savage et al. (1962b, p. 309)

George E.P. Box added:

> "I believe, for instance, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem."
> George E.P. Box in Savage et al. (1962b, p. 311)

With regard to statistical hypothesis testing, one of the most important comments was made by A.P. Dempster in the discussion of Birnbaum's paper, who stressed:

> "(L) implies the exclusion from any role in evidential meaning of significance tests, confidence statements, and even so basic a concept as the mean square error of an estimator. To eradicate such concepts from the thought process of statisticians would require d prodigious brain-washing program."
> A.P. Dempster in Savage et al. (1962b, p. 318)

11.3.4 Refinement of Birnbaum's theorem

Birnbaum's theorem presented a shock for frequentist statisticians who were routinely using Fisher's significance tests, Neyman-Pearson tests, confidence intervals and, in general, methods which uses data that was not actually observed in an experiment. Acceptance of the sufficiency and conditionality principle built the foundation of the frequentist mode of statistical inference, and thus the equivalence to the likelihood principle presented a serious challenge to proceed with the current practice of statistical inference. However, in 1962, the only alternative was to use Bayesian methods which were difficult to use because the lack of MCMC theory and computing resources, and pure likelihood inference was of little use as stressed in the comment of Irwin Bross above.

Birnbaum's result was later refined and extended by Berger and Wolpert (1988) to the continuous case. In this section, Birnbaum's result is translated into modern notation and more precise definitions are provided. Illustrating examples also clarify why (S) and (C) should be accepted as axioms. In this section, the underlying joint probability space Ω[7] is therefore assumed to be a discrete space like in Birnbaum's paper, and an experiment E is now defined as follows:

Definition 11.2 (Experiment (Berger and Wolpert, 1988)). A statistical experiment E is a triple $(X, \theta, \{f(\cdot|\theta\})$, where X is a random vector on Ω with probability mass function $f(\cdot|\theta)$ for $\theta \in \Theta$.

The experiment is thus modeled by a family of probability densities for a random variable X according to which the observed data x is assumed to be generated. However, the probability densities are parameterized by the parameter(s) θ, which remain unknown. The observed data x provide information about the unknown parameter θ and are used to estimate it, or test a hypothesis about θ. For more details see Appendix C. Birnbaum (1962) denoted the inference or conclusion about θ the evidence $Ev(E, x)$ about θ which arises from E and x, based on the observed data $X = x$ and the experiment E. He presupposed nothing specific about what this evidence might be.[8] Thus, his definition was:

[7] See Appendix C for a measure-theoretic perspective: The space Ω in this section can be interpreted as the joint probability space $\Omega := \mathcal{X} \times \Theta$ as given in Appendix C.

[8] Birnbaums understanding of mathematical and statistical evidence has been debated in the literature, for details see Giere (1977).

Definition 11.3 (Evidence (Berger and Wolpert, 1988))**.** When experiment E was performed and $X = x$ is observed

$$Ev(E, x) \tag{11.7}$$

denotes the evidence about θ arising from E and x.

Evidence in the sense of Birnbaum (1962) is therefore quite abstract. While 'evidence' could be associated with traditional statistical measures of evidence like p-values, significance levels or Bayes factors, this is purposely not done in Birnbaums derivations. Thus, Birnbaum (1962) stayed as general as possible. As outlined above, Birnbaum started with the *Conditionality Principle* (*CP*) (henceforth abbreviated as (CP) instead of (C) as in Birnbaum's original notation), which essentially states that if an experiment is selected by a random mechanism (independent of θ) out of many experiments, only the experiment actually performed is relevant for the evidence obtained. While Birnbaum (1962) used the CP in his derivations, Basu (1975) showed that the even *Weak Conditionality Principle* (*WCP*) suffices for Birnbaum's derivations.[9] The WCP is weaker than the CP because it does not allow for an arbitrary number of mixture experiment components, but exactly two components with equal mixing probabilities:

Weak Conditionality Principle (WCP, (Berger and Wolpert, 1988)). *Suppose that $E_1 = (X_1, \theta, \{f_1(\cdot|\theta)\})$ and $E_2 = (X_2, \theta, \{f_2(\cdot|\theta)\})$ are two experiments, where only the unknown parameter θ needs to be common between the two experiments. Consider the mixed experiment in which the random variable J is observed, where $P(J = 1) = P(J = 2) = \frac{1}{2}$ (independent of θ, X_1 or X_2), and then experiment E_J is performed. Formally, the experiment performed is $E^* = (X^*, \theta, \{f^*(\cdot|\theta)\})$, where $X^* = (j, X_j)$ and $f^*(x^*|\theta) = f^*((j, x_j)|\theta) = \frac{1}{2}f_1(x_1|\theta) + \frac{1}{2}f_2(x_2|\theta)$. Then*

$$Ev(E^*, (j, x_j)) = Ev(E_j, x_j) \tag{11.8}$$

One may ask why the WCP should be accepted from an axiomatic perspective. There are indeed many examples which show that rejecting the WCP would be completely unreasonable and even contradict common sense. Here, an example given by (Berger and Wolpert, 1988, p. 6) is repeated for clarification:

Example 11.4 (Berger and Wolpert (1988))**.** Suppose a substance (e.g. a blood sample of a patient) needs to be analyzed and can be sent either to a laboratory in New York or California. Both labs are equally good, so a fair coin is flipped to decide between them, where heads denotes the lab in New York will be chosen and tails denotes the lab in California will be chosen. The coin is flipped and comes up tails, so the California lab is chosen. Finally, the results from the lab arrive and a conclusion needs to be reached about the sample. Should the conclusion take into account the fact that the coin could have been heads, and therefore that the experiment in New York might have been performed instead?

Of course, common sense allows only to incorporate the information from the experiment actually performed. This implies to use only the information from the lab the sample was actually sent to. Note, that this example of Berger and Wolpert (1988) is a copy of the famous example of Cox (1958) as discussed in Part I, Chapter 3. Cox argued that a mixed experiment in which a fair coin is flipped and either a test for level

[9] See also Berger and Wolpert (1988, p. 25) and Casella and Berger (2002, p. 293)

$\alpha = 0$ or $\alpha = .10$ is performed, leads to a total test level of $\alpha = .05$ (see Section 3.2.3). Also, according to Cox (1958), in this case, the conditional test which bases inference only on the experiment actually conducted should be used. This test either has the test level $\alpha = 0$ or $\alpha = 0.10$ instead of $\alpha = 0.05$. While the unconditional test will attain the level $\alpha = 0.05$ under repetition eventually, for the actual inference at hand either $\alpha = 0$ or $\alpha = 0.10$ holds, which is why the original argument of Cox (1958) is a strong argument for *conditional* inference instead of unconditional inference. The WCP formalises the example of Cox (1958) in a certain sense and states that evidence should only depend on the experiment conducted. Based on this reasoning, Cox (1958) argued that experiments which were not realized should be irrelevant to the inference obtained.[10] The WCP, therefore, can be attributed to Cox (1958) and is the first cornerstone in the development of Birnbaum (1962). The second cornerstone of Birnbaum's development as outlined above was the sufficiency principle. Next to the conditionality principle, also the sufficiency principle can be weakened and the proof of Birnbaum's theorem still holds. This weaker version of the sufficiency principle was called the *Weak Sufficiency Principle* (*WSP*), and the name differs between authors: While Casella and Berger (2002) call it the *Formal Sufficiency Principle*, Berger and Wolpert (1988) denote it the *Weak Sufficiency Principle*, which is also used here. Held and Sabanés Bové (2014, p. 47) denote it simply the Sufficiency Principle. The original notation of *Weak Sufficiency Principle* is attributed to Dawid (1977), and before stating the WSP it is useful to consider the definition of a sufficient statistic T, see Definition C.50:

Definition 11.5 (Sufficiency)**.** A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Thus, a statistic $T(X_{1:n})$ is sufficient for θ if the conditional distribution of $X_{1:n}$ given $T = t$ is independent of θ, that is, if

$$f(X_{1:n}|T = t) \tag{11.9}$$

does not depend on θ, compare Held and Sabanés Bové (2014). Here, $X_{1:n}$ denotes the sample $(X_1, ..., X_n)$ of size n. Thus, when the statistic T has been observed for the sample X, the distribution of the data conditional on the value of this statistic does not depend on the unknown parameter θ anymore, because all information about θ is already contained in T which has been observed. The WSP makes the following statement:

Weak Sufficiency Principle (WSP, (Berger and Wolpert, 1988)). *Consider the experiment $E = (X, \theta, \{f(x|\theta)\})$ and let $T(X)$ a sufficient statistic for θ. If x and y are sample points satisfying $T(x) = T(y)$, then $Ev(E, x) = Ev(E, y)$.*

The intuitive motivation for the trustworthiness of the WSP is the fact that a sufficient statistic T captures *all* information in the data without any loss, compare Definition C.50.[11] The main reason for accepting the WSP, therefore, is that a statistic which

[10] However, Example Example 11.4 differs from the example of Cox in the detail that the hypothesis tests considered by Cox are not 'equally good', when equally good is interpreted as the specified test level α. The labs in Example 11.4 are equally good, but one could easily omit this detail and the conclusion would stay identical: Any inference can only be based on the data observed and the experiment actually performed, even when the quality of both labs is different.

[11] Note that Fisher had a talent to give his concepts very appealing and useful names. What better name could one give a statistic, which captures all information in the data, than sufficient?

summarizes all relevant information in the observed data so that it only compresses the information but does not lose any of it suffices to draw inferences. However, this motivation serves primarily for accepting the original sufficiency principle as used by Birnbaum. The WSP, however, follows immediately from the original sufficiency principle, because from the sufficiency principle in Birnbaum's version, it follows that $Ev(E,x) = Ev(E,T(x))$ when T is a sufficient statistic in the experiment E for the parameter of interest. As a consequence, one obtains for two sample points x,y with $T(x) = T(y)$ from the sufficiency principle that $Ev(E,x) = Ev(E,T(x)) = Ev(E,T(y)) = Ev(E,y)$. Thus, whenever $T(x) = T(y)$ for two sample points x,y it follows that $Ev(E,x) = Ev(E,y)$, which is precisely the statement of the WSP, and the WSP follows from the original sufficiency principle. According to the WSP, if in an experiment the sufficient statistic T yields the same value for two hypothetical sample realizations x and y, the statistical evidence $Ev(E,x)$ and $Ev(E,x)$ provided by x and y is the same. The following example shows why the WSP is reasonable as an axiom:

Example 11.6. Suppose a person flips a single coin twenty times. Consider two hypothetical realizations A and B of the experiment:

$$A : 10110110001000010010$$
$$B : 01000010010010101101$$

Here, success is defined as 'coin comes up heads' or 1, and failure as 'coin comes up tails' or 0. In A, the twenty flips yield 8 successes in the first 8 flips and 12 failures in the last 12 flips. In B, the order is reversed. Assume the situation is modelled as the observed data X being binomially distributed with $n = 20$ and unknown success probability $\theta \in [0,1]$. The value $X_i = 1$ indicates a success and $X_i = 0$ indicates no success. It is well known that the number of successes $T(X_1,...,X_{20}) := \sum_{i=1}^{20} X_i$ is a sufficient statistic for the unknown parameter θ of the coin yielding a success.[12] Now, should the inference drawn about the parameter θ be different in A and B?

As $T((X_1,...,X_{20})) := \sum_{i=1}^{20} X_i = 8$ is a sufficient statistic for the unknown parameter θ, the statistical evidence about θ of course should be the same in both hypothetical realizations A and B. In both cases 8 successes are observed, and the only difference is the ordering when these are observed. The evidence obtained for the success rate parameter θ is the same and θ would be estimated as $8/20$ in both A and B. It would contradict common sense to infer a different conclusion about θ in both possible realizations. The WSP formalizes this intuition.

Now, Berger and Wolpert (1988) refined Birnbaum's original proof by weakening the original sufficiency and conditionality principles to the WCP and WSP. They showed that then the *Likelihood Principle* (*LP*) still follows, and (WSP) and (WCP) together are equivalent to (LP).

Formal Likelihood Principle (LP, (Berger and Wolpert, 1988)). *Suppose that we have two experiments, $E_1 = (X_1, \theta, \{f_1(x_1|\theta)\})$ and $E_2 = (X_2, \theta, \{f_2(x_2|\theta)\})$ where the unknown*

[12] This is shown by calculating $f_\theta(X|\sum_{i=1}^{20} X_i = t)$, which is the conditional density of the sample data X given $T(X) = t$. From the definition of conditional probability, it follows that

$$f_\theta(X|\sum_{i=1}^{20} X_i = t) = \frac{\theta^t(1-\theta)^{20-t}}{\binom{20}{t}\theta^t(1-\theta)^{20-t}} = \frac{1}{\binom{n}{t}}$$

Thus, the conditional density $f_\theta(X|T(X) = t)$ is independent of θ and it follows that T is sufficient for θ.

parameter θ is the same in both experiments. Suppose x_1^* and x_2^* are sample points from E_1 and E_2, respectively, such that

$$L(\theta : x_2^*) = C \cdot L(\theta : x_1^*) \tag{11.10}$$

for all θ and for some constant C that may depend on x_1^* and x_2^* but not θ. Then

$$Ev(E_1, x_1^*) = Ev(E_2, x_2^*) \tag{11.11}$$

In its simplest case $C(x,y) = 1$, the LP states that if two sample points result in the same value of the likelihood function for all parameter values θ, the inference made should be identical for both points. This special case of course can hardly be rejected, but of more interest is the general case in which $C \neq 1$. The more farreaching consequences of accepting the LP become clear when considering this general case, which is highlighted in the following classic example, taken from (Berger and Wolpert, 1988, Chapter 3):

Example 11.7. In this example, two experiments are compared where the only difference between them is the rule when to stop experimenting. Therefore, let $Y_1, Y_2, ...$ be independent and identically distributed (iid) Bernoulli random variables with success parameter θ. In experiment E_1, we assume that a fixed sample size of 12 observations is taken (the experiment is stopped after $n = 12$ observations), and suppose the sufficient statistic $T_1(y_1, ..., y_{12}) = \sum_{i=1}^{12} y_i$ to be 9. Experiment E_2 proceeds by taking indefinitely many observations until a total amount of 3 zeros has been observed and is then stopped. We assume that by coincidence, $T_2(y_1, ..., y_{12}) = \sum y_i$ again turns out to be 9. The distribution of T_1 in E_1 is binomial with density

$$t_1 \mapsto f_1(t_1|\theta) = \binom{12}{t_1} \theta^{t_1} (1-\theta)^{12-t_1} \tag{11.12}$$

which leads for $t_1 = 9$ to the likelihood function

$$\theta \mapsto L_1(9;\theta) = \binom{12}{9} \theta^9 (1-\theta)^3 \tag{11.13}$$

The distribution of T_2 in E_2 is negative binomial with density

$$t_2 \mapsto f_2(t_2|\theta) = \binom{t_2+2}{t_2} \theta^{t_2} (1-\theta)^3 \tag{11.14}$$

which leads for $t_2 = 9$ to the likelihood function

$$\theta \mapsto L_2(9;\theta) = \binom{11}{9} \theta^9 (1-\theta)^3 \tag{11.15}$$

First, the LP states that for experiment E_i all information about θ is contained in $L_i(9 : \theta)$ alone. The more striking consequence here is that as $L_1(9 : \theta)$ and $L_2(9 : \theta)$ are proportional as functions of θ, $L_1(9 : \theta) = C \cdot L_2(9 : \theta)$ with proportionality constant $C = \binom{12}{9} / \binom{11}{9}$, the information about θ in E_1 and E_2 is identical. According to the LP, this implies that it does not matter if the fixed sample size 12 is chosen pre-experimental, or if the experiment is conducted until enough successes have been observed. The evidence in both cases must be identical. While it seems reasonable to base any inference only on the observed likelihood function and not on the intentions of the researchers or design of the experiment, accepting or rejecting the likelihood principle can lead to entirely different conclusions in the above example.

Now, consider a frequentist who conducts a significance test according to Fisher or Neyman-Pearson. The p-value or test statistic is of course different for both experiments, because in E_1 the p-value or test statistic is calculated based on the *binomial distribution*, while in E_2, it is based on the *negative binomial distribution*. It now can happen that the test in E_1 based on the *binomial distribution* will become significant while the test in E_2 based on the *negative binomial distribution* does not. Then, *different* conclusions are drawn for E_1 and E_2, although the likelihood functions are proportional. Therefore, frequentist significance tests in the sense of Fisher or Neyman and Pearson violate the LP. In the above example this is shown as follows: A p-value for the null hypothesis of a fair coin $H_0 : \theta = 0.5$ versus $H_1 : \theta < 0.5$ is calculated as the probability of obtaining a result equal to or more extreme than the one observed under assumption of $H_0 : \theta = 0.5$. We define the observed zeros as the result of interest here. In the binomial experiment, three zeros were observed in twelve iterations. The probability of three or fewer zeros (we compare $H_0 : \theta = 0.5$ against $H_1 : \theta < 0.5$) under the assumption of $H_0 : \theta = 0.5$ is

$$\mathbb{P}(t_1 \geq 9|H_0) = \left(\binom{12}{9} + \binom{12}{10} + \binom{12}{11} + \binom{12}{12}\right)\left(\frac{1}{2}\right)^{12} \approx 0.073$$

This is the type I error probability. Using a significance threshold of $\alpha = 0.05$, the null hypothesis $H_0 : \theta = 0.5$ cannot be rejected. In contrast, for the negative binomial experiment, the type I error probability is the probability of needing to conduct twelve or more experiments to obtain three zeros. The probability density of the negative binomial density is given by $f(k) = \binom{k+r-1}{k} \cdot p^r \cdot (1-p)^k$, where $k \in \{0, 1, 2, 3, ...\}$ is the number of failures, $p \in (0, 1)$ the single-toss success probability and $r > 0$ the number of successes until sampling stops. In the above example, a success is a zero, so sampling is done until $r = 3$ zeros are obtained. The type I error probability therefore is $\mathbb{P}(k + r \geq 12|H_0, r = 3)$, which equals the probability $\mathbb{P}(k \geq 9|H_0)$ of obtaining nine or more failures (twelve or more experiments are conducted until three zeros are obtained):

$$\begin{aligned}
&\mathbb{P}(k+r \geq 12|H_0, r=3) = \mathbb{P}(k \geq 9|H_0) = 1 - \mathbb{P}(0 \leq k \leq 8|H_0) \\
&= 1 - [\binom{8+3-1}{8}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^8 + \binom{7+3-1}{7}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 + ... + \binom{0+3-1}{1}\left(\frac{1}{2}\right)^3] \\
&= 1 - [\binom{10}{8}\left(\frac{1}{2}\right)^{11} + \binom{9}{7}\left(\frac{1}{2}\right)^{10} + \binom{8}{6}\left(\frac{1}{2}\right)^9 + \binom{7}{5}\left(\frac{1}{2}\right)^8 + \binom{6}{4}\left(\frac{1}{2}\right)^7 \\
&+ \binom{5}{3}\left(\frac{1}{2}\right)^6 + \binom{4}{2}\left(\frac{1}{2}\right)^5 + \binom{3}{1}\left(\frac{1}{2}\right)^4 + \binom{2}{0}\left(\frac{1}{2}\right)^3] \approx 0.0327
\end{aligned}$$

Using $\alpha = 0.05$, the null hypothesis $H_0 : \theta = 0.5$ is rejected this time. Therefore, although the likelihood functions of the binomial and negative binomial experiment are proportional to each other, a significance test rejects the null hypothesis $H_0 : \theta = 0.5$ in one experiment, while it does not in the other. The above example therefore highlights the following problematic fact:

Fact. Frequentist significance tests in the interpretation of Fisher and frequentist hypothesis tests in the interpretation of Neyman and Pearson violate the likelihood principle.

As stressed above, the axiomatic basis of frequentist statistical hypothesis tests was given by the sufficiency and conditionality principles, and frequentists readily accepted these two cornerstones of statistical inference. The likelihood principle was, on the contrary, accepted by few statisticians as an axiom (even Fisher violated it via his significance tests although he advocated it, compare Part I). The violation of the LP by frequentist statistical hypothesis tests thus implied that the likelihood principle needed to be rejected to be able to perform hypothesis tests in the spirit of Fisher or Neyman and Pearson. However, due to the equivalency relationship between the LP and WSP and WCP, Birnbaum's theorem demonstrated that when rejecting the likelihood principle, either the WSP or the WCP or possibly even both need to be rejected. As the WSP and WCP presented the cornerstone of frequentist hypothesis testing, this presented a major challenge for frequentist statistics.

A direct corollary expresses the conflict between frequentist null hypothesis significance testing the LP even more radically:

Likelihood Principle Corollary (Berger and Wolpert (1988)). *If $E = (X, \theta, \{f(x|\theta)\})$ is an experiment, then $Ev(E, x)$ should depend on E and x only through $L(\theta : x)$.*

Clearly, null hypothesis significance tests in the spirit of Fisher or Neyman and Pearson are not based solely on the likelihood function.

Dealing with Example 11.4, as an alternative consider a Bayesian who faces the same situation: As the likelihood functions are proportional, and θ is the same parameter in both E_1 and E_2 (compare the assumption of the LP), a Bayesian can select only a single prior distribution $p(\theta)$ for θ. The posterior $p_1(\theta|9)$ in E_1 is then given as

$$\begin{aligned} p_1(\theta|9) &= \frac{L_1(9;\theta)p(\theta)}{f_1(x)} = \frac{L_1(9;\theta)p(\theta)}{\int_\Theta f_1(x|\theta)p(\theta)d\theta} \overset{(\Delta)}{=} \frac{\cancel{C} \cdot L_2(9;\theta)p(\theta)}{\cancel{C} \cdot \int_\Theta f_2(x|\theta)p(\theta)d\theta} \\ &= \frac{L_2(9;\theta)p(\theta)}{f_2(x)} = p_2(\theta|9) \end{aligned} \tag{11.16}$$

where $f_i(x) = \int_\Theta f_i(x|\theta)p(\theta)d\theta$ is the marginal likelihood under $E_i, i = 1, 2$ and $L_1(9;\theta) = C \cdot L_2(9;\theta)$ was used. Thus, the above shows that the posterior density $p_1(\theta|9)$ is identical to $p_2(\theta|9)$. As a consequence, the posterior distributions in E_1 and E_2 are equal (up to null sets). As for Bayesians, all inference follows from the posterior distribution, all subsequent steps like the computation of a Bayes factor will yield identical results in E_1 and E_2.

The assumption that θ is the same in both experiments is crucial here, as otherwise two priors $p_1(\theta)$ and $p_2(\theta)$ could reasonably be chosen by a Bayesian without violating the LP. Then, different conclusions could be drawn even though likelihoods are proportional even when opting for a Bayesian approach. This also manifests in Equation (11.16), because when θ is not the same in E_1 and E_2, the prior density $p(\theta)$ can be selected as $p_1(\theta)$ in E_1 and $p_2(\theta)$ in E_2, and the equality (Δ) in Equation (11.16) does not hold anymore.

A Bayesian analysis thus accords with the LP.[13]

[13] Note that this follows even when not specifying which Bayesian method is used: In the above, a

Berger and Wolpert (1988) then proved Birnbaum's fundamental theorem with weakened versions of the sufficiency and conditionality principle, the WSP and WCP:

Theorem 11.8 (**Birnbaum (1962)**). The Formal Likelihood Principle follows from the Weak Sufficiency Principle and the Weak Conditionality Principle. The converse is also true.

Proofs of Birnbaum's Theorem based only on the WSP and WCP and the Likelihood Principle Corollary can be found in (Casella and Berger, 2002) and (Berger and Wolpert, 1988, Chapter 3) and do only require basic probability theory, so they are not repeated here. However, from the outline of Birnbaum's original proof above it is immediate that Birnbaum actually used only the WCP (his mixture experiment consisted of two sub-experiments with equal mixing probabilities), and he made use of the WSP only via the Fisher-Neyman-factorization.

11.3.5 Arguments for and against Birnbaum's development

After the original publication of Birnbaum (1962), a variety of criticisms were offered, most of which did not stand the test of time. While the original derivations of Birnbaum (1962) lacked some clarity, these were quickly resolved afterwards and did not limit the validity of his theorem, see also Birnbaum (1972), Basu (1975), Joshi (1976) and Godambe (1979) for further details. Importantly, from a philosophic perspective, Basu (1975) argued that the discrete case handled by Birnbaum (1962) suffices for any practical purposes. The reason for this is that in reality, any sample space Ω is finite in any physically realisable experiment because one can only observe data with *finite* precision. Therefore, one may be completely satisfied with the discrete case. Still, there are various other ways to attack the LP, which principally consist of questioning the axiomatic assumption of the WSP and WCP.

There are also some criticisms which arise from the misapplication or misinterpretation of the LP, which are also detailed in Berger and Wolpert (1988) and which are less convincing as serious criticism against the LP:

1. *Criticism:* The LP applies only when θ includes all unknowns relevant to the problem, but in practice, there is a palette of techniques like latent-variable analysis, models including nuisance parameters, sequential analysis, and many more, where important unknowns often include more than just θ, the parameter in the probability model.
   Solution: The LP can be reformulated to include such unknowns, see (Berger and Wolpert, 1988, Section 3.5), especially the details about nuisance parameters, the *Marginalization Principle* (*MP*) and the *Noninformative Nuisance Parameter Principle* (*NNPP*). These additional principles guarantee that the LP also holds in models with nuisance parameter. Furthermore, the LP makes only a statement conditioned on two selected statistical models for the experiments E_1 and E_2. Thus, if the model does not parameterize important unknowns, this is no defect of the LP but a model choice problem which comes before any inference.

2. *Criticism:* In some cases like quality control or medical tests for a specific disease, long-run performance is the main target. Therefore a frequentist measure like a

hypothesis test based on the Bayes factor or the computation of the posterior median could be chosen. Other options would be to calculate an interval estimate. No matter which kind of analysis is selected, the results will be identical in E_1 and E_2, as all inference is based on the posteriors $p_1(\theta|x_1^*)$ and $p_2(\theta|x_2^*)$.

test level in the Neyman-Pearson theory is the appropriate object to proceed with. This thought is sometimes used as a counterexample to the LP.
Solution: Such situations cause no counterexample to the LP but are situations in which indeed the interest is not in evidence about the situation at hand in every single case, but about the long-run performance. Therefore, the LP does not apply in such situations, because the evidence is not of interest here. However, this criticism is considered in more detail later as it aims at the appropriate application context of the LP.

3. *Criticism:* There can be ambiguities in the definition of the likelihood function, especially in the continuous case detailed below.
Solution: These problems can be resolved via measure-theoretic considerations, especially by treating sets of measure zero similar to the identification of measures in $\mathcal{L}^p$ which differ only on Lebesgue null-sets by changing to the quotient space L^p; for details see (Berger and Wolpert, 1988, Section 3.4)

4. *Criticism:* There are periodical attempts to prove the LP wrong. Most of these are using likelihood-based methods which give bad results.
Solution: The LP does not state which method to use and also nothing about the efficiency of solutions obtained with any particular method. A bad result is no argument against the LP itself.

5. *Criticism:* The LP is not applicable to situations in which information is conveyed via different parameters from different experiments, for example two binomial-distributed experiments with parameter θ_1 and θ_2 in E_1 and E_2. The LP states that the conclusions reached need to be identical, but as θ_1 and θ_2 could measure entirely different things, this is a contradiction to the LP.
Solution: The LP only applies when θ_1 and θ_2 are the same parameter in both experiments E_1 and E_2 and are physically (or at least conceptually) the same quantity, compare **??** . If θ_1 measures the success rate of a flipped coin, and θ_2 the efficacy of a drug, observing 10 successes out of 15 may lead to entirely different conclusions about the coin and the efficacy of the drug, as the LP does not apply. Still, even in these cases the statistical evidence is obtained by the likelihood function. The *statistical* evidence obtained in both experiments could, therefore, argued to be the same, while the *scientific* evidence (what can be learned when incorporating other elements than only the statistical analyses) has not to be identical in the case the parameters θ_1 and θ_2 mean entirely different things.

Next to these more superficial misconceptions, the main criticisms against Birnbaum's proof can be structured roughly into four areas:

1. Criticisms targeting the model assumption

2. Criticisms targeting the evidence assumption

3. Criticisms targeting the WCP

4. Criticisms targeting the WSP

Criticisms targeting the model assumption

Often the LP is criticized because it assumes a particular parametric model with a density for X. This assumption does not always hold, for example, when considering nonparametric statistics. Still, even for situations in which no particular parametric model can be assumed, there do exist multiple models which are under consideration. The LP then still states that all information is in the data for any model under consideration, even if the information now cannot be associated with information about a parameter θ in a parametric model, but only with the information about which model is the most suitable one. Berger and Wolpert (1988) formalized this by letting θ represent various models, and in the situation when X is discrete and the sample space Ω is given as $\Omega := \{x_1, x_2, ...\}$, they denoted $\theta = (\theta_1, \theta_2, ...)$ as a point in the infinite-dimensional simplex $\Theta := \{\theta : 0 \leq \theta_i \leq 1, \sum \theta_i = 1\} \subseteq \mathbb{R}^{\mathbb{N}}$. Then, Berger and Wolpert (1988) defined $\mathbb{P}_\theta(x_i) = \theta_i$ on $\Omega = (x_1, x_2, ...)$, which is the class $\{\mathbb{P}_\theta\}$ of all probability distributions on Ω. This formalization suffices for a completely nonparametric setup in which no parametric model can be assumed for X.

The philosophical argument of finite-precision measurements detailed above indicates that the discrete case here suffices because even for continuous nonparametric settings the measurement precision will be finite, so a discrete sample space Ω suffices. When comparing an infinite number of models, a discrete approximation again suffices due to the inability to separate in practice between, for example $\mathcal{N}(0,1)$ and $\mathcal{N}(0, 1.00000000001)$ depending on the measurement precision. While it could be argued that there are cases in which one deals with a non-dominated family, so that there is no Radon-Nikodym derivative, there is the relative likelihood principle detailed in section 11.3.5 below which establishes the LP also in this situation. Also, in almost all realistic applications the statistical model $\mathcal{P}$ is dominated, see Kleijn (2022). Otherwise, the underlying metric space $(\mathcal{P}, d_r)$ is not separable with regard to the total variation norm d_r, compare Rüschendorf (2014, Theorem 3.1.17), which quickly narrows down the options to model the observations via a metric space.

Criticisms targeting the evidence assumption

Another branch of criticisms of the derivation of Birnbaum (1962) aims at his definition of evidence, which is questioned. Option one is to question the existence of evidence at all, which leads inevitably to philosophical discussions. However, the general notion is that the data, in some form, provide that evidence. Otherwise, the whole statistical enterprise should be stopped. The second option is to question the meaning and especially the uniqueness of $Ev(E, x)$, which is not defined precisely. Still, Birnbaum (1962) never argued that there needs to be a *single measure of evidence*. There could be multiple to proceed with, and therefore this argument does not limit the validity of the LP. Other authors have argued to replace evidence with inference patterns (Dawid, 1977), which makes it possible to include specific inference patterns like p-values, significance tests, Bayes factors or others in the set of patterns considered. In practice, however, this is rarely if ever done so statistical evidence as an abstract concept is still widely used (Giere, 1977).

Criticisms targeting the WCP

Criticisms about the WCP are more severe and show an essential reconnection to the historical reconstruction in Part I, especially Chapter 4. Frequentists can reject the weak conditionality principle if a position is taken in which one says the WCP is based on the mistaken belief that it is possible to obtain evidence about a particular parameter value θ from a particular experiment. Interestingly, Neyman (1957) took this perspective years after publishing his theory with Egon Pearson. His argument consisted of stating that it is only possible to guarantee the performance of a procedure in repeated use, and this should in some kind include averaging over both experiments E_1 and E_2 in the mixture experiment E^* defined in the WCP. Still, in all his applications, long-term performance or quality control were the main goals and this position is in sharp contrast to the scientist who obtains single results which cannot be reproduced in precisely the same fashion and for which it is necessary to draw conclusions in direct succession to the single experiment or study conducted. This situation is the routine in scientific practice, at least in medical and social sciences. One may argue that natural sciences like physics or chemistry yield experiments which can be repeated and indeed should deliver nearly the same results so that long-term performance could be a reasonable goal. However, in most of these cases a causal model exists and experiments serve only to investigate or confirm these causal relationships. Here, interest lies in evidence about a scientific theory or causal model after all, too. Therefore, Neyman's position is questionable from a perspective which focusses on the application of hypothesis testing in *scientific* contexts. This also answers the second criticism above about the appropriate context of the LP in routine applications like quality control: There, the primary interest is not statistical evidence, but long-term minimisation of a prespecified loss. In scientific contexts, however, the argument that successful replications with simultaneous minimization of type I errors are the primary goal, for example in experimental sciences like physics or chemistry, does not hold. From a decision-theoretic perspective, following such a behaviour is inferior regarding the incurred losses, as was first shown by Berger and Wolpert (1988). These details will be discussed in Section 11.4, where Neyman's position is identified with the *Confidence Principle*. As will be shown there, next to the decision-theoretic problems the confidence principle also lacks an axiomatic justification in contrast to the (relative) likelihood principle. The relative likelihood principle is the extension of the LP for continuous densities.

Criticisms questioning the WSP

The last category of criticism targets the weak sufficiency principle, and this is indeed the most serious criticism. The first issue is that if one faces a decision in which the consequences depend on the observed data x, and not just on the action taken and unknown parameter θ, the WSP needs not be valid.

Example 11.9 (Continuation of Example 11.6)**.** Reconsider Example 11.6 and assume a decision rule δ which always reports the estimate $\hat{\theta} = \frac{1}{2}$ whenever the last flip is a success, and otherwise the percentage of successes in the twenty coin flips. Thus, the decision rule $\delta : \mathcal{X} \to \Theta$ which maps from the sample space $\mathcal{X}$ to the parameter space

Θ now depends on the observed data and can be written as follows:

$$\delta(X_1, ..., X_{20}) := \begin{cases} \frac{1}{20}\sum_{i=1}^{20} X_i \text{ if } X_{20} = 0 \\ \frac{1}{2}, \text{ if } X_{20} = 1 \end{cases} \tag{11.17}$$

Based on δ we can easily construct two sequences as given in Example 11.6 which yield different decisions (or estimates) for the unknown success rate parameter θ: For sequence A we have $X_{20} = 0$ and thus we arrive at $\delta(X_1, ..., X_{20}) = \frac{8}{20}$, while for B we have $X_{20} = 1$ and we arrive at $\delta(X_1, ..., X_{20}) = \frac{1}{2}$. Thus, when we interpret the estimates provided by δ as the statistical evidence provided by the experiment E, we have $Ev(E, A) \neq Ev(E, B)$ for the sequences A and B, but still $T(A) = T(B)$. The WSP would require to assert evidential equivalence due to $T(A) = T(B)$, but we can violate the WSP.

The reason, however, is that the decision rule (or the statistic) δ is not sufficient: While for the case $X_{20} \neq 0$ it uses the sufficient statistic $T(X_1, ..., X_{20}) = \frac{1}{20}\sum_{i=1}^{20} X_i$, for $X_{20} = 1$ it picks the arbitrary constant $\frac{1}{2}$ as an estimate. Thus, δ violates the WSP explicitly, and when the WSP is adopted, selecting δ would not be allowed. Luckily, these situations are quite rare, and in fact they can be handled by reformulating the LP to $Ev(E, x)$ should depend on $L(\theta; x)$ *and* x. Details are omitted here because most examples are quite artificial and reformulating the LP solves the appearing problems (Berger and Wolpert, 1988, p. 46-50).

A second criticism targets the legitimacy to apply the concept of sufficiency to mixture experiments. Kalbfleisch (1975) argued that this is not allowed, but gave no argument why a restriction of the concept of sufficiency to non-mixture experiments should hold. Also, a strong argument against this criticism is raised by (Berger and Wolpert, 1988, p. 47), who argue that it is 'impossible to clearly distinguish between mixture and non-mixture experiments'[14] so that if the criticism of Kalbfleisch (1975) would be taken seriously, sufficiency as a concept could not be used anymore. This, in turn, would severely limit the results of classical statistics, which seems absurd, because sufficiency is one of the most elementary and central concepts to classical statistics, as detailed in Chapter 3. This would imply that there are no statistics which compress the data without any loss of information, and there are various easy counterexamples of exactly the statistics fulfilling precisely the definition of sufficiency by including all information available in the data. In summary, the criticism of Kalbfleisch (1975), therefore, is not well justified.

The most severe criticism targets the fact that Birnbaum (1962) represented the experimental structure solely via probability distributions on Ω indexed by the unknown parameter θ. Dawid (1977) entitled this assumption the *Distribution Principle* (*DP*), see also the discussions in Birnbaum (1962), as well as Basu (1975), Fraser (1963, 1969, 1972) and Wilkinson (1977). The representation of the experimental structure via probability distributions seems natural because probability distributions are the central building blocks of parametric models in statistics and probability theory after all. Still, one could question this assumption, especially when considering nonparametric models. No matter if this criticism is taken seriously or not, the implications only hold

[14] Of course, only unless one explicitly uses a randomization device like flipping a coin to decide which experiment to conduct. In practice however, data are observed and the true data generating mechanisms remain unknown. As a consequence, the true nature of this mechanism – mixture or non-mixture – can never be known with certainty.

for settings in which a description of the experiment with probability distributions is judged to be inappropriate.

The answer to this criticism is quite involved. First attempts included a theory of an axiomatic development of the LP which incorporated structural information by Berger (1984). Four years later Berger and Wolpert (1988) noted that this attempt was "something of a failure, containing a suspect axiom from the above viewpoint." (Berger and Wolpert, 1988, p. 47). Finally, it was shown that from the decision-theoretic perspective, violation of the LP leads to inadmissible decisions which do not minimise the incurred loss. Berger and Wolpert (1988, Section 3.7) entertained the project to demonstrate that violating the LP leads to such inadmissible decisions under repeated use. While their derivations are not a direct answer to the criticism of the DP, it unarms the criticism because not following the LP is eventually shown to lead to inadmissible or incoherent behaviour, so that a larger loss is incurred from a decision-theoretic perspective. Therefore, when not following the distribution principle, the decision-theoretic analysis of Berger and Wolpert (1988) shows that this behaviour is inferior to following the distribution principle and the likelihood principle. This situation can, in turn, be seen as an argument *for* the DP, or at least as an argument that the DP needs not to be questioned from a decision-theoretic perspective which judges the losses incurred by following the LP. Berger and Wolpert (1988) summarized the idea as follows:

> "We will not argue that measures of long-run performance have an important practical role in statistics (as frequentists would argue), but we will argue that they have the important theoretical role of providing a test for proposed methodologies: it cannot be right (philosophically) to recommend repeated use of a method if the method has "bad" long run properties. Both of the main approaches to long run evaluation, decision theory and betting coherency, will be discussed."
> (Berger and Wolpert, 1988, p. 51)

Berger and Wolpert (1988) started from the WCP which implies

$$Ev(E^*,(j,x_j)) = Ev(E_j,x_j)$$

where E^* is again the mixture experiment, in which $J = 1$ or 2 is performed with probability 0.5 each and subsequently experiment E_J is performed afterwards. Then, they assumed the LP is violated intentionally which means that

$$f(x_1;\theta_1) = C \cdot f(x_2;\theta_2)$$

for all θ but the evidence obtained is not identical in both experiments:

$$Ev(E_1,x_1) \neq Ev(E_2,x_2) \tag{11.18}$$

Combining the last equations leads to conclusions

$$Ev(E^*,(1,x_1)) \overset{WCP}{=} Ev(E_1,x_1) \overset{\text{Equation (11.18)}}{\neq} Ev(E_2,x_2) \overset{WCP}{=} Ev(E^*,(2,x_2))$$

To illustrate their point, Berger and Wolpert (1988) showed that such behaviour is inferior under repeated use, thereby indicating that even from a frequentist perspective, the DP cannot be questioned. Note that the only way out of the implications of the results shown by Berger and Wolpert (1988) is to reject the WCP, which again leads to the setting described in Section 11.3.5.

The derivations of Berger and Wolpert (1988) are formally no answer to the criticism of assuming the distribution principle. However, they provide a strong counterargument against the position of Neyman (1957): His long-term performance goal was shown to be inferior from a decision-theoretic perspective, leading to inadmissible decisions and a higher incurred loss compared to when following the DP and LP. This proven fact disarms the criticism of Neyman (1957) even when considering experimental sciences like physics or chemistry, in which long-term control of type I errors could be the primary goal after all. Next to the analysis of Berger and Wolpert (1988), another appealing solution to the criticism is to note that the LP makes a statement conditional on two selected statistical models in E_1 and E_2. Questioning that the distributional assumptions made in these models suffice to model the real-world problem correctly is no defect of the LP itself. It just implies that the model – parameterized in this form – may not capture all relevant information to draw any conclusion.

Criticisms questioning the discreteness assumption

While the discrete case handled by Birnbaum (1962) suffices from a philosophical perspective, in practice continuous probability distributions are useful and necessary to simplify computations and obtain a variety of results. Therefore, Berger and Wolpert (1988) extended the discrete proof of Birnbaum (1962) to the continuous case, leading to the *Relative Likelihood Principle* (*RLP*). One main reason for developing such an extension is the question if a likelihood function of a continuous model differs from that of the discrete model it is intended to approximate, which cannot be safely rejected in all generality. Therefore, the validity of the LP in discrete problems may extend to validity in approximating continuous problems, but it also may not. Berger and Wolpert (1988) therefore extended the original results of Birnbaum to the continuous case, which led to the relative likelihood principle. The relative likelihood principle ensures that Birnbaum's theorem holds also in continuous probability spaces, with modified versions of the WSP and WCP. The extension to the continuous case is mostly of technical nature and presents no conceptual challenge to the validity of Birnbaum's result, and details are provided in Appendix B.

11.4 Implications of the (relative) Likelihood Principle

The implications of the LP (or RLP) are farreaching, and the most important consequences are concerned with hypothesis testing, stopping rules and censoring of observed data.

11.4.1 Implications on Hypothesis Testing

The most striking implication of the likelihood principle is the incompatibility with frequentist hypothesis testing in the sense of Fisher's significance testing detailed in Chapter 3 and the Neyman-Pearson testing framework described in Chapter 4. Also, the hybrid approach which evolved out of both methodologies – see Chapter 5 – is incompatible with the likelihood principle. Berger and Wolpert (1988) expressed the problem with all these approaches as follows:

> "The philosophical incompatibility of the LP and the frequentist viewpoint is clear, since the LP deals only with observed x, while frequentist analyses involve averages over possible outcomes."
> (Berger and Wolpert, 1988, p. 65)

Thus, as p-values or rejection regions are calculated as the tails of the test statistic's probability distribution under assumption of the null hypothesis, the evidence obtained depends not only on the likelihood function but on also on values which were *not actually observed*. In Example 11.7, the p-values for the fixed size experiment were based on the binomial distribution and in the variable size experiment on the negative binomial distribution. In the binomial experiment, the probability $\mathbb{P}(t_1 \geq 9|H_0)$ which was calculated for the p-value is based on data $t_1 = 9$, $t_1 = 10$, and so on, until $t_1 = 20$, but the data that was observed was only $t_1 = 9$. Thus, data that was not observed during the experiment is used. The same holds for the calculation of the p-value $\mathbb{P}(k \geq 9|H_0)$ in the negative binomial experiment. Therefore, the resulting evidence in both experiments differs, and in general, different conclusions are reached when frequentist hypothesis testing is applied. Of course, Example 11.7 is not just a special case in which this happens, but in much wider generality frequentist test statistics or p-values will be calculated differently even though the likelihood functions may be proportional in any two given experiments. Even in cases when the proportional likelihoods lead to identical conclusions in frequentist hypothesis testing – in Example 11.7 this would mean that the tail probabilities of the binomial and negative binomial coincide for the observed data x – frequentist hypothesis testing still violates the LP because not only the likelihood function is used for obtaining the evidence $Ev(E,x)$ as required by the Likelihood Principle Corollary . In general, frequentist reasoning is based on a test statistic, which averages in some way over the possibly obtained data, or a p-value, which is defined as a probability over unobserved data. In both cases, more than just the likelihood function is used for drawing conclusions. The problem with averaging over 'more extreme' observations in the spirit of p-values or rejection regions is highlighted in another example of Cox (1958).

Example 11.10 (Cox (1958))**.** Consider that the random variable X has the distributions as specified in the table below under $\mathbb{P}_0$ and $\mathbb{P}_1$:

x	0	1	2	3	4
$\mathbb{P}_0(x)$	.75	.14	.04	.037	.033
$\mathbb{P}_1(x)$	.70	.25	.04	.005	.005

Table 11.1: Example of Cox (1958) against averaging over observations more extreme in frequentist hypothesis testing

Cox (1958) used the test statistic $T(x) = x$ then for a significance test between the two competing hypotheses $\mathbb{P}_0$ and $\mathbb{P}_1$. Large values of x are considered as extreme and when $x = 2$ is were observed, the significance level against $\mathbb{P}_0$ could be written as

$$\mathbb{P}_0(X \geq 2) = .04 + .037 + .033 = .11$$

and the significance level against $\mathbb{P}_1$ as

$$\mathbb{P}_1(X \geq 2) = .04 + .005 + .005 = .05$$

Cox (1958) did not really want to decide between $\mathbb{P}_0$ and $\mathbb{P}_1$, but instead wanted to focus on both the significance tests against $\mathbb{P}_0$ and $\mathbb{P}_1$ with significance levels .11 and .05. The first thing to note based on the above is that $\mathbb{P}_1$ can be rejected at the 5% level after observing $x = 2$, while $\mathbb{P}_0$ cannot even be rejected at the 10% level. Cox (1958) pinpointed the paradox now by looking at the likelihood ratio when considering $\mathbb{P}_0$ and $\mathbb{P}_1$ simultaneously as possible models: The likelihood ratio between $\mathbb{P}_0$ and $\mathbb{P}_1$ is

$$\frac{\mathbb{P}_0(2)}{\mathbb{P}_1(2)} = \frac{.04}{.04} = 1$$

so that both $\mathbb{P}_0$ and $\mathbb{P}_1$ are equally supported by $x = 2$. The indifference which is based *only* on the observed data is in contrast to the strong evidence against $\mathbb{P}_1$ when conducting significance tests instead, which average over more extreme observations than the observation actually observed. This questionable logic behind averaging over more extreme observations was already criticized by Jeffreys (1939) in his famous quote as detailed in Part II, Chapter 6:

> "...a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred."
> (Jeffreys, 1939, p. 316)

In the example of Cox (1958), $\mathbb{P}_1$ is rejected because it does not predict the values $x = 3$ and $x = 4$ which have not occurred. The only reason that $\mathbb{P}_1$ is rejected is that $\mathbb{P}_1$ predicts the unobserved values $x = 3$ and $x = 4$ even less than does $\mathbb{P}_0$. This problem is important in practice because averaging over more extreme observations virtually always has a profound effect on the obtained results of statistical analysis, see also Berger and Wolpert (1988) and Edwards et al. (1963).[15] Even Fisher himself accepted this criticism, one of the very few situations in which he admitted problems with his own work:

> "Objection has sometimes been made that the method of calculating Confidence Limits by setting an assigned value such as 1% on the frequency of observing 3 or less (or at the other end of observing 3 or more) is unrealistic in treating the values less than 3, which have not been observed, in exactly the same manner as the value 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation."
> (Fisher, 1956b, p. 56)

This quote shows that Fisher was aware that his significance tests violated conditional inference. However, the connection between the LP and WCP was unknown at that time and thus he was not forced to decide between conditional inference as mandated

[15] Note that an often stated criticism that the LP fails to consider which observations might have occurred is based on shallow grounds: The LP can incorporate this by specifying the likelihood function for the random variable X observed in a way so that the variation over possibly observed values is expressed in the form of the likelihood function itself. This way, no averaging over (more extreme) values is necessary at all, as all information about this is made explicit in the design of the experiment, which determines the likelihood function.

by the WCP and his significance tests, which are denied when accepting the LP. Another excellent example which highlights the untrustworthiness of results based on averages over more extreme values was given by Berger and Sellke (1987), who adapted the example from Edwards et al. (1963).

Example 11.11 (Edwards et al. (1963))**.** Consider $X = (X_1, ..., X_n)$ is observed with $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. and σ^2 is known. The standard test statistic for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is (following the CLT, see also Rüschendorf (2014))

$$T(X) = \sqrt{n}|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_0|/\sigma \tag{11.19}$$

with $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ being the sample mean. If $T(x) = t$ is observed, the significance level is

$$p = 2(1 - \Phi(t)) \tag{11.20}$$

with $\Phi(t)$ being the standard normal cumulative density function. Reconsidering the situation from a likelihood perspective with H_1 given as $H_1 : \mu = \mu_1$, it would be natural to use the likelihood ratio

$$L_{\mu_1} = f_{\mu_0}(x)/f_{\mu_1}(x)$$

As H_1 consists of all $\mu \neq \mu_0$, this does not suffice of course, but a lower bound on the above likelihood ratio L can simply be given as

$$\underline{L} = f_{\mu_0}(x)/\sup_{\mu \neq \mu_0} f_{\mu_1}(x)$$

Therefore, the evidence against H_0 is no stronger than $\underline{L}$, and an easy calculation[16] shows that in this example $\underline{L} = \exp(-\frac{1}{2}t^2)$, where $t := (x - \mu_0)/\sigma$.

One can now calculate the lower bound $\underline{L}$ for various values of t, and give the associated significance levels for these t-values, see Equation (11.19) and Equation (11.20). Berger and Sellke (1987) did this and the above lower bound is therefore often called the *Berger-Sellke lower bound* in the literature. A summary of their calculations is given in (Berger and Wolpert, 1988, p. 108), who provided the following table:

The striking difference which reveals itself now is that while a p-value of .05 ascertains that the odds are 20 to 1 for the alternative hypothesis H_1 (once in 20 trials one expects a type I error at the .05 level), the lower bound $\underline{L}$ is much larger: The evidence against H_0 certainly is no stronger than $\underline{L}$, but for $p = .05$, the lower bound $\underline{L}$ equals

[16] Note that

$$\underline{L} = \frac{\bcancel{\frac{1}{\sqrt{2\pi}\sigma}} \exp[-\frac{1}{2\sigma^2}(x-\mu_0)^2]}{\sup_{\mu \neq \mu_0} \bcancel{\frac{1}{\sqrt{2\pi}\sigma}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]} \geq \frac{\exp[-\frac{1}{2\sigma^2}(x-\mu_0)^2]}{\exp(0)} = \exp(-\frac{1}{2}t^2)$$

because $\frac{1}{2\sigma^2}(x-\mu)^2 \geq 0$ due to $\sigma \geq 0$ and therefore $\exp[-\frac{1}{2\sigma^2}(x-\mu_0)^2] \in (0,1]$ so that $\sup_{\theta \neq \theta_0} \exp[-\frac{1}{2\sigma^2}(x-\mu_0)^2] = 1 = \exp(0)$. See also Edwards et al. (1963, p. 227), substituting μ and μ_0 for λ and λ_0 in their notation.

t	1.645	1.960	2.576	3.291
p-value	.10	.05	.01	.001
$\underline{L}$	.258	.146	.036	.0044
$\underline{L}g$	.644	.409	.123	.018

Table 11.2: Different t-values and the corresponding lower bound values $\underline{L}$, corresponding p-values, and Sellke-Berger lower bound values $\underline{L}g$ (see below)

.146, so the odds are at best 7 to 1 for the alternative hypothesis H_1, and not 20 to 1 as indicated by the p-value of .05. The situation gets even worse when considering the results of Berger and Sellke (1987), who argued that the lower bound $\underline{L}$ is misleadingly small due to the maximization of the likelihood under the alternative in the likelihood ratio. Therefore, Berger and Sellke (1987) considered to use an average of $f_\theta(x)$ over all values $\theta \neq \theta_0$, which led them to a weighted likelihood ratio

$$Lg := \frac{f_{\mu_0}(x)}{\int_{\{\mu \neq \mu_0\}} f_\mu(x) g(\mu) d\mu} \tag{11.21}$$

where g is some density, which in a Bayesian interpretation would be chosen to be the conditional prior density on H_1, so that Equation (11.21) becomes the Bayes factor as given in Definition 6.11.[17] Using any density g which is nonincreasing as a function of $|\theta - \theta_0|$, Berger and Sellke (1987) showed that the weighted likelihood ratio Lg is at least as large as $\underline{L}g$ given in Table 11.2.[18] Therefore, if one considers $p = .05$, a more realistic interpretation of the true odds of H_1 against H_0 is not 7 to 1 as indicated by the lower bound $\underline{L}$, but 2.5 to 1, as indicated by the corresponding value $\underline{L}g = .409$ for $p = .05$. So, a significant observed t-value of 1.960 (which is based on the original sample $X = (X_1, ..., X_n)$), which corresponds to a p-value of exactly $p = .05$ is barely stating strong evidence against H_0 when interpreted from a realistic perspective, which incorporates a weighted likelihood ratio.[19] Even if any Bayesian reasoning is rejected, the lower bound $\underline{L}$ states no more than odds of 7 to 1 for H_1 instead of 20 to 1 as indicated by the frequentist p-value.

The examples of Edwards et al. (1963) and Berger and Sellke (1987) show how severely frequentist averaging over more extreme values deteriorates the results obtained from hypothesis tests. A more realistic approach uses *all* results instead of only *more extreme* results as it is, for example, done by employing the weighted likelihood ratio Lg. Here, the conditional prior density g on H_1 incorporates all other results, and not only more extreme results. Also, the incorporation happens in the parameter space,

[17] To identify Equation (11.21) with the Bayes factor in this case, it is necessary to assign a mixture prior to the parameter θ which assigns a positive amount of probability mass to the point null value μ_0, see Robert (2007, p. 229) and compare Chapter 7.

[18] The assumption that the prior g is nonincreasing as a function of $|\theta - \theta_0|$ is reasonable as it can be interpreted as a prior which is centered on the null hypothesis value θ_0. Also, the restriction is not severe, as for example flat priors can be chosen in the Berger-Sellke derivation, too.

[19] The Berger-Sellke lower bound therefore gives the maximum Bayes factor which can be obtained under the class of priors g – as chosen in Berger and Sellke (1987) – from a p-value at the specified level. The relationship shows how overreadily frequentist significance measures state evidence against a null hypothesis compared to Bayesian methods, see also Edwards et al. (1963).

as g is a prior distribution, instead of the sample space. The average is thus not taken over samples not actually observed, so no violation of the LP occurs.

Nevertheless, there exist frequentist procedures like Fisher's conditional inference – see Section 3.2.3 – which are not in conflict with the LP. Still, these procedures do not include hypothesis testing because both main frequentist hypothesis testing theories use more than just the likelihood function to obtain evidence $Ev(E, x)$ by observing data x in an experiment E. As a consequence, frequentist null hypothesis significance testing is not allowed when the LP or RLP is accepted. It does not matter if hypothesis testing is interpreted as significance testing according to Fisher (1925a) via p-values, or as hypothesis testing in the spirit of Neyman and Pearson (1933) via test statistics and rejection regions.

Berger and Wolpert (1988) even argued that in many situations frequentist normal distribution theory inference yields the same numerical measures as non-informative prior conditional Bayesian inference, and therefore noted that a "cynic might argue that frequentist statistics has survived precisely because of such lucky correspondences." (Berger and Wolpert, 1988, p. 65). In fact, Bayesian theory offers more than just a few striking examples for this phenomenon, compare Appendix C, Theorem 6.7 and (Held and Sabanés Bové, 2014, Chapter 6).

11.4.2 Implications on Stopping rules

One of the most important consequences of the LP is the *Stopping Rule Principle* (*SRP*), which states that the reason for stopping the experimentation, also called the stopping rule of the experiment, is not relevant for the conclusions drawn about the unknown parameter θ. The SRP follows directly from the LP, and in the continuous case, a more general version can be derived from the RLP. Without exaggeration, Berger and Wolpert (1988) note that

> "The theoretical and practical implications of the SRP to such fields as sequential analysis and clinical trials are enormous."
> Berger and Wolpert (1988, p. 74)

The reason is that when the SRP is adopted, researchers are allowed to stop recruiting participants when already a fraction of the data show overwhelming evidence for either of both hypotheses under consideration and report their results. However, when the SRP is violated, the dependence of the outcome of statistical inference on the stopping rule implies that in the former situation researchers are forced to continue their study as otherwise all calculations will be invalidated.

The first introduction of the SRP goes back to Barnard G.A. (1947, 1949) in the context of sequential analysis. Barnard's position was that an experimenter's intention should not influence the conclusions drawn from the data. The intention to stop after a fixed sample size or to stop only when money or time runs out should not influence the inference. Based on this idea, the SRP was shown to be a consequence of the LP by Barnard et al. (1962). There are various discussions about the implications of the principle, for example, in medical contexts by Anscombe (1963). General discussions are given in Bartholomew (1967), Basu (1975), Berger (1980) and Edwards et al. (1963). The SRP is concerned with the stopping rule in a sequential experiment. Therefore, the concept of a stochastic process is required. A stochastic process is defined as follows (Brémaud, 2020, Definition 5.1.1):

Definition 11.12 (Stochastic process). A stochastic process is a family $\{(X_t)\}_{t\in I}$ of random elements defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking their values in the measurable space $(Z, \mathcal{Z})$.

Thus, Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $(Z, \mathcal{Z})$ a measurable space with σ-algebra $\mathcal{Z}$ and $T \neq \emptyset$ an index set. A stochastic process X is then a family of random variables $X_t : \Omega \to Z$ for $t \in T$, that is

$$X : \Omega \times T \to Z, (\omega, t) \mapsto X_t(\omega) \tag{11.22}$$

is an $\mathcal{F} - \mathcal{Z}$-measurable map for all $t \in T$. Z is called the state space of the process and contains the values X can take. In practice, most often the index set $T := \mathbb{R}_+$ or $T := \mathbb{N}_0$, and the state space Z often is equal to $\mathbb{R}$ with the Borel-σ-algebra $\mathcal{B}(\mathbb{R})$. A sequence of σ-algebras $(\mathcal{F}_t)_{t\in I}$ is called a filtration (in $\mathcal{F}$), when $\mathcal{F}_t \subseteq \mathcal{F}_s$ when $t \leq s$ (Brémaud, 2020, Definition 5.3.4). A process is called adapted to the filtration $\mathbb{F} := (\mathcal{F}_t)_{t\in I}$ when each X_t is $\mathcal{F}_t$-measurable for all $t \in I$ (Brémaud, 2020, Definition 5.3.5). A stopping rule is then defined as follows (Brémaud, 2020, Definition 5.3.10):

Definition 11.13 (Stopping rule). Let $\{F_t\}_{t\in I}$ be a filtration with non-empty index set I. The random variable $\tau : \Omega \to T$ defined on the probability space $(\Omega, \mathcal{F}, P)$ with values in $T := [0, \infty)$ is called a stopping rule with respect to the filtration $\mathbb{F} := (\mathcal{F}_t)_{t\in I}$ when

$$\{\tau \leq t\} \in \mathcal{F}_t \tag{11.23}$$

for all $t \in T$ holds.

Most often, the index set $I := \mathbb{R}$, and the condition that $\{\tau \leq t\} \in \mathcal{F}_t$ can be interpreted as at each time t is is known whether the event of interest has happened or not. Based on the above, a sequential experiment can be formalised as follows:

Definition 11.14 (Sequential Experiment). A sequential statistical experiment E^τ is a triple $(X, \theta, \{f(\cdot|\theta)\})$ where $X = (X_1, ..., X_n)$ is a stochastic process on Ω with probability density function $f(\cdot|\theta)$ for $\theta \in \Theta$ and unknown sample size n which depends on a stopping rule τ.

The SRP states the following (Berger and Wolpert, 1988, p. 76):

Stopping Rule Principle (SRP Berger and Wolpert (1988)). *In a sequential experiment E^τ, with observed final data x^n, $Ev(E^\tau, x^n)$ should not depend on the stopping rule τ.*

In the above, x^n denotes the realization $X_1(\omega) = x_1, X_2(\omega) = x_2, ..., X_n(\omega) = x_n$ of X for $\omega \in \Omega$, and the sample size n is determined by the stopping rule $\tau : \Omega \to [0, \infty)$ with $\tau(\omega) = n$. Recall Example 11.7, where a fixed sample size binomial experiment was compared with a variable sample size experiment. It demonstrated that null hypothesis significance testing could yield different conclusions based on the tail probabilities of the binomial and negative binomial distribution under the null hypothesis H_0, which violates the LP. What is more, the SRP now states that the stopping rule τ in the example – that is, stopping after fixed or variable sample size – should not influence the evidence $Ev(E^\tau, x^n)$ in the sequential experiment E^τ. Therefore, null hypothesis testing does violate not only the LP but also the SRP. This fact was to be expected as the SRP is a consequence of the LP. The stopping rules for fixed sample size can be expressed as

$$\tau : \Omega \to [0, \infty), \tau(\omega) := 12 \tag{11.24}$$

which always stops at $n = 12$.[20] For the variable sample size setting, τ can be expressed as

$$\tau : \Omega \to [0, \infty), \tau(\omega) := \mathbb{1}_{\sum_{i=1}^{n} X_i = n-3}(X_1, ..., X_n) \tag{11.25}$$

and we stop when $\tau(\omega) = 1$. In Example 11.7, the experiment was stopped after observing $r = 3$ zeros, which is equal to observing $n - 3$ successes, which is equivalent to $\sum_{i=1}^{n} X_i = n - 3$ for n flipped coins.[21]

Indeed, the SRP follows immediately from the LP in the discrete case. Therefore, one just needs to assume a sequential experiment E^τ yielding a sequence $(X_1, X_2, ...)$ of observations with common density f_θ and a stopping rule τ, which takes values $(A_n)_{n \in \mathbb{N})} \subseteq \mathscr{P}$, where $\mathscr{P}$ is the power set of Ω, and we stop if and only if $x^n = (x_1, ..., x_n) \in (A_1, ..., A_n)$, and sampling continuous if $x^n \notin (A_1, ..., A_n)$. The stopping time τ then corresponds to the random index n for which $(x_1, ..., x_n) \in (A_1, ..., A_n)$. Then the probability density of the random outcome $X^N = (X_1, X_2, ..., X_{\tau(\omega)})$ is

$$f_\theta^\tau(x^n) = \mathbb{1}_{A_1,...,A_n}(x^n) \prod_{i=1}^{n} f_\theta(x_i) \tag{11.26}$$

Now, Equation (11.26) is the likelihood function when interpreted as function of θ, which is proportional[22] to $\prod_{i=1}^{n} f_\theta(x_i)$. This latter product term does not depend on the stopping rule, and therefore the likelihood principle states that if two different stopping rules τ and τ' are chosen, the only term influenced in Equation (11.26) is $\mathbb{1}_{A_1,...,A_n}(x^n)$. As both likelihood functions therefore are proportional as functions of θ, the evidence is the same according to the LP, no matter if τ or τ' is used. The proportionality constant in the LP in this case is just a quotient of indicator functions not depending on θ.

As already noted, the difference between fixed sample size and variable sample size (also called optional stopping, see Hendriksen et al. (2020) and Kelter (2020b)) in Example 11.7 is striking when viewed from a frequentist perspective. Optional stopping, which is presumably the rule rather than the exception in a variety of research, therefore causes a major problem for classical frequentist statistics, especially frequentist hypothesis testing. The situation is complicated even further:

> "Honest frequentists face the problem of getting extremely convincing data too soon (i.e. before their stopping rule says to stop), and then facing the dilemma of honestly finishing the experiment, even though a waste of time or dangerous to subjects, or of stopping the experiment with the prematurely convincing evidence and then not being able to give the frequency measures of evidence."
> (Berger and Wolpert, 1988, p. 77)

[20] As $\{\tau \leq t\} = \{\omega \in \Omega : \tau(\omega) \leq t\} = \{\omega \in \Omega : 12 \leq t\} = \{\Omega, \emptyset\}$ (either $12 \leq t$, so it holds for all $\omega \in \Omega$, or $12 > t$, then it holds for no single ω), τ is a stopping rule.

[21] As each X_i is $\mathcal{F}_t$-measurable by the assumption that X is a stochastic process adapted to $\mathbb{F} := (\mathcal{F}_t)_{t \in [0,\infty)}$, the sum $\sum_{i=1}^{n} X_i$ is also $\mathcal{F}_t$-measurable, and therefore $\mathbb{1}_{\mathbb{1}_{\sum_{i=1}^{n} X_i = n-3}(X_1,...,X_n)}$ is also $\mathcal{F}_t$-measurable. Thus, $\{\mathbb{1}_{\sum_{i=1}^{n} X_i = n-3}(X_1, ..., X_n) \leq t\} \in \mathcal{F}_t$ for all $t \in [0, \infty)$ and τ is a stopping rule.

[22] Note that Equation (11.26) when interpreted as the likelihood which treats x as fixed and maps $\theta \mapsto f_\theta^\tau(x^n)$ is proportional to $\prod_{i=1}^{n} f_\theta(x_i)$, while the probability density which treats θ as fixed and maps $x \mapsto f_\theta^\tau(x^n)$ of course is not.

It is therefore undesirable from a scientific perspective to violate the SRP. In some situations, especially in the biomedical sciences it may even be questioned if it is ethically legitimate to use frequentist measures of evidence. In presence of the risk,participants are exposed to in clinical trials and given the need for fast development of new drugs or treatments the use of frequentist measures of evidence seems to be hardly justifiable. Even when no ethical conflicts are implied by violating the SRP, the costs and time invested are much higher compared to the situation where optional stopping is allowed (that is, when the SRP and in turn the LP is accepted): When it is mandatory to run the analysis until the fixed sample size n is reached, even if a tenth of the samples planned to be taken already yield overwhelming evidence, the experiment has to be conducted until all n observations are taken because otherwise, the frequentist measure of evidence stated will be false in the sense that the reported measure would not be the correct one as shown in Example 11.7. There, starting with the plan of a fixed sample size n which implies a frequentist measure of evidence based on the binomial setting, and stopping then after some $m < n$ observations make it impossible to use the binomial frequentist measure anymore. The correct frequentist measure needs to be based on the negative binomial setting.[23]

The quintessence of the SRP is expressed nicely by Edwards et al. (1963) in their discussion about the relevance of stopping rules to statistical inference:

> 'The irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by emphasis on significance levels (in the sense of Neyman and Pearson) (...). Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience.'
> (Edwards et al., 1963, p. 239)

What is more, they stressed:

> 'The irrelevance of stopping rules is one respect in which Bayesian procedures are more objective than classical ones. Classical procedures (...) insist that the intentions of the experimenter are crucial to the interpretation of data, that 20 successes in 100 observations means something quite different if the experimenter intended the 20 successes than if he intended the 100 observations. According to the likelihood principle, data analysis stands on its own feet. The intentions of the experimenter are irrelevant to the interpretation of data once collected, though of course they are crucial to the design of the experiments.'
> (Edwards et al., 1963, p. 239)

The distinction made by Edwards et al. (1963) is subtle but important, in that the experimenter may very well design different experiments, for example a fixed and variable

[23] One could argue that the correct sampling distribution could be selected after performing the experiment in frequentist hypothesis testing. It is illusionary to assume that this is routinely done in practice, in particular, because the standard tests developed by Fisher and Gosset assume fixed sample sizes, compare Part I. Also, Neyman-Pearson tests are conducted in practice after a power analysis, wherein the necessary sample size n is determined to obtain the desired test power β for a fixed test level α. Therefore, most available frequentist tests assume fixed sample sizes and practitioners often use optional stopping as funding or time runs out (Ioannidis, 2016; Kruschke and Liddell, 2018b).

sample size experiment as in Example 11.7. However, as soon as data analysis starts the intentions of the experimenter need to be irrelevant, and data analysis has to stand on its own feet.

For completeness, it should be mentioned that the SRP can be generalized to the continuous case (Berger and Wolpert, 1988, p. 86-87):

Continuous Stopping Rule Principle (CSRP (Berger and Wolpert, 1988)). *From the RLP, it follows that for any (proper) stopping rule τ,*

$$Ev(E^{\tau},(n,x^n)) = Ev(E^n,x^n)$$

for $\{\mathbb{P}_{\theta}^{\tau}\}$-almost everywhere (n,x^n), that means the evidence concerning θ in E^{τ} is identical with that for the fixed sample size experiment E^n (with the observed n), so that τ is irrelevant.

A proof can be found in Berger and Wolpert (1988, Section 4.2.6). In their treatment Berger and Wolpert stressed that while a Bayesian may also not be able to avoid such considerations as the prior probability construction may be influenced by the experimenters' intentions, too, these intentions are made *explicit* in the prior formulation. The prior elicitation is available and made transparent in any Bayesian analysis.[24] Frequentist hypothesis testing shoves these intentions under the rug when reporting the results. The intentions are at best *implicit*, while it may be reasonable to argue that in many cases, no distinction – although necessary – between different stopping rules is made at all by (especially statistically untrained), the majority of researchers in the empirical sciences belong to. If no new analysis is conducted when changing the stopping rule, the reported frequentist measure of evidence will be incorrect.

The above analysis explains various of the experienced problems observed in the replication crisis detailed in Chapter 1 on a purely axiomatic basis. Importantly, these errors are forced by the scientific system, because when money or time runs out in a study so that the planned sample size cannot be reached, optional stopping happens. If no new analysis is conducted, an error occurs inevitably based on the above considerations (unless the statistical analysis is completely recalculated, which is very unlikely to happen in practice, given the standard tests of Fisher and Neyman-Pearson, compare Part I).

11.4.3 Implications on Censoring

Next to the strong implications on hypothesis testing and stopping rules, the (relative) LP also has consequences for censoring. Censored data are often observed in medical and social science, when an event like death, job change or divorce can be observed in a specified time window, but if it is not data are censored. The status of the event remains unknown to the experimenter. This is the case when time or money runs out, and a study ends, but the event was not observed until then. For example, a participant may not show up for the second and third of three follow-up examinations of a study so it remains unknown if the event of interest (for example, death) has happened or not when the study ends. Thus, the observation is censored after the first follow-up examination and the investigators do now know whether and if so, when the event happened. The most popular method for analyzing censored data is survival analysis, and details can be found in Klein et al. (2014) and Ibrahim et al. (2001).

[24] They also noted concerning frequentist inference, that '...we have to know what the experimenter's intentions were. Trying to analyze hard data by guessing what the experimenter was thinking before doing the experiment seems rather strange.' (Berger and Wolpert, 1988, p. 79)

Definition 11.15 (Censoring)**.** Censoring happens when instead of observing the random variables $X_1, ..., X_n$ for $n \in \mathbb{N}$, only the following variables are observed:

$$(Y_i, \delta_i) \text{ with } Y_i := X_i \text{ and } \delta_i = 1 \text{ if } X_i \text{ is actually observed} \tag{11.27}$$

$$(Y_i, \delta_i) \text{ with } Y_i < X_i \text{ and } \delta_i = 0 \text{ if } X_i > Y_i \tag{11.28}$$

In the first case, $\delta_i = 1$ indicates that no censoring happened and the original X_i is observed. In the second case, $\delta_i = 0$ indicates that it is known that $X_i > Y_i$, but only Y_i is observed. In the latter case, Y_i for $X_i > Y_i$ is called a censoring time (Klein et al., 2014). In practice, often fixed censoring is assumed, so that the censoring time is not a random variable itself. This happens, for example, when a study ends after a fixed time and this time is known in advance:

Definition 11.16 (Fixed censoring)**.** Fixed censoring occurs, when instead of the random variable X, the censored variable $Y := g(X)$ is observed, where g is a known function which maps from Ω to $\tilde{\Omega}$. The experiment performed therefore changes from $E = (X, \theta, \{\mathbb{P}_\theta\})$ to $E^g = (Y, \theta, \{\mathbb{P}_\theta \circ g^{-1}\})$ where for $A \subset \tilde{\Omega}$, $g^{-1}(A) = \{x \in \Omega : g(x) \in A\}$.

Thus, a fixed censoring mechanism can be interpreted as a function which maps the original data $(X_1, ..., X_n)$ to the data $(Y_1, ..., Y_n)$, which are complemented by the variables $(\delta_1, ..., \delta_n)$ and which indicate whether observation X_i is a censoring time. From the perspective of a practitioner it is problematic that the mechanisms which cause the censoring can be quite complicated. For example, in survival analysis, a censored observation may happen due to the death of a patient (no more measurements are taken after the death), or due to the last follow-up session in the study, after which the study was terminated due to financial limits (no more measurements are taken either in this case). In most cases, the LP will imply the *Censoring Principle* (CeP), which states that only the results of the censoring and not of the censoring mechanism itself are relevant for inference about the unknown parameter θ. Maybe the most important implication of the CeP is that the impact of an uncensored observation on the obtained evidence is the same, no matter if the observation was observed in an experiment in which censoring was possible or in an experiment with no censoring. Originally, this principle can be attributed to the discussion of Pratt (1961) of Erich Lehmann's influential monograph *Testing Statistical Hypotheses* and to Pratt (1965). While the CeP seems reasonable to be assumed, the original example of Pratt in the discussion of Birnbaum's 1962 paper strikingly shows why the CeP should hold in practice to avoid absurd situations:

Example 11.17 (Pratt (1962))**.** An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that the measurement error is negligible to the variability in tubes. A statistician takes a look at the measurements which seem to be normally distributed and vary from 75 to 99 volts with mean $\mu = 87$ and standard deviation $\sigma = 4$. He makes an ordinary normal analysis which yields a confidence interval for the mean μ. Later, he visits the engineer's laboratory and notices that the volt-meter measures only up to 100 volts, so the data technically are now "censored". Therefore, a new analysis based on the censored data is necessary if the statistician is conservative. The engineer now says he has another volt-meter of equal accuracy which reads up to 1000 volts, which he would have used if any voltage would have been above 100. Therefore, the statistician is relieved because the population was not censored at all technically. Next day, the engineer calls the statistician and tells him on the telephone that his high-range voltmeter was broken at the day he did his experiment. The statistician then ascertains that a new analysis will be required, as now the data are censored at 100 volts. The engineer is astonished, replying that the experiment turned out just the same as if the high volt-meter had been working and the precise voltages were obtained anyway, so he learned exactly what he would have learned if the high volt-meter had been working correctly.

The example of Pratt demonstrates that traditional frequentist methods lead to paradoxical conclusions when data are censored.[25] This conflict is an immediate consequence of frequentist confidence intervals violating the likelihood principle. As the likelihood function changes, depending on which censoring mechanism is applied, the evidence obtained is different depending on which censoring mechanism exactly is at work in the current experiment. Therefore, first, the analysis needs to be changed from an uncensored likelihood to a censored one when the statistician gets to know that the volt-meter reads up only to values of 100 volts (the influence to the likelihood function of any observation x yielding ≥ 100 volts is then from a truncated normal distribution, see Klein et al. (2014)). When the engineer ascertains him that the high volt-meter would have been used, if any particular value had been over 100 volts, the situation is now uncensored. No measurement happened to be censored at 100 as the values ranged from 77 to 99 volts, so no censoring occurred at all (the influence to the likelihood function of any observation x is not truncated, so the uncensored case is recovered, see Klein et al. (2014)). More importantly, the engineer would have been technically able to proceed with the high volt-meter if any censoring had happened. When the next day, the engineer now calls and states that the high volt-meter was broken, the likelihood conceptually changes back to a censored likelihood at value 100. The reason is that the engineer now would not have been able to use the high volt-meter, even if an observation turned out to yield 100 volts on the low volt-meter.

Now, in the censored case the $1-\alpha$ confidence interval for θ has no longer a coverage probability of at least $1-\alpha$, because a different analysis is required based on the now censored likelihood function

$$L(x;\theta) = \prod_{i=1}^{n} \mathbb{1}_{x_i<100} f(x_i|\theta) + \mathbb{1}_{x_i\geq 100} \int_{100}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} dx \tag{11.29}$$

As the likelihoods are different, they can (and most probably will) lead to different evidence in each case when using frequentist analyses. Pratt argued that this situation is absurd because the functionality of an instrument which was *not* actually used in the experiment changes the evidence obtained. It is remarkable that Pratt found such a compelling example in direct succession to the publication of Birnbaum's theorem. He showed that when the LP is violated, the censoring mechanisms can play a crucial role. Also, he clarified that if the LP is accepted the censoring mechanisms play no role for the evidence obtained and the paradoxical situation in his example vanishes. This is because the CeP below follows from the LP, and according to the CeP, the evidence of an observation from the censored experiment $Ev(E^g, g(x))$ is equal to the evidence of an observation from the uncensored experiment $Ev(E, x)$, if the censoring transformation g is bijective, that is, $g^{-1}(g(x)) = x$ for all $x \in A$ holds. The latter equality holds if

[25] Savage et al. (1962b) pinpointed the problem with frequentist CIs which occurs in Pratt's example as follows: "The only use I know for a confidence interval is to have confidence in it.", also referring to the inability of confidence intervals to make probabilistic statements about the parameter of interest except for a coverage probability, compare Chapter 5.

and only if no observation is censored at all, because only then the reconstruction of x from $g^{-1}(g(x))$ is possible.[26] In the example of Pratt, the evidence therefore is always the same, because both for the low and high volt-meter, no real censoring occurred and thus $g^{-1}(g(x))$ for all $x \in A$ holds, whereby $Ev(E^g, g(x)) = Ev(E, x)$ follows from the CeP below. Thus, although the censored and uncensored likelihood functions may not be proportional to each other, as can be seen from Equation (11.29), the evidence is the same according to the CeP when the censoring mechanisms are equivalent. When considering two fixed censoring mechanisms g_1 and g_2, these are said to be *equivalent* on $A \subset \Omega$ if

$$g_1^{-1}(g_1(x)) = g_2^{-1}(g_2(x)) \qquad \text{for all } x \in A \tag{11.30}$$

A special case is when a single fixed censoring mechanism g is considered, which is said to be *equivalent to no censoring* on $A \subset \Omega$ if for all $x \in A$, $g^{-1}(g(x)) = x$ holds. The CeP now makes the following statement:

Censoring Principle (CeP (Berger and Wolpert, 1988)). *If E^{g_1} and E^{g_2} are two experiments arising from censoring mechanisms equivalent on A for an experiment E, then*

$$Ev(E^{g_1}, g_1(x)) = Ev(E^{g_2}, g_2(x)) \tag{11.31}$$

for all $x \in A$. In the special case when $g^{-1}(g(x)) = x$ for all $x \in A$, then Equation (11.31) *can be replaced by*

$$Ev(E^g, g(x)) = Ev(E, x) \tag{11.32}$$

In the continuous case, the CeP holds $\{\mathbb{P}_\theta\}$-almost everywhere. The CeP itself follows from the LP as shown by Berger and Wolpert (1988, p. 94-95). Therefore, when accepting the LP, the problem of Pratt's example vanishes by employing the CeP. To see that CeP is a consequence of the LP it suffices to note that the censored experiment $E^g = (Y, \theta, \{\mathbb{P}_\theta \circ g^{-1}\})$ uses the family of measures $\{\mathbb{P}_\theta \circ g^{-1}\}$. Therefore, when mechanism g_1 is used, the censored experiment E^{g_1} uses the family of measures $\{\mathbb{P}_\theta \circ g_1^{-1}\}$, and if g_2 is used, the censored experiment E^{g_2} uses the family of measures $\{\mathbb{P}_\theta \circ g_2^{-1}\}$. Now, as g_1 and g_2 are equivalent by assumption of the CeP, Equation (11.30) holds. Therefore, in E^{g_1} the probability of the set $g_1(x)$ is

$$\mathbb{P}_\theta \circ g_1^{-1}(g_1(x)) = \mathbb{P}_\theta(g_1^{-1}(g_1(x)) \overset{\text{Equation (11.30)}}{=} \mathbb{P}_\theta(g_2^{-1}(g_2(x)) = \mathbb{P}_\theta \circ g_2^{-1}(g_2(x)) \tag{11.33}$$

so that the probability of $g_1(x)$ in E^{g_1} and $g_2(x)$ in E^{g_2} are identical, where in eq. (11.33) the equivalency of g_1 and g_2 is used. Now, as the probabilities of the left and right hand side are identical for all θ, the resulting likelihood functions (where it is implicitly assumed Radon-Nikodym derivatives to the measures $\{\mathbb{P}_\theta\}$ exist), which are the

[26] The motivation behind this is clear, since when $g^{-1}(g(x)) = x$ holds for all $x \in A$, the original observation $x \in A$ can be retrieved from $g(x)$ by applying the inverse g^{-1} of the censoring mechanism g on $g(x)$, that is, $g^{-1}(g(x))$. If any true censoring occurs, so that g e.g. censors data at the value 10, then for $x_1, x_2 \in A, x_1 \neq x_2$, $g(x_1) = g(x_2) = 10$ is possible. Then, $g^{-1}(g(x_1)) = g^{-1}(10) = \{x_1, x_2\}$, so it is unclear if the original observation was x_1 or x_2 and the original data x_1 cannot be reconstructed from the censored data $g(x_1)$, if g truly censors the data in any form. The same idea holds for Equation (11.30), where the reconstructions of g_1 and g_2 need to be identical for all $x \in A$.

corresponding densities when interpreted as functions of θ for fixed x, are identical, and the LP with proportionality constant $C = 1$ then states that the evidence obtained needs to be the same in E^{g_1} and E^{g_2}.[27] The general case follows from the RLP in a similar way, for details see Berger and Wolpert (1988, Chapter 4), where also the case of random censoring is treated, which leads to a modified version of the CeP (Berger and Wolpert, 1988, Theorem 7). Finally, it should be stressed that the CeP does *not* state that censoring, in general, is irrelevant for statistical inference. It only states that the evidence provided through an uncensored observation is identical in the experiment $(E^g, g(x))$ where censoring was possible and in the experiment (E, x) where censoring was not possible: This is precisely the statement in Equation (11.32), because when $g^{-1}(g(x)) = x$ for all $x \in A$, no censoring occurred in $(E^g, g(x))$, and then from the CeP the evidence $Ev(E^g, g(x))$ in the experiment E^g where censoring was possible (but where no single observation was censored by applying the censoring mechanism g to the observed data x) is equal to the evidence $Ev(E, x)$ in the experiment (E, x) where no censoring was possible. The contradiction to this rational principle was the cause of the paradoxical situation in the example of Pratt.

11.5 An axiomatic basis for Frequentists – The Confidence Principle

Next to the LP (or RLP), which can be derived axiomatically from the WCP (or CWCP) and WSP (or CWSP), one may reason what axiomatic basis can be revealed behind the long-term performance perspective taken by the Neyman-Pearson theory. Maybe the most natural motivation for a frequentist perspective is to argue that frequency measures are objective and can be assigned a physical interpretation. As science needs objective measures, frequency measures are the natural candidate to proceed with. Also, the use of frequency measures incorporates the desire of repeatable experiments, at least at first glance, because the evidence obtained should be obtained again under exact repetition of the experiment. As a consequence, long-term performance is not an unimportant aspect to consider. Berger and Wolpert (1988, p. 66,67,71) coined the following principle, which goes back to Neyman (1957) and should be followed if long-term performance is the goal:

Confidence Principle (CoP (Berger and Wolpert, 1988)). *A procedure δ is to be used for a sequence of problems consisting of observing $X_i \sim \mathbb{P}_{\theta_i}$. A criterion, $L(\theta_i, \delta(x_i))$, measures the performance of δ in each problem (with small L being good). One should report, as the confidence in use of δ,*

$$\overline{R}(\delta) = \sup_{\tilde{\theta}} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L(\theta_i, \delta(x_i))$$

assuming that the limit exists with probability one.

[27] Suppose no censoring happens and consider Equation (11.29) again: When no data are censored (which is a special case of equivalent censoring mechanisms, see the CeP), the indicator function $\mathbb{1}_{x_i \geq 100}$ in the second summand in Equation (11.29) is always zero. Thus, the likelihoods are even identical with proportionality constant $C = 1$. The LP mandates that evidence is the same in the experiment where censoring was possible but did not occur and in the experiment where no censoring was possible from the beginning.

In the above, $L(\theta_i, \delta(x_i))$ can be associated with a loss function, and $\delta(x_i)$ with a decision rule (compare Appendix C). The term $\frac{1}{n}\sum_{i=1}^{n} L(\theta_i, \delta(x_i))$ can be interpreted as the average loss for n repetitions of the experiment (or making n instead of a single observation), and thus $\overline{R}(\delta)$ describes an upper bound on the asymptotic risk of using the decision rule δ. The benefit of the CoP is that when following it, the actual average performance of the procedure δ is assured to be at least as good as the reported performance $\bar{R}(\delta)$. For example, in the Neyman-Pearson theory the reported test level α may be larger than the actually calculated p-value.

Formally, the CoP does not contradict the principle of adequacy, but its most prominent implementation, the NP-theory, violates the principle of adequacy. By now, it has not been shown that the CoP follows from one of the previously introduced principles, that is for example from the WSP and WCP (or any other principles, which are reasonable to assume). Thus, the confidence principle lacks a solid axiomatic foundation.

An even severer problem with the CoP is that its motivations are questionable: For example, Box (1980) argued that the objectivity assumption of frequentist measures does not hold, simply because choosing a specific model is precisely as subjective as choosing a specific prior in Bayesian inference.[28] In total, objectivity is not an argument for frequentist measures, at least not more as it is for noninformative Bayesian inference. For a detailed discussion about the truthfulness of the statement that repeatability is desirable when the goal is to judge the evidence, see Berger and Wolpert (1988, Section 4.1.5). There, counterexamples are provided which demonstrate that it is improbable to be able to reproduce results at all.

Due to the model assumptions, the reported performance is only an upper bound if the model assumptions made are true, which is hard or even impossible to judge in practice. Therefore, the claim that a Neyman-Pearson test errs in only α per cent of the cases needs to be reduced to the statement that it errs in α per cent of the cases *only if* the model assumptions made by the test are correct. In practice, most often this does not hold, and even slight violations of the model assumptions increase the upper bound drastically, see for example Rochon et al. (2012); Kelter (2021a). Consider the following example, which is adapted from Colquhoun (2014):

Example 11.18 ((Colquhoun, 2014))**.** Consider hypothesis testing between two point hypotheses H_0 and H_1. The UMP level $\alpha := 0.05$ test is used, so one would be tempted to state that with 5% probability a false-positive result is obtained in the long run, that is, $\bar{R}(\delta) = 0.05$ for θ equal to the null value of H_0. Assume the test has power of $\beta := 0.05$, so it does reject H_0 in 5% of the cases if H_1 is true. Taking n hypotheses $H_0^1, ..., H_0^n$ where half of them is true, and half is false, the number of rejections of hypotheses is equal to $n/2 \cdot \alpha + n/2 \cdot \beta$, where the number of false-positive results, is $n/2 \cdot \alpha$. Using n hypotheses leads to a false-positive rate of

$$\frac{\alpha \cdot n/2}{\alpha \cdot n/2 + \beta \cdot n/2} = \frac{0.05 \cdot n/2}{0.05 \cdot n/2 + 0.05 \cdot n/2} = 0.5$$

that means half of the rejections will be in error, not 5% of them. Note also, that while it may be argued that the test power β is artificially small, in practice, it is much more

[28] Also, Berger (1985) and Berger and Wolpert (1988) argued that even nonparametric frequentist statistics are prone to this criticism because the choice of a statistical procedure instead of a parametric model causes the same problems. On the other hand, the often observed equality of results between frequentist and nonsubjective Bayesian procedures makes a strong argument for the use of these nonsubjective Bayesian procedures, because statements about uncertain quantities should be made probabilistically, see de Finetti (2017).

realistic to assume that only a few of the hypotheses proposed are true. So even when increasing β to .8, and assuming 1 out of 20 hypotheses is true, the situation does not become much better.

Therefore, the upper bound $\overline{R}(\delta)$ often says little about the true long-term performance. Next to this problem, it needs to be computed first. Except for traditional Neyman-Pearson tests, derivation of the upper bound $\bar{R}(\delta)$ is complicated for most problems. Putting these conceptual problems aside for a moment, the most severe problem with the CoP was maybe given by Berger and Wolpert (1988), who stressed that when interpreted as the basis for Fisher's significance testing or the Neyman-Pearson theory

> "...it conflicts with the LP. (...) In choosing between the LP and the Confidence Principle, it is important to recall the simple axiomatic basis of the LP, and to realize that no such basis has been found for the Confidence Principle."
> (Berger and Wolpert, 1988, p. 73-74)

Clearly, reporting $\bar{R}(\delta)$ violates the LP, as $\bar{R}(\delta)$ is not based solely on the likelihood function.

11.6 Axiomatic Map

Figure 11.1 visualizes the axiomatic foundations of statistics. In it, arrows indicate that the target principle is implied by the source principle(s). For example, the sufficiency principle implies the weak sufficiency principle. Filled circles indicate that the principle is not implied by the source principle. For example, the sufficiency principle is not implied by the weak sufficiency principle as shown by Birnbaum (1972). Filled squares indicate that the principles are compatible with each other, that is, neither the source principle implies that the target principle does not hold, nor does the target principle imply that the source principle does not hold. For example, the principle of adequacy and the likelihood principle are compatible. Empty squares indicate that all source principles with empty squares are required to imply the target principle with an arrow, and that each of the principles with an empty square is implied by the principle with the arrow. For example, the sufficiency and conditionality principle together imply the likelihood principle, and the likelihood principle implies both the sufficiency principle and the conditionality principle. However, neither the sufficiency principle nor the conditionality principle alone imply the likelihood principle. Empty circles indicate that the source principles are required to solve certain criticisms against the proof of Birnbaum's theorem that the weak sufficiency principle and weak conditionality principle imply the likelihood principle. For example, the distribution principle is required to answer the assumption of Birnbaum that the experiment can be represented by a family of probability distributions. The most elementary principle is the *Adequacy Principle* (AP), which requires a measure of evidence to be able to separate between different magnitudes of evidence. Based on this principle, both the *Relative Likelihood Principle* (RLP) and the *Confidence Principle* (CoP) could be used. In what follows, three perspectives are discussed.

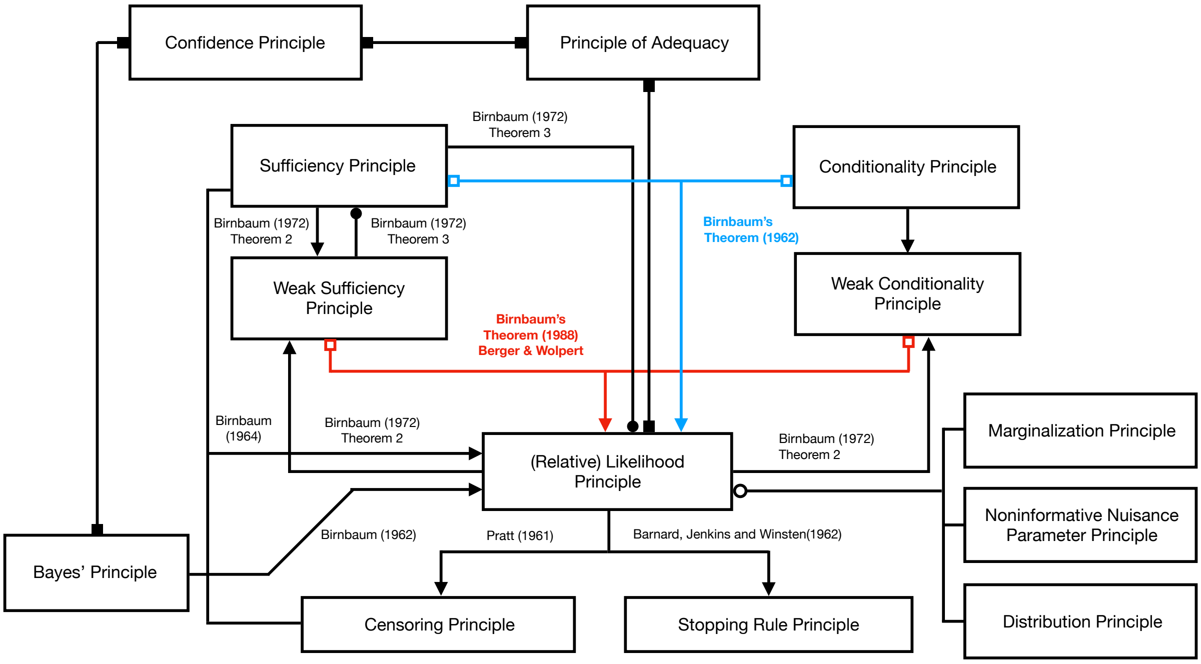

Figure 11.1: Overview and connections between the principles of statistical inference

The Fisherian perspective

The p-value in Fisher's significance tests fulfills the requirement of the AP. Fisher interpreted his p-value as a continuous measure of evidence against the assumed null hypothesis, and thus his theory complies with the AP. As shown in Chapter 3, Fisher strongly advocated conditional inference, so his position would agree with the conditionality and weak conditionality principle. Also, he created the concept of a sufficient statistic and made it a cornerstones of his theory of estimation. Therefore, he would also agree with the sufficiency principle and weak sufficiency principle. Together, these principles imply the RLP, and the evidential interpretation of a p-value would also locate Fisher's theory of significance testing at the RLP because primary interest for him was judging the statistical evidence about a scientific experiment. When interpreting Fisher's position in this way, acceptance of the RLP implies also the stopping rule and censoring principle.

However, Fisher also had long-term performance in mind, which is why he coined the standard threshold of .05. Also, his significance tests violated the RLP as shown above. Therefore, one would locate Fisher's significance testing at the CoP. Still, Part I showed that Fisher's concept was much more in the spirit of the RLP, as he preferred to interpret his p-values continuously and rejected the Neyman-Pearson theory. Thus, Fisher's position conflicts with both the confidence principle and Bayes' principle, which states that all inference follows from the posterior distribution (Grossman, 2011).

Fisher's position is thus somewhat self-contradictory: While he accepted the sufficiency and conditionality principle, his significance tests violated the consequence of these two principles, which is the likelihood principle. Also, his tests violate the stopping rule principle as shown in Example 11.7, and Example 11.17 demonstrated that basing the statistical analysis on observations that were not actually observed violates the censoring principle. Fisher's p-value bases the statistical analysis on observations which were not actually made, so Fisher's perspective conflicts with the CeP, too. Thus, Fisher's perspective can be seen as the precursor of a frequentist perspective which incorporates the results of Birnbaum (1962): While Fisher implicitly accepted the RLP as noticed by Birnbaum (1972, p. 271), he did not strictly adhere to it and this presented no axiomatic inconsistency because Birnbaum's theorem was not proven then. In the same year Birnbaum published his theorem, Fisher died, so it remains speculation if Fisher would have abandoned his significance tests that were in conflict with the RLP to save the cornerstones of his maximum likelihood theory, the sufficiency and conditionality principle.

The Neyman-Pearson perspective

As shown in Chapter 4, the Neyman-Pearson theory separates only between significant and non-significant results, so that test result which corresponds to a Fisherian p-value of $p = 0.04$ is interpreted identically to a test result which corresponds to a Fisherian p-value of $p = 0.01$ for a fixed test level $\alpha = 0.05$. Thus, the NP-theory needs to be located at the CoP, and this is the perspective advocated by Neyman (1957). Neyman's and Pearson's perspective conflicts with the conditionality principle and weak conditionality principle. While they would agree with the sufficiency and weak sufficiency principle, this does not suffice to imply the RLP, and also the AP is violated by their theory of hypothesis testing. As shown in Chapter 4, Neyman and Pearson did not reject Bayes' principle as Fisher did, and they considered it even as an alternative solution for testing statistical hypothesis and estimating confidence sets, but eventually, their theory ended up being a frequentist one.

The Bayesian perspective

Bayesian indices of significance like the Bayes factor which was detailed in Chapter 7 match the requirement of the AP, too. As Jeffreys' Bayes factors is concerned with the evidence provided by the data about both hypotheses under consideration, the Bayesian approach is located at the RLP. Also, as stressed by Birnbaum and shown above, Bayes' principle implies the RLP, so the Bayesian perspective does not require to accept the sufficiency principle or the conditionality principle as axioms. However, the conditionality and sufficiency principle are implied by the RLP, so Bayesians accept these principles as *consequences* of Bayes' principle. Interestingly, the Fisherian perspective – accepting the sufficiency and conditionality principle – are thus immediate consequences of accepting Bayes' principle. This may be seen as the primary reason why Bayesian inference and frequentist inference in Fisher's interpretation often agree, at least asymptotically (which was also noted by Jeffreys' when he compared Fisher's solutions with his own, compare Chapter 7). Bayes' principle does not conflict with the CoP, but it also lacks a clear relationship with it. Furthermore, Bayes' principle is compatible with the AP and the censoring principle and stopping rule principle are implied by Bayes' principle through the RLP.

An axiomatic basis for hypothesis testing in scientific contexts

Based on Figure 11.1, the following axiomatic perspective for hypothesis testing in scientific contexts presents itself: Based on the above sections, the AP is a natural requirement for hypothesis testing in scientific contexts, compare Example 11.1, and violation of the RLP implies that the SP and the CeP are violated, the undesirable consequences of which were highlighted by Example 11.7 and Example 11.17. Thus, any axiomatic basis thus must obey the RLP and AP. Based on this requirement, Fisher's perspective is inadequate as an axiomatic basis for hypothesis testing in scientific contexts. As shown above, although Fisher accepted the AP, his theory of significance tests violates the RLP, and thus also the CeP and SP are violated. Also, Neyman's and Pearson's perspective rejects the AP and opts for the CoP and is thus inadequate as an axiomatic basis, too. Neyman-Pearson tests violate the SP as shown in Example 11.7, conflict with the RLP and with the WCP. Acceptance of the RLP is the only option to benefit from the SP and CeP when these are not accepted as axioms.

From a frequentist perspective, the RLP can thus either be accepted as an axiom or as a consequence of accepting the WSP and WCP. However, the RLP escapes a direct axiomatic motivation as noted by Birnbaum (1962), while acceptance of the WSP and WCP are more natural for frequentists. The Neyman-Pearson theory conflicts with the WCP, so the only option for frequentists is to follow Fisher's perspective. However, Fisher's significance tests conflict with the RLP, too, so this change in perspective presents no solution. In summary, neither the Fisherian perspective nor the Neyman-Pearson perspective obey the RLP and AP.

From a Bayesian perspective, the RLP follows immediately from Bayes' principle as shown above. The CeP and SP are implied by the RLP, and Bayesian hypothesis testing based on the Bayes factor or on posterior probabilities as outlined in Chapter 7 are compatible with the AP. Interestingly, the conditionality and sufficiency principle follow from Bayes' principle, too, as they are consequences of the RLP. Thus, Bayes' principle presents an axiomatic basis for hypothesis testing in scientific contexts which does not conflict with the RLP or AP.

However, some additional principles need to be assumed to answer some criticisms against Birnbaum's proof when accepting Bayes' principle as an axiom: The *Distribution Principle* (DP) as proposed by Dawid (1977) may be seen as a requirement for the RLP, but one can also interpret the principle as superfluous as discussed above. Also, to handle models which include nuisance parameters, the *Noninformative Nuisance Parameter Principle* and the *Marginalization Principle* need to be assumed to answer some criticisms, see Berger and Wolpert (1988).

11.7 Implementation of the Likelihood Principle

Andrei Kolmogorov stressed in his *Foundations of the Theory of Probability* that

> "The theory of probability, as a mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their re-

> lations."
> Kolmogorov (1950, p. 1)

The axiomatic analysis above showed that, in particular, due to the results of Birnbaum a similar development is possible for the theory of statistics. Such a development from statistical axioms or principles strongly encourages the acceptance of the likelihood principle, and thus serves as a first step towards solving some of the problems observed in the replication crisis today. Based on the axiomatic analysis above, all further exposition must be based exclusively on compatibility with the likelihood principle and the adequacy principle to recite Kolmogorov. Bayes' principle as an axiom implies both of these principles, while frequentist theories like Fisher's theory of significance tests or the NP-theory violate either the RLP, the AP, or both.

When following the LP, the remaining question is how to implement it in practice. There are essentially two options available:

1. Conduct purely likelihood-based reasoning, which does only involve maximum likelihood estimation based on $L(\theta; x)$ and reasoning which is based on likelihood ratios $L(\theta_1; x)/L(\theta_0; x)$. The likelihood function, as well as corresponding maximum likelihood estimates, are reported. This mode of inference is called pure likelihoodism[29] (Grossman, 2011; Royall, 1997).

2. Conduct Bayesian inference, which combines the likelihood function $L(\theta : x)$ with a prior $p(\theta)$ to produce the posterior $p(\theta|x)$ of θ given the data x. Subsequently, use posterior indices for significance or the size of an effect like the Bayes factor. Alternatively, one can report posterior point or interval estimates like the posterior mean, median or mode and credible or highest density intervals.

While the first option sounds appealing, concerning hypothesis testing, it prevents practitioners from using any of the tests developed by Fisher or Neyman and Pearson. Even when refraining from hypothesis testing and focussing on parameter estimation alone, the proposal brings multiple problems with it: Informative nuisance parameters often occur in realistic models in practice and make these difficult to analyse via pure maximum likelihood methods (Berger and Wolpert, 1988, Chapter 5). While multiple methods have been developed to deal with these problems occurring in purely likelihood-based reasoning – for example maximizing over nuisance parameters – these often lead to degenerate solutions or even fail to provide solutions at all. Examples include the famous Kiefer-Wolfowitz mixture (Kiefer and Wolfowitz, 1956), which is also detailed in Frühwirth-Schnatter (2006). As highlighted there, this problem holds especially for increasingly high-dimensional models (without the need of any nuisance parameters at all), making the situation even worse as the dimensionality of models is getting larger and larger in the modern era of big data. Especially statistical models in the life sciences include hundreds to thousands of parameters in practical applications, causing much trouble to analytic or numerical purely likelihood-based solutions to work properly. Due to this restrictions, Berger and Wolpert (1988) noted that:

> "The only situations in which pure likelihood methods are completely convincing are simple ones (such as testing two simple hypotheses), where

[29] Mayo (2018) called people proceeding this way *Likelihoodists*. Note that a Likelihoodist will not accept any form of null hypothesis significance testing in the veins of Fisher or Neyman-Pearson, because comparison of pointwise likelihoods in the form of likelihood ratios can cause contradictions to hypothesis tests, especially when involved with composite hypotheses, see (Mayo, 2018).

> they, in fact, correspond to Bayes procedures."
> Berger and Wolpert (1988, p. 125)

As shown in Part I, from a historical perspective these simple low-dimensional models are exactly what these procedures were developed for in the first place. Now, given that the dimensionality of models has exploded and simple (and often, even complex) statistical models are at best an approximation to reality, these methods are questionable from both a conceptual and practical perspective.

The second option is to use Bayesian inference instead to implement the LP. Indeed, Bayesian inference could be used to implement the LP: Birnbaum (1962) noted in his original derivation of the LP, that the likelihood principle

> "...is an immediate consequence of Bayes' principle, when the latter (with any interpretation) is adopted."
> (Birnbaum, 1962, p. 283)

When using Bayes' theorem as a grounded scientific theory which implements probabilistic enumerative induction as detailed in Chapter 10, the LP follows immediately: By incorporation of a prior density $p(\theta)$ on θ, the posterior density $p(\theta|x)$ is given as

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \tag{11.34}$$

where $f(x) := \int_{\Theta} f(x|\theta)p(\theta)d\theta$ is the marginal density. Therefore, the posterior $p(\theta|x)$ depends on the experiment E only through the likelihood $f(x|\theta)$ above, and not through the prior $p(\theta)$, assuming that the selection of $p(\theta)$ is independent of E and the observed x. Note that the elicitation of the prior $p(\theta)$ involves the a priori knowledge about the parameter, for example information from previous studies or incorporation of subject-domain knowledge. The prior elicitation does not depend on the experiment E *itself*.[30]

Therefore, when using Bayes' theorem as a grounded theory of science which implements probabilistic enumerative induction, the evidence $Ev(E, x)$ depends *only* on the likelihood function $L(x;\theta)$ ($f(x|\theta)$ in the above), as required by the RLP. As all Bayesian inference is obtained from the posterior distribution, "the LP is an immediate consequence of the Bayesian paradigm." (Berger and Wolpert, 1988, p. 23). As noted above, the principle to draw all inference from the posterior distribution, which is inherent in any mode of Bayesian analysis, is called Bayes' principle Grossman (2011). Phrased in this way, Bayes' principle implies the RLP.

What is more, if the likelihood functions $L_1(\theta; x)$ and $L_2(\theta; x)$ are proportional for two experiments E_1 and E_2 under consideration, and the same prior $p(\theta)$ is used (which must necessarily be the case as one cannot have two differing beliefs about the same parameter at the same time), the resulting posteriors are identical, compare Equation (11.16). Therefore, the resulting evidence in the form of point or interval estimators like the posterior median, mean or mode, or a posterior highest density interval will be the same for both experiments.

[30] Importantly, this demonstrates that approaches like empirical Bayes methods in which the prior distribution is estimated from the data x are, in general, problematic whenever the same data x is used twice for prior elicitation and the final inference (Kleijn, 2022). When the same data x is used to elicit $p(\theta)$ and obtain $p(\theta|x)$ subsequently, the posterior $p(\theta|x)$ depends on the data x not solely via the likelihood $f(x|\theta)$, but also via the prior $p(\theta)$ that itself now depends on x. Thus, empirical Bayes methods violate the RLP unless data x is split into two sets and the set used for estimating the prior is discarded for the final analysis.

Attempts to argue that Bayesian inference violates the RLP, for example because one could choose two different priors in E_1 and E_2 with proportional likelihood functions $L_1(\theta; x)$ and $L_2(\theta; x)$ also fail. As the parameter θ is required to be the same in both experiments according to the RLP, it is not allowed to use two different priors for the same parameter. Thus, the RLP implicitly denies to select two *different* priors in the two experiments for the *same* parameter θ. As long as the parameter resembles the same physical quantity in both experiments, or is at least identical conceptually in both statistical models, only a single prior can be chosen.[31] Remember the analogy of two experiments, in the first of which a coin is flipped twenty times to estimate the success probability, and in the second of which a heart surgery is conducted at twenty patients. Although the parameters are mathematically the same in both experiments, namely the success parameter in a binomial model, they do not model the same real-world quantity. As a consequence, the RLP does not apply. For example, if all patients who have undergone surgery were high-risk patients, one would judge eight successes out of twenty surgeries as more scientifically relevant than eight successes out of ten coin flips. This observation shows that it is intuitively allowed to choose different priors in the coin flips and heart surgery experiment[32] which is again related to the fact that the parameters are not the same in both experiments and the RLP thus does not apply. When the priors are identical in both experiments, the statistical evidence coincides when the same number of successes is observed in both experiments. However, when different priors are used, the evidence can differ even when the same number of successes is observed. The scientific evidence about the surgery procedure or the coin will, in general, be different.

Furthermore, even when an extremely subjective prior is used in E_1 and E_2, the evidence in form of the obtained posteriors $p(\theta|x)$ both in E_1 and E_2 depends only on the likelihood functions, and not on the extremely subjective prior chosen, as indicated by Equation (11.34). Thus, subjective Bayesian inference does not contradict the RLP.

Summing up, Bayesian inference has multiple advantages compared to a purely likelihood-based view:

1. *Conceptual advantages:* The Bayesian paradigm treats the likelihood $L(\theta : x)$ as a probability density with respect to the presumed prior measure for θ. As Berger and Wolpert (1988) noted:

   > "...probability is the language of uncertainty, so the uncertainty about θ, reflected in $L(\theta : x)$, should be expressed probabilistically."
   > Berger and Wolpert (1988, p. 126), notation changed for consistency

   This argument is appealing, and few if any objections can be raised against it. Note that misinterpretations of the likelihood as a probability distribution are common, a phenomenon which already Fisher (1932) was aware of:

   > "It [the likelihood] is not a probability and does not obey the laws of probability."
   > Fisher (1932, p. 259)

[31] Note that when θ does not describe the same physical quantity in E_1 and E_2, choosing different priors is allowed, and of course the evidence $Ev(E_1, x)$ and $Ev(E_2, x)$ can differ. However, then the RLP does not apply.

[32] Judging eight successes in the heart surgery experiment as stronger evidence for the procedure to work when all patients had high risk is equivalent to assuming a prior for the success rate θ which is not uniform on $[0, 1]$. In fact, the prior will be shifted towards values < 0.5 then.

Fisher's approach of fiducial probability showed that he also acknowledged the advantages of obtaining probabilistic statements about a parameter. However, the fiducial approach turned out to be not successful, see Howie (2002).

2. *Practical advantages:* Statistical inference requires to compare subsets of Θ to another, for example, a null and alternative hypothesis H_0 and H_1, both of which are subsets of Θ. Therefore, some kind of averaging over $L(\theta; x)$ is necessary when H_0 or H_1 is a composite hypothesis, which consists of more than a single parameter value. Comparing every parameter value with another does not work anymore in the common non-discrete setting, which is why heuristics like the likelihood ratio test as given in Definition C.75 were developed. There, the idea is to use at least lower or upper bounds for the likelihood ratio.[33] Therefore, a purely likelihood-based perspective is left with (1) providing a unique maximum likelihood estimate, which is not always possible in practice or (2) sticking with simple likelihood ratio comparisons, which is a stab in the dark in non-discrete high-dimensional models.[34]

   In contrast to the problems inherent to likelihood-based averaging, the Bayesian approach simply proceeds by averaging over the range of credible values θ determined by the form of the prior distribution $p(\theta)$, which in turn influences the values of $L(\theta; x)$ over which the averaging is done when producing the posterior subsequently. Whether it is a posterior probability of the subset $H_0 \subset \Theta$ or $H_1 \subset \Theta$ or a posterior index like the Bayes factor to compare H_0 against H_1, the averaging in the Bayesian approach is explicit in the prior formulation and causes the problems inherent to purely likelihood-based methods to resolve. Unreasonable prior selection may very well end up with the same problems as encountered in averaging over more extreme observations. This holds especially for completely uninformative, flat priors like $p(\theta) = 1$.[35] Still, any rational Bayesian analysis will report the used prior, so that the conclusions reached simply can be rejected if the prior seems unreasonable. Also, for a large variety of models, there is wide agreement on which priors to use. These turn out to be weakly informative in most cases to make complex models treatable via a slight restriction of the parameter space through the prior $p(\theta)$, see Gelman et al. (2013) and McElreath (2020). What is more, incorporating prior information is often seen as a drawback of the

[33] The same holds for the Berger-Sellke lower bound as given by (Berger and Sellke, 1987), but here the averaging is done over the parameter space by introducing a prior $g(\theta)$. Thus, no violation of the RLP occurs.

[34] Note that one could argue that due to the finite precision of measurements the sample space is always finite so a finite number of comparisons of likelihood ratios suffices. The problem with this argument is that even when the parameter space is discretized as an approximation to the true continuous parameter space, the number of likelihood ratio comparisons necessary to compare all parameter values with each other grows exponentially with the dimension of the model. Assume any reasonable measurement precision, and suppose n grid points are used for each dimension with $n = 1000$. A 100-dimensional model (which is not untypical in biomedical research) would produce a 100-dimensional grid with 1000^{100} points. The number of likelihood ratio comparisons is the number of unordered sequences (likelihood ratios $L(\theta_1; x)/L(\theta_2; x)$ are not distinguished for practical purposes from $L(\theta_2; x)/L(\theta_1; x)$ here) without replacement of size $n = 2$, which is $\binom{1000^{100}}{2}$. The number of necessary comparisons quickly becomes prohibitively large.

[35] Schervish (1995, p. 21) stresses that "an alternative to using improper priors is to do a robust Bayesian analysis.", a suggestion which is followed in this thesis. Note also that a flat prior $p(\theta) = 1$ is no probability density anymore, as $\int_\theta f(\theta) d\theta \neq 1$ in general.

Bayesian approach, but frequentist procedures also include subjectivity in form of selecting a parametric family of distributions (or even nonparametric families of distributions):

> "Some people seem to think that choosing a prior distribution introduces subjectivity into the analysis of data but choosing a parametric family does not. These people are mistaken. Each choice one makes introduces subjectivity."
> Schervish (1995, p. 19-20)

Importantly, even if Bayesian inference adds a layer of subjectivity by introducing a prior on the relevant parameters, the influence of this modeling vanishes for large samples (which still can be problematic for small to moderately sized samples), and the influence can be kept minimal.[36]

Due to these conceptual and practical advantages, Bayesian inference is the most attractive option when searching

(i) a philosophically grounded scientific theory, which is given by Bayes theorem' as an implementation of probabilistic enumerative induction

(ii) an axiomatically justified procedure which follows the RLP to avoid violation of the SRP and CeP for practical reasons and which complies with the AP

(iii) an (easy) applicable method for a wide variety of models and settings in scientific and statistical practice, in particular for hypothesis testing

The remaining problem is the selection of the prior distribution $p(\theta)$. The elicitation of the prior is not easy, which calls out for a *robust Bayesian analysis* as detailed in Berger (1985). The idea of robust Bayesian analysis is to use a class $\mathcal{K}$ of suitable prior distributions and use this class of priors instead of a single guess $p(\theta)$. If the conclusions drawn are essentially the same for all the posteriors produced by all the priors in $\mathcal{K}$, one can regard the problem as solved, or robust to the prior selection. However, this is not always the case, and in practice, making a decision may be warranted without further delay. In many cases, as also noted by Berger and Wolpert (1988), it is possible to restrict further the class of considered priors $\mathcal{K}$ based on prior knowledge so that a unified answer is provided by the class of priors considered. If this is not the case, collecting more data is another option, even if costly.

Carrying out such a robust Bayesian analysis has for a long time been impossible for practitioners. It is argued in this thesis, that this situation has dramatically changed through the advent of modern MCMC algorithms and available software in the last decade. Therefore it is shown that carrying out a robust Bayesian analysis is straightforward. As Berger and Wolpert (1988) noted already 30 years ago

> "A second reason for possible violation of the LP (...) is that many users of statistics will be unable to perform careful robust Bayesian analyses. For these users we must provide simple Bayesian procedures with "built in"

[36] For example, Schervish (1995) noted: "Philosophy aside, suppose that one finds it difficult to specify a prior distribution because one does not have much idea where the parameter is likely to be located. In such cases, one may wish to do calculations based on a prior distribution that spreads the probability very thinly over the parameter space." (Schervish, 1995, p. 20).

> robustness. In part, this robustness should be measured in frequency sense, since the procedures will be used repeatedly (i.e., for different X)."
> Berger and Wolpert (1988, p. 139-140)

Therefore, the biggest obstacle in following the RLP via robust Bayesian analysis is to enable practitioners to conduct such robust Bayesian analyses, and to guarantee that these robust Bayesian procedures also enjoy a good, stable long-term performance.

INTERMEDIATE CONSIDERATIONS

The last chapters showed that Bayesian inference is justified from a philosophical point of view and even more importantly, strongly mandated by an axiomatic analysis of the foundations of statistical inference. The axiomatic analysis of the preceding chapter implies that there is – with the exception of purely likelihood-based reasoning – no alternative to Bayesian inference for quantifying statistical evidence in scientific research. This includes, in particular, statistical hypothesis testing in scientific contexts. Thus, the following Part V contributes several statistical solutions to the ongoing replication crisis from a Bayesian perspective.

First, Chapter 12 shows that computational tools make it possible to perform robust Bayesian hypothesis tests easily for the majority of widely used statistical models in the biomedical sciences. This is a direct and straightforward option to improve the reliability of research by shifting towards robust Bayesian hypothesis tests instead of null hypothesis significance tests in the sense of Fisher or Neyman and Pearson.

Second, Chapter 13 shows that the more advanced Bayesian HMC algorithms provide richer insights than frequentist counterparts, in particular for models like survival analysis which are widely used in the biomedical sciences. Furthermore, it is demonstrated that even complex and highly customised statistical models can be analysed by employing the HMC algorithms which were detailed in Chapter 8.

Third, Chapter 14 studies the behaviour of Bayesian evidence measures and posterior indices and obtains long-term performance results. Although these long-term performance results are not of theoretical nature but based on Monte-Carlo simulations because of the strongly differing theory of the available posterior indices for Bayesian hypothesis testing, they remain beneficial in practice and allow for quantification of long-term performance and error rates of Bayesian hypothesis tests in the two-sample setting, which is among the most widely used research designs in the biomedical sciences.

Fourth, as noted by Berger and Wolpert (1988, p. 139-140), robust Bayesian methods need to be evaluated with regard to their long-term properties to measure the asserted robustness also in a frequency sense. Chapter 14 shows that it is possible via simulation-based approaches to obtain type I and type II error rates for Bayesian procedures, in particular for Bayesian hypothesis tests. This provides a strong justification for using such Bayesian procedures even without formal adoption of the confidence principle. In fact, such results can show that Bayes' principle in Figure 11.1 is compatible with the confidence principle. Although the U.S. Food and Drug administration stressed in its section on Bayesian Adaptive Designs for Clinical Trials of Drugs and Biologics that "Bayesian statistical properties are more informative than Type I error probability" (U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, 2019, p. 20), it is also acknowledged that "Bayesian adaptive and complex trials (...) rely on computer simulations for their design." (U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, 2019, p. 1). Thus, investigating the resulting error rates of Bayesian hypothesis tests can help in judging the reliability of these tests in biomedical research. Furthermore, Chapter 14 shows that Bayesian methods do not only respect the likelihood principle, but often also enjoy long-term performance properties like balancing the type I and II errors more evenly than their frequentist counterparts.

The results show that Bayesian evidence measures – in general – treat the hypothesis testing problem more symmetrical than traditional Neyman-Pearson tests.

Fifth, Chapter 15 provides a new Bayesian solution to the Behrens-Fisher problem and a new method to test for differences between groups. The results indicate that a convenient solution to solve various problems of the replication crisis is to consider small interval hypotheses instead of precise point null hypotheses in practice. This approach goes back at least to Hodges and Lehmann (1954), and theoretical results demonstrate the superiority of the approach to traditional frequentist hypothesis tests.

Finally, Chapter 16 discusses the results, revisits the replication crisis and the provided solutions and presents a perspective on future research venues.

Part V

Bayesian Statistical Solutions to the Replication Crisis in the Biomedical Sciences

Chapter 12

Bayesian Alternatives to Null Hypothesis Significance Testing in the biomedical Sciences with JASP

> It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
>
> Pierre-Simon Laplace
> Théorie Analytique des Probabilités

12.1 Introduction

Null hypothesis significance testing (NHST) remains the dominating inferential approach in medical research (Altman, 1982; Altman et al., 1983; Altman, 1991b,a). The results of medical research therefore stand on the shoulders of the frequentist statistical philosophy, which goes back to the early days of Fisher (1925a) and Neyman and Pearson (1936) as reconstructed in Part I. This chapter presents Bayesian alternatives to null hypothesis significance testing based on p-values and demonstrates, that for the majority of statistical models used in the biomedical sciences robust Bayesian hypothesis tests are available. The use of these tests is directly motivated by the results of the axiomatic analysis in Part IV.

The centerpiece of frequentist inference is a test statistic T, which can be computed from the raw data, and which is known to have a specific distribution F under the null hypothesis H_0. If the observed value of the test statistic passes a given threshold, which is located in the tails of F, then the null hypothesis H_0 is rejected, because observing such a value would be quite unplausible if H_0 were true. The well known p-value states exactly the probability of observing a result as extreme as the one observed or even more extreme when the null hypothesis H_0 were true. To solve the problems inherent to NHST (compare Chapter 1), researchers from the University of Amsterdam have developed the open-source statistical software JASP (JASP Team, 2019), which is an acronym for *Jeffreys Awesome Statistics Package*, referring to Harold Jeffreys (compare Part II). JASP is available for all common operating systems and provides both frequentist NHST as well as Bayesian hypothesis tests. Also, to foster reproducible medical research, JASP offers a seamless integration to the Open Science Framework (Center for Open Science, 2020) as well as shareable JASP-files which include all data and analyses, to promote collaboration between researchers and transparency of statistical data analysis.[1]

In both the hypothesis testing as well as parameter estimation perspective in Bayesian inference, the role of the prior is crucial. The prior distribution quantifies the prior information about any parameters in the model *before* the data x are actually observed. While this may bring a subjective flavour with it, selecting an appropriate prior is a topic of huge relevance in Bayesian literature, as extreme priors can shrink the posterior estimates of a parameter or the obtained Bayes factor into a desired direction specified by the prior shape. Luckily, there is an unspoken agreement to use uninformative priors in most cases (McElreath, 2016; Kruschke, 2015), especially when no prior information (for example in form of results of pilot studies) is available. This makes it easy for most standard tests and methods to select a suitable prior. For example, in the biomedical and cognitive sciences most often the effect size d of Cohen (1988) is important. The effect size is used to quantify the effect of a treatment (e.g. between a treatment and control group) and a priori it is reasonable to assume that very large effects $|d| > 1$ are less probable than small effects $|d| \leq 1$, as often in the biomedical and cognitive sciences small to medium effect sizes $(0.2 \leq |d| < 0.5)$ are observed (Rouder et al., 2009). Common choices of prior distributions for the effect size are the normal distribution (Rouder et al., 2009), t-distribution and the Cauchy distribution (Jeffreys, 1961). A common approach also includes to use uniform priors or priors with extremely large scale parameters like $\mathcal{N}(0, 500)$ if no information is available for the parameter of interest (Kruschke, 2015). However, this approach is problematic and should be avoided, as it can be shown that the a priori assumption then often degenerates to statements which believe much more probability mass in the tails as in the center of the distribution, essentially making the prior distributional assumption questionable. For example, a $\mathcal{N}(0, 500)$ prior will tend to put much more probability mass on unreasonable parameter values than reasonable ones. To be more specific, this prior implies that one believes a priori that $\mathbb{P}(|\theta| < 250) < \mathbb{P}(|\theta|) > 250)$, which is easily shown by calculating $\mathbb{P}(-250 < \theta < 250) \approx 0.38$. Even worse, pioneers of Bayesian inference like Jeffreys (1961) already noticed that such unrealistic overdispersed priors can lead to situations in which the Bayes factor always signals evidence for the null hypothesis H_0, even if the data x are indeed generated by the alternative H_1, a situation which has been entitled the Jeffreys-Lindley-paradox, see Lindley (1957) and Robert (2014). To prevent such problems, often slightly informative or weakly informative priors are used, which span a realistic range of values of the parameter a priori, but are not completely flat (Gelman et al., 2013, 2015; McElreath, 2020).

[1] Recently, various papers have emerged both in the statistical and methodological literature which detail certain aspects of JASP. The latter range from reporting guidelines (van Doorn et al., 2021) to discussions of how to carry out specific analyses which are popular in certain scientific areas (e.g. in psychiatry, compare Quintana and Williams (2018)). These articles include van Doorn et al. (2021), Faulkenberry et al. (2020), Ly et al. (2021), Quintana and Williams (2018), van den Bergh et al. (2021), and this chapter focusses on discussing Bayesian hypothesis testing in JASP for biomedical research. Therefore, a sample of the most widely used statistical tests in medical research is taken and it is shown how to carry out these tests from a Bayesian perspective in JASP and report the analysis.

If a reasonable weakly informative prior is selected, typically Bayes factors between 1/100 and 100 are observed in the biomedical and cognitive sciences, and the reporting guidelines for JASP are therefore built on this scale (van Doorn et al., 2021). While there are multiple offers for translating a Bayes factor into a qualitative statement about the evidence it resembles (Jeffreys, 1961; van Doorn et al., 2021; Good, 1950; Kass and Raftery, 1995; Held and Ott, 2018; Goodman, 1999; Lee and Wagenmakers, 2013), these proposals do not differ drastically (compare Chapter 6) and one benefit is that by reporting the actual Bayes factor instead of "moderate evidence" or "strong evidence" researchers can quantify the evidence based on the reported Bayes factor themselves if desired. The oldest classification or labeling scheme goes back to Jeffreys (1961), and the reporting guidelines of JASP are an adoption of the original Jeffreys scale. The JASP guidelines seperate between "anecdotal", "moderate", "strong", "very strong" and "extreme" relative evidence for a hypothesis based on the size of the Bayes factor obtained. Details about the scale can be found in van Doorn et al. (2021), see also Kelter (2021b) for an overview about the various scales that exist and Table 6.1. While any scale is arbitrary, the scheme of Jeffreys offers a good starting point for judging the relative evidence for the alternative hypothesis compared to the null hypothesis in light of the observed data x. Note that not all circumstances and research contexts require the same scaling: The obtained Bayes factor depends on the prior selected, so that heavily unrealistic hypothesis should require much larger Bayes factors to confirm the a priori unprobable statement[2] in contrast to highly likely hypotheses, which have been confirmed in multiple previous studies already. A research hypothesis with low prior probability will therefore require a convincing Bayes factor such that the evidence overcomes the initial skepticism and the model attains considerable posterior credibility. Therefore, it is important to consider the prior odds carefully when performing such analyses instead of using isolated Bayes factors only. Nevertheless, the scheme proposes a consensus which researchers can use to orient at when reporting results. In particular, it is a good starting point when a weakly informative prior is used. Such priors are prebuilt into JASP and can be selected there.

JASP includes both frequentist and Bayesian methods, and this is a particular strength, as few competitors include that broad a palette of Bayesian methods. Next to this flexibility, ease-of-use is supported through an interactive live view where analyses are performed in real time and added to the results page. The interface of JASP is intuitive and consists of a data page displaying the loaded data set, an analysis page, displaying the analyses which are carried out on this data set, and a results page which includes all results and plots of conducted analyses. In summary therefore, JASP can be judged as flexible and easy to use.

12.2 Methods and Results

To demonstrate how straightfoward the application of Bayesian hypothesis tests in JASP is, three typical questions arising in biomedical research are used as a scaffold: (1) Do multiple groups (treatment one, treatment two, control) differ on an observed metric variable, and if so, how large is the effect size? (2) Do two groups (treatment, control) differ on an observed metric variable, and if so, how large is the effect size between both

[2] See the discussion about Popper's counterexample against probability as a reasonable measure of confirmation of a hypothesis in Chapter 10.

groups? (3) How strong is the relationship between two observed variables? Usually NHST in form of (1) an analysis of variance (ANOVA) (2) a two-sample t-test and (3) linear regression is used to reject a null hypothesis via the use of p-values. In the following, it will be shown that Bayesian versions of these statistical procedures can complement NHST and provide even richer information. A compelling feature here is, that both traditional as well as the Bayesian methods can be run in JASP seamlessly (Wagenmakers et al., 2018; Etz and Vandekerckhove, 2016), so that methodological flexibility is guaranteed.

The results show that the transition from NHST and p-values towards robust Bayesian analyses can be achieved almost effortlessly, as JASP offers an intuitive graphical interface and covers a wide range of Bayesian counterparts for commonly used tests in medical research with rich annotations for correct interpretation and reporting.

Three datasets from medical research were used to compare NHST and Bayesian tests in JASP. The first dataset is from Moore and colleagues (Moore et al., 2012), and consists of 800 patients which had to exercise for six minutes. After the six minutes, heart rates of male and female patients were recorded. All patients were additionally classified as runners or sedentary patients, depending on averaging more than 15 miles per week or not, so that in total two treatment and two control groups of size 200 each sum up to 800 participants.

12.2.1 Question (1) – Analysis of variance (ANOVA)

A typical question in medical research would be to find out any differences between gender as well as both groups, leading to the setting of a 2×2 between subjects ANOVA for the variables group and gender. More specifically, a test for the hypothesis of differing average heart rates between gender and control and treatment groups is desired. The results of the frequentist ANOVA conducted in JASP are shown in Table 12.1. The output shows that both gender and group are significant variables as well as the interaction term for gender and group. All quantities of the ANOVA calculations, sum of squares, degrees of freedom, mean square, F-statistic, η^2 and the p-value are given. Also, the Vovk-Sellke Maximum Ratio (VS-MPR) is given based on the p-value, which is the maximum possible odds in favor of H_1 over H_0.[3]

Table 12.1: ANOVA - Heart Rate

Cases	Sum of Squares	df	Mean Square	F	p	VS-MPR*	η^2
Gender	45030.005	1.000	45030.005	185.980	< .001	1.296e+35	0.110
Group	168432.080	1.000	168432.080	695.647	< .001	1.264e+107	0.413
Gender * Group	1794.005	1.000	1794.005	7.409	0.007	11.062	0.004
Residual	192729.830	796.000	242.123				

Note. Type III Sum of Squares

[3] The Vovk-Sellke Maximum Ratio is similar to the Berger-Sellke upper bound as detailed in Chapter 11, and its name is attributed to the seminal papers of Vovk (1993) and Sellke et al. (2001b). Sellke et al. (2001b, p. 66) showed that a lower bound on the Bayes factor of testing $H_0 : p \sim U(0,1)$ against $H_1 : p\ f(p|\xi)$ is given as $-ep\log(p)$ for $p < e^{-1}$, where p is the p-value, and $f(p|\xi)$ is a beta-density $B(\xi,1)$. The result uses the fact, that the p-value is uniformly distributed under H_0, and the inverse of their bound $-1/(ep\log(p))$ for $p \leq 0.37$ provides an upper bound on the Bayes factor for H_1 against H_0. For the relationship of the Vovk-Sellke maximum ratio to the Berger-Sellke lower bound see (Sellke

One nice feature of JASP is that it offers the option to include assumption checks for the tests conducted: For the ANOVA, homogeneity of variance is required, and the included assumption check in form of Levene's test is given in Table 12.2, showing that the assumption is violated. Still, investigating the provided Q-Q-plot in JASP (see Figure 12.1a) shows that due to the balanced design of 200 participants in each sample and a high power due to 800 participants in total, the ANOVA will be relatively robust to the violations. Conducting a Bayesian ANOVA on the same data in JASP yields the results given in Table 12.3. There are five distinct models for each of which the prior probability $P(M)$, the posterior probability $P(M|data)$, the change from prior odds to posterior odds BF_M for each model, and the Bayes factor BF_{10} for the relative evidence

Table 12.2: Test for Equality of Variances (Levene's)

F	df1	df2	p	VS-MPR*
5.562	3.000	796.000	$< .001$	59.104

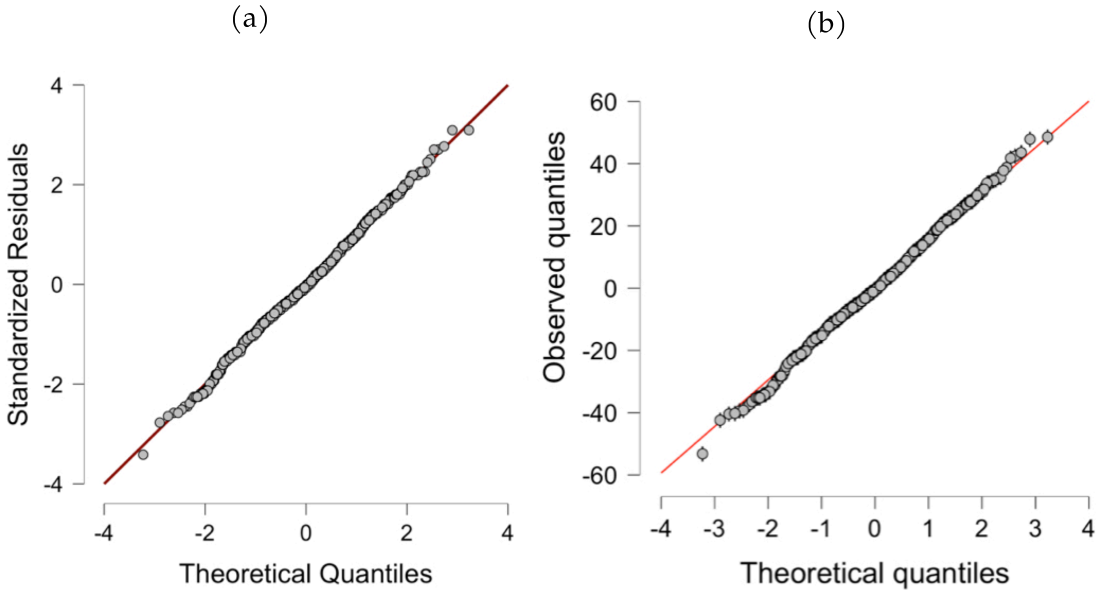

Figure 12.1: Q-Q-plots for the traditional and Bayesian ANOVA for the heart rate dataset of Moore and colleagues produced by JASP

Table 12.3: Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model	0.200	2.281e-126	9.124e-126	1.000	
Gender + Group + Gender * Group	0.200	0.790	15.047	3.463e+125	2.485
Gender + Group	0.200	0.210	1.063	9.207e+124	1.068
Group	0.200	6.651e-36	2.661e-35	2.916e+90	2.683e-95
Gender	0.200	1.797e-107	7.186e-107	7.876e+18	2.699e-23

et al., 2001b, Example 3).

of the alternative hypothesis H_1 compared to the null hypothesis H_0 as well as the error in percent is given. This is necessary, because for some analyses the results are based on numerical algorithms such as Markov chain Monte Carlo (MCMC), which yields an error percentage (for more details on the computation see van Doorn et al. (2021)). The error percentage thus is an estimate of the numerical error in the computation of the Bayes factor via Gaussian quadrature in the BayesFactor R package (Morey and Rouder, 2018) JASP uses internally, and values below 20% are deemed acceptable (Bergh et al., 2019). If the error percentage is deemed too high, the number of samples can be increased to reduce the error percentage at the cost of longer computation time. Also, the BF_M column shows the change from prior odds to posterior odds for each model. For example, for the full model including both main effects as well as their interaction effect, the prior odds are $0.2/(1-0.2) = 0.25$, while the posterior odds are $0.790/(1-0.790) = 3.761905$, leading to a ratio of $3.761905/0.25 = 15.04762$, as shown in the BF_M column. All models are compared to the null model here, where the null model includes no predictor variables at all, and the full model includes both variables gender and group as well as their interaction term. It is clear that the BF_{10} of $3.463e+125$ is largest for this last most complex model, indicating extreme evidence for this model according to the scale of Jeffreys and the reporting guidelines for JASP (van Doorn et al., 2021). Also, the BF_{10} column contains the Bayes factor that quantifies evidence for this model relative to the null model with no variables included, therefore it is 1 for the null model row. While the BF_M column thus states that the most complex model is the most probable a posteriori (because the prior odds were identical for all models, so that BF_M is largest iff $P(M|data)$ is largest), the BF_{10} column also adds that the data support this model best when ignoring the prior odds (the Bayes factor is, in general, independent of the prior odds, see Kleijn (2022)). It may be of interest to obtain a Bayes factor $BF_{10}(\mathcal{M}_{\text{main effects vs. full}})$ for comparison of the full model including the interaction effect, and the model with both main effects. This is straightforward, as due to the transitivity of the Bayes factor, it is clear that

$$\frac{BF_{10}(\mathcal{M}_{\text{main effects}})}{BF_{10}(\mathcal{M}_{\text{full}})} = \frac{\frac{p(x|H_1^{\mathcal{M}_{\text{main effects}}})}{p(x|H_0^{\mathcal{M}_{\text{null}}})}}{\frac{p(x|H_1^{\mathcal{M}_{\text{full}}})}{p(x|H_0^{\mathcal{M}_{\text{null}}})}} = \frac{p(x|H_1^{\mathcal{M}_{\text{main effects}}})}{p(x|H_1^{\mathcal{M}_{\text{full}}})} = BF_{10}(\mathcal{M}_{\text{main effects vs. full}})$$

because the denominators $p(x|H_0^{\mathcal{M}_{\text{null}}})$ cancel each other out, so that dividing the main effects model Bayes factor $BF_{10}(\mathcal{M}_{\text{main effects}}) = 9.207e+124$ by the full models Bayes factor $BF_{10}(\mathcal{M}_{\text{full}}) = 3.463e+125$ yields a Bayes factor $BF_{10}(\mathcal{M}_{\text{main effects vs. full}}) \approx 0.2658677$ for comparing the main effects model to the full model, which also indicates that the full model is to be preferred. This Bayes factor can also be calculated in JASP by selecting *compare to best model* instead of *compare to null model* in the user interface. Figure 12.1b shows a Q-Q-plot for the residuals of the Bayesian ANOVA, showing that it is quite robust to the deviations from normality.

A compelling feature of the Bayesian way now is that posterior credible intervals on all variables of interest are easily obtained. While often frequentist confidence intervals are interpreted as containing the true parameter θ with 95% probability, this is actually the correct interpretation of a Bayesian credible interval, after observing the data x as discussed in Chapter 4. Table 12.4 shows the model averaged posterior summaries of the full model for both variables and the interaction term. From the table, one can easily see that females have a posterior mean of 7.448, that is an increased heart rate of 7.448

Table 12.4: Model Averaged Posterior Summary

				95% Credible Interval	
Variable	Level	Mean	SD	Lower	Upper
Intercept		124.490	0.551	123.168	125.426
Gender	Female	7.448	0.559	6.339	8.553
	Male	-7.448	0.559	-8.586	-6.373
Group	Control	14.474	0.557	13.334	15.551
	Runners	-14.474	0.557	-15.584	-13.367
Gender * Group	Female & Control	1.465	0.547	0.378	2.577
	Female & Runners	-1.465	0.547	-2.586	-0.387
	Male & Control	-1.465	0.547	-2.586	-0.387
	Male & Runners	1.465	0.547	0.378	2.577

beats per minute, while males have a posterior mean of -7.448, indicating a decreased heart rate of the same magnitude compared to the global mean. Thus, the heart rate seems to be differing between males and females. Specifically, after observing the data x the average heart beat of females lies in the range of values $[6.339, 8.553]$ with 95% probability, so that with 95% we can be sure that females have an increased heart rate of at least $6.339 \approx 6$ beats per minute after exercising 6 minutes compared to the global mean. The 95% credible intervals of males and females do not overlap, so we can be quite confident that there is a true difference.

Other inferences are obtained in identical manner from Table 12.4. Note that the frequentist MLE estimates and confidence intervals cannot offer this flexibility. The values in Table 12.4 can also be obtained as plots in JASP, showing the posterior densities, see Figures 12.2a , 12.2b and 12.2c.

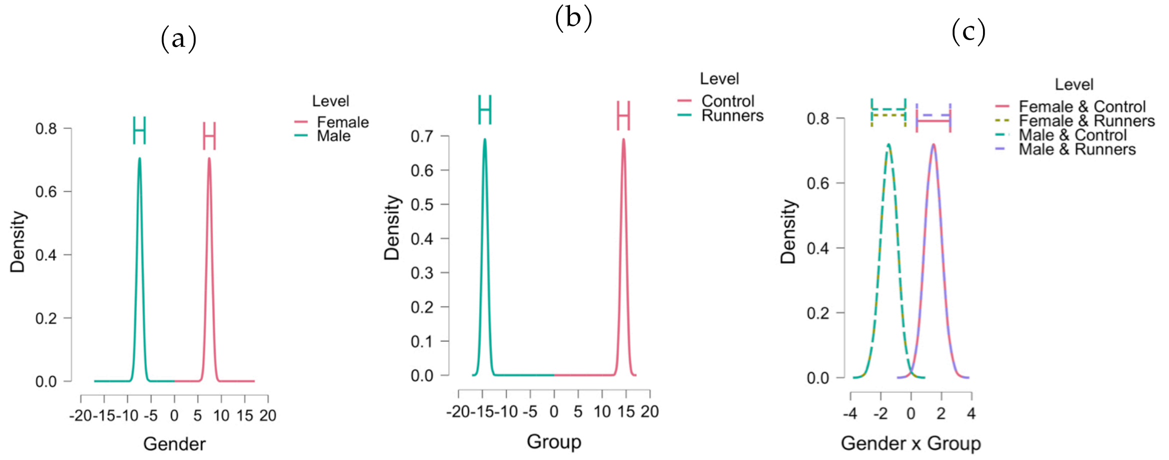

Figure 12.2: Posterior plots for all variables and interaction terms for the heart rate data of Moore and colleagues produced by JASP

12.2.2 Question (2) – Paired samples t-test

Another common situation in medical research is the paired samples t-test which compares the means μ_1 and μ_2 of the same population at two different timepoints (pre-treatment vs. after treatment). The dataset used is again from Moore and colleagues (Moore et al., 2012), and provides the number of disruptive behaviours by dementia patients during two different phases of the lunar cycle. The hypothesis tested is H_0:"Average number of disruptive behaviours in patients with dementia does not differ between full moon and other days" against the alternative H_1 of a differing average numbers of disruptive behaviours. Table 12.5 shows the results of the frequentist paired-samples t-test, indicating with $p < .001$ that H_0 can be rejected. The paired samples t-test therefore suggests that the data (or more extreme data) are unlikely to be observed if the average number of disruptive behaviours was identical during full moon days and other days in patients with dementia. Note that this is not what researchers actually want to know: The desired answer is which hypothesis is more probable after observing the data, which is exactly quantified by the posterior odds $\mathbb{P}(H_1|x)/\mathbb{P}(H_0|x)$, of which the BF_{10} is a key ingredient (the posterior odds are the product of the Bayes factor and the prior odds). A large BF_{10} therefore necessitates a change in beliefs towards H_1. Assumption checks include a Shapiro-Wilk test on normality, which is not significant with $p = .148$. Now, the Bayesian paired-samples t-test shown in Table 12.6 yields $BF_{10} = 1521.058$, indicating extreme evidence for H_1. JASP produces also a plot of the prior and posterior distribution of the effect size δ according to Cohen (1988), which is of interest in most medical research settings (van Doorn et al., 2021). Figure 12.3a shows this prior and posterior plot of the effect size δ as well as the corresponding BF_{10}. A large advantage of the Bayesian paradigm reveals itself here: The posterior of the effect size δ precisely estimates which effect size is most probable after observing the data x. The frequentist paired-samples t-test did not yield any information about the effect size. Although the test was significant, it did not state anything about whether the observed effect is small, medium or large. The prior-posterior plot shows how the prior probability mass is reallocated to the posterior via observing the data and shows that with 95% probability, the true effect size δ is in $[0.818, 2.345]$ and the posterior median is 1.527, indicating a large effect. Another benefit is given by the robustness check plot given in Figure 12.3b: Different prior distribution widths are used for the effect size δ and the Bayes factor BF_{10} is computed. Specifically, the prior width γ of the Cauchy prior $C(0, \gamma)$ on the effect size δ is increased gradually, showing how the prior shape influences the resulting BF_{10}. Figure 12.3b shows that even when changing the prior

Table 12.5: Paired Samples t-Test

			t	df	p	Mean Difference
Moon	-	Other	6.452	14	$< .001$	2.433

Table 12.6: Bayesian Paired Samples t-Test

			BF_{10}	error %
Moon	-	Other	1521.058	5.014e-7

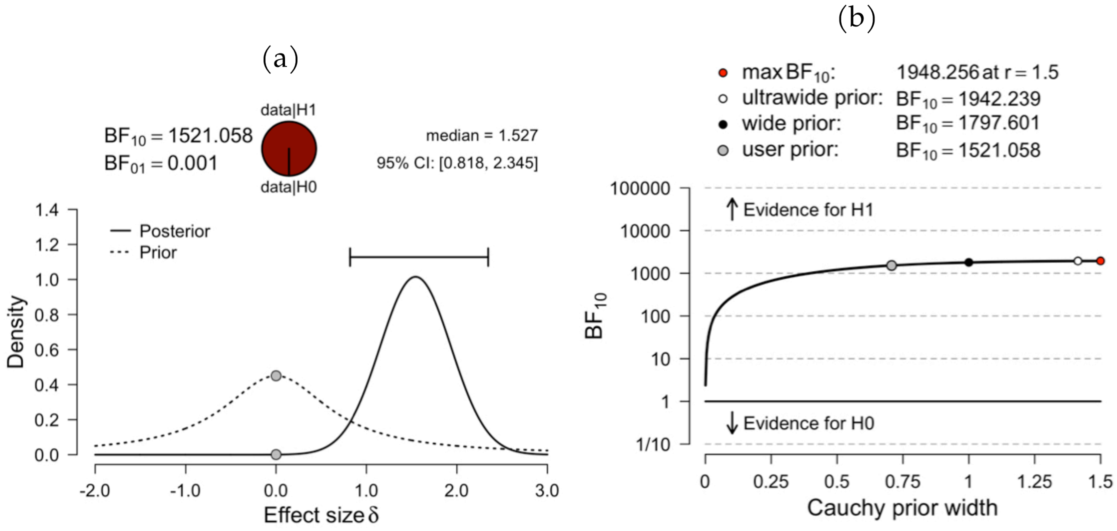

Figure 12.3: Prior and posterior plot and robustness check for the heart dementia data of Moore and colleagues produced by JASP

from the user prior, which equals a medium $C(0, \sqrt{2}/2)$ prior, to a wide $C(0,1)$ or even ultrawide $C(0, \sqrt{2})$ prior, the Bayes factor for H_1 stays above 1000. Thus, the influence of the prior is negligible here, so that only an inconsequential amount of subjectivity goes into the analysis. Such an analysis shows how straightforward an implementation of robust Bayesian hypothesis tests can be.

12.2.3 Question (3) – Linear Regression

One of the most widespread methods in biomedical research and clinical trials is linear regression (Altman, 1991a). The dataset used here is from Mestek et al. (2008) published in the *Journal of American College Health*. The study provided 100 participants' Body Mass Index (BMI) and average daily number of steps, investigating this relationship with linear regression models. A traditional linear regression with the BMI as dependent variable and the average number of daily steps (in thousands) of participants as explanatory variable yields the results given in Table 12.7. The table shows that physical activity (PA) is a significant predictor of the BMI of participants, as $p < .001$. While

Table 12.7: Coefficients

	Unstandardized	Std. Error	t	p
(Intercept)	29.578	1.412	20.948	$< .001$
PA	-0.655	0.158	-4.135	$< .001$

JASP also offers to provide confidence intervals, these are counterintuitive to interpret, and therefore the Bayesian linear regression given in Table 12.8 is preferred. Again, the change from prior to posterior odds for the model BF_M and the Bayes factor for the alternative BF_{10} are given, as well as the models prior probability $P(M)$ and the posterior model probability $P(M|data)$ after observing the data. One can conclude from the results, that the $BF_M = 284.327$ of the physical activity model shows extreme evidence

Table 12.8: Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Null model	0.500	0.004	0.004	1.00	0.00
PA	0.500	0.996	284.327	284.33	0.15

for the model including the variable. Also, the identical BF_{10} for the alternative H_1 relative to H_0, where H_1 states that the regression coefficient for the PA variable differs from zero, shows that the coefficient for the variable is most probably non-zero. The null hypothesis H_0 of a regression coefficient of size zero for the PA variable can thus be rejected based on this result, and even better, the alternative H_1 can be regarded as *confirmed*, which would *not* be allowed when using p-values because accepting hypotheses is generally not allowed in frequentist NHST when interpreted in the sense of Fisher's significance testing. Note that when interpreted from the Neyman-Pearson theory of hypothesis testing, accepting a hypothesis is allowed, but as the Neyman-Pearson theory is only concerned with long-term type I error control, nothing can be said about the hypothesis tested in the performed study or experiment. As Neyman and Pearson (1933, p. 291) state explicitly, their theory "tells us nothing as to whether in a particular case H is true". In fact, the only meaning one can associate to accepting a hypothesis in the Neyman-Pearson sense is to act as if it were true to minimise the long-term loss, but not to actually believe in it, compare Chapter 5.

Furthermore, the PA model explains 15% of the variance observed in the data as can be seen from Table 12.8. Table 12.9 shows the posterior summary of coefficients for the Bayesian linear regression, yielding 95% credible intervals so that inference about the most probable range of coefficient values given the data x can be made. Figure 12.4a shows a plot of the posterior coefficients obtained from the Bayesian linear regression for the BMI data produced by JASP. The Mean and 95% credible intervals are shown, indicating that the PA coefficient is with 95% probability in $[-0.908, -0.326]$, compare Table 12.9. Figure 12.4b shows a residual plot to check the assumption of normally distributed residuals, which seems fine for the Bayesian linear regression model. JASP internally uses the BAS package for R (Clyde, 2020) for these computations.

Table 12.9: Posterior Summaries of Coefficients

						95% Credible Interval	
Coefficient	Mean	SD	P(incl)	P(incl|data)	$BF_{inclusion}$	Lower	Upper
Intercept	23.939	0.366	1.000	1.000	1.000	23.244	24.615
PA	-0.609	0.157	0.500	0.996	284.327	-0.908	-0.326

12.3 Discussion

The comparison of NHST and Bayesian methods in JASP revealed that robust Bayesian analysis, in particular robust Bayesian hypothesis tests are straightforward to perform for the majority of statistical models used in biomedical research. Not only does robust Bayesian inference complement traditional frequentist hypothesis tests and pro-

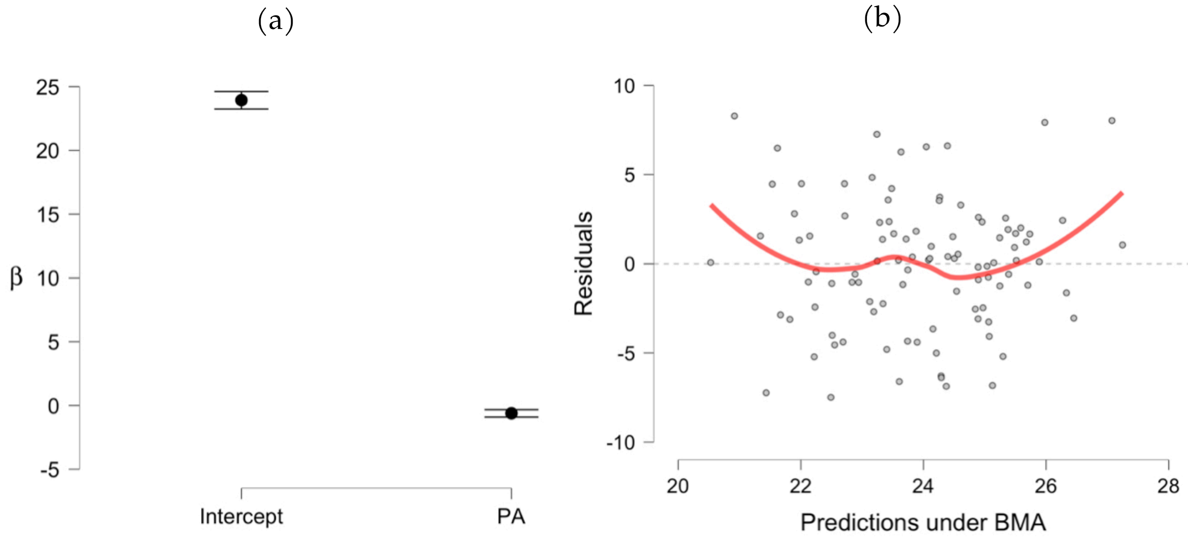

Figure 12.4: Posterior coefficients with credible intervals and residual plot for the BMI data of Mestek et al. (2008) produced by JASP

vide richer information. Robust Bayesian hypothesis tests also avoid the axiomatic conflicts with the likelihood principle, so researchers can benefit from the irrelevance of censoring mechanisms and stopping rules as discussed in Chapter 11. Both of these benefits can be achieved with JASP easily, and the transition from NHST and p-values towards robust Bayesian analyses as an implementation of the likelihood principle is seamless.

Not only can Bayes factors be used to quantify the relative evidence for the alternative hypothesis H_1 compared to H_0 in JASP, but additional parameter estimation with easy to interpret credible intervals allows for richer and easier-to-interpret inference compared to traditional methods. Also, model comparisons and robustness checks can be included into the main analysis to assess the degree to which the conclusions change with background assumptions like the chosen priors, no matter if a t-test, an analysis of variance or a linear regression model is the method of choice. Also, detailed plots and visualisations of results can be created, allowing simple interpretation and communication of the results of a Bayesian hypothesis test. Furthermore, a complete analysis in JASP can be saved in a single JASP-file, which makes it possible to send a conducted analysis to a colleague or even share it publicly. This fosters reproducibility and makes checking results easier for colleagues and reviewers of journals.

There is a large palette of more options for each method (like prior specification, descriptive statistics, providing BF_{01} instead of BF_{10}, inclusion probability for coefficients, and so on) not described here due to space reasons.

Still, although a good spectrum of statistical tests and methods is available in JASP, there are also limitations. Especially for medical research there are some important methods missing. For example, JASP offers no options for survival analysis, which is strongly important in clinical trials (Klein et al., 2014; Ibrahim et al., 2001). Also, more complex generalized linear models are missing, for example there is no Bayesian logistic regression available, a method of large importance for the biomedical and cognitive sciences (Faraway, 2016). Recently, machine learning algorithms like clustering, penalized regression models, linear discriminant analysis and classification and regression trees have been added in form of a machine learning module.

12.4 Conclusion

To demonstrate how straightforward it is to carry out a robust Bayesian hypothesis test, the open-source software JASP was presented, and three worked out examples of common situations in the biomedical and cognitive sciences were provided. These consisted of an ANOVA, a paired t-test and a linear regression model. Conducting and interpreting an analysis in JASP is straightforward and guided by an intuitive interface, and assumptions of a wide variety of tests can be included into the main analysis.

In summary, the results show that JASP provides easy access to advanced (Bayesian) statistical methods, and the transition from NHST towards Bayesian hypothesis tests via the Bayes factor is thus straightforward for practitioners. Also, the effect size which often is of large relevance in biomedical research can be easily estimated in JASP alongside a hypothesis test. In summary, in its current state JASP offers a wide range of Bayesian versions of hypothesis tests which are routinely used in the biomedical and cognitive sciences, and allows seamless transition from NHST to robust Bayesian analysis, in particular, robust Bayesian hypothesis testing.

Chapter 13

Bayesian Survival Analysis in Stan via Hamiltonian-Monte-Carlo

> The most important questions of life are indeed, for the most part, really only problems of probability.
>
> Pierre-Simon Laplace
> Théorie Analytique des Probabilités

The last section demonstrated that robust Bayesian analysis (including Bayesian hypothesis tests) in most standard models in the biomedical and cognitive sciences can be conducted via JASP. However, there exist of course more complex and specialised statistical models, which are not implemented by now. This section shows that for such models, the availability of Hamiltonian Monte Carlo samplers like Stan (Carpenter et al., 2017) provides a straightforward option to implement robust Bayesian analysis. As an example, this section demonstrates how parametric survival models can be analysed via these methods. Survival analysis is an important method in the biomedical and cognitive sciences. Also known under the name time-to-event analysis, this method is also of use in the social sciences and model fitting as well as parameter estimation commonly is conducted via maximum-likelihood. Bayesian survival analysis offers multiple advantages over the frequentist approach, but computational difficulties have mitigated interest in Bayesian survival models in the last decades. This section shows that even complex statistical models like Bayesian survival models can be fitted in a straightforward manner via the probabilistic programming language Stan, which offers full Bayesian inference through Hamiltonian Monte Carlo algorithms. Illustrations show the benefits of a robust Bayesian analysis in contrast to traditional frequentist methods, which highlights that due to the advent of capable HMC algorithms, a robust Bayesian analysis is possible even in complex statistical models.

13.1 Introduction

Survival analysis or time-to-event analysis deals with censored data. This type of data is most often observed in clinical trials where the event often equals death, or in social science, where the event could be divorce or job change of a person. Censored data usually consists of the time $x_i \in \mathbb{R}_+$ and the censoring status ν_i. If $\nu_i = 1$, x_i is observed without censoring (e.g. death, divorce), and if $\nu_i = 0$, x_i is censored so it is unclear what happens after x_i (e.g. because the study time ends or a patient is lost to follow-up). The usual approach for survival data analysis is based on maximum likelihood estimation (MLE), the most prominent approach being the Cox proportional hazards model (Klein et al., 2014). Bayesian analysis on the other hand uses posterior distributions of model parameters to draw inference about them. These posterior distributions are obtained via Markov-Chain-Monte-Carlo (MCMC) algorithms in realistic settings (compare Part III), and Bayesian survival models also rely on MCMC (Ibrahim et al., 2001). In practice, algorithms like Gibbs sampling are necessary to provide posterior inference. This fact made good knowledge of probability theory a must for researchers willing to apply Bayesian methodology to survival analysis. Still, in the last decades, flexible modeling languages for Bayesian inference have grown in popularity. The BUGS language (Lunn et al., 2009) was the first widely used language (Monnahan et al., 2017), and was then made platform-independent in the OpenBUGS language (Lunn et al., 2009). Also, JAGS (Plummer, 2003) was a popular alternative, and these approaches have made Bayesian inference more accessible for practitioners. Stan (Carpenter et al., 2017) can be seen as a relatively new successor to these modeling languages, implementing a new and more efficient Hamiltonian Monte Carlo (HMC) algorithm than its competitors. Instead of Gibbs sampling, most often used by JAGS or OpenBUGS, Stan uses the No-U-Turn sampler as introduced by Hoffman and Gelman (2014) and outlined in Chapter 9. This section focusses on Stan and demonstrates that survival analysis can be carried out in Stan following the Bayesian paradigm.

Stan requires the user first to specify a log density function in its own probabilistic programming language. After that, parameter estimation can be achieved via full Bayesian inference with HMC sampling. Next to this, parameter estimation can be done by approximative Bayesian inference via variational inference (Azevedo-Filho and Shachter, 1994) and the third option is to conduct penalized maximum likelihood estimation with optimization (Carpenter et al., 2017; Gelman et al., 2015).

13.2 Flexibility and Application

Next to its competitive algorithms, Stan offers a highly flexible built-in probabilistic programming language which makes it possible to code nearly arbitrarily complex models for inference. This has benefits and drawbacks, as users can adapt a given model to their specific needs but require at least some theoretical and programming knowledge to do so. Also, Stan's palette of algorithms does include multiple MCMC algorithms like NUTS or plain HMC. The additional possibility to conduct optimization and variational inference offer a wide range of application contexts. Regarding statistical modelling, with a particular focus on survival models, Stan offers to recreate a multitude of models, to modify or extend existing models and thereby can foster a flexible modelling process. In summary Stan can be judged as a highly flexible but equally complex solution, which requires some time to become acquainted with, and to successfully incorporate it into a data analysis workflow based on one of the supported programming languages.

13.3 A detailed Example – Parametric Survival Analysis

To illustrate how robust Bayesian analysis can be achieved even for complex models like survival models, a parametric survival model is used as an example. Most Bayesian survival analyses in clinical research are carried out using parametric survival models (Brard et al., 2017). Therefore, the example presented below uses the parametric exponential model. The exponential model is the most basic model for Bayesian survival analysis and assumes that the survival times $y := (y_1, y_2, ..., y_n)$ are each distributed exponentially with parameter λ, that is

$$f(y_i|\lambda) := \lambda \exp(-\lambda y_i) \qquad \text{for } i = 1, ..., n \tag{13.1}$$

Denoting the censoring indicators as $\nu := (\nu_1, \nu_2, ..., \nu_n)$ where $\nu_i = 0$ if y_i is right censored (lost to follow-up) and $\nu_i = 1$ if y_i is a failure time (death, divorce, job change), the survival function, which is the probability of surviving past the time point y_i is given by

$$S(y_i|\lambda) := \mathbb{P}(T \geq y_i|T \geq 0) = 1 - F(y_i|\lambda) = 1 - [1 - \exp(-\lambda y_i)] = \exp(-\lambda y_i) \tag{13.2}$$

where $F(\cdot|\lambda)$ is the cumulative distribution function of the exponential distribution with parameter λ, and T a random variable modeling the survival time. The observed data D are composed of the number of observations n, the observations y themselves and the censoring status ν, and the likelihood can be written as

$$L(\lambda|D) = \prod_{i=1}^{n} f(y_i|\lambda)^{\nu} S(y_i|\lambda)^{1-\nu_i} \tag{13.3}$$

The likelihood is simply a product of $f(y_i|\lambda)$ for all observations with censoring status $\nu_i = 1$ (death observed) and the survival function $S(y_i|\lambda)$ for all observations with censoring status $\nu_i = 0$. It is possible to use conjugate priors to reach a closed-form posterior, but if one does not want to limit modeling to the Gamma conjugate family of prior distributions for λ, Stan can be used for more flexible modeling.

Covariates need to be incorporated into the model now. One could for example set $\lambda = x_i'\beta$ for a $p \times 1$ covariate vector x_i and a $p \times 1$ regression coefficients vector β, where x_i' is the transposed vector to x_i. The reason that usually $\lambda = \exp(x_i'\beta)$ is used as a predictor instead of $\lambda_i = x_i'\beta$ is simple: Typically, one wants to interpret increasing coefficients β as increasing risk, that is as a decreasing survival function. First, if $\lambda = \exp(x_i'\beta)$, then λ is larger than zero for all coefficients β. Second, if β increases, λ does, too. Third, if β and subsequently λ increases, the survival function $S(t|\lambda) = \exp(-\lambda \cdot t)$ decreases, leading to the desired behaviour. In summary, the exponential survival model can be written as

$$y_i|\nu_i \sim f(y_i|\lambda)^{\nu} + S(y_i|\lambda)^{1-\nu_i} = (\lambda \exp(-\lambda y_i))^{\nu_i} + (\exp(-\lambda y_i))^{1-\nu_i} \tag{13.4}$$

$$\lambda \sim p(\lambda) \tag{13.5}$$

$$\lambda = \exp(x_i'\beta) \tag{13.6}$$

where $p(\lambda)$ is the prior on λ. Listing 1 shows the Stan model code for the exponential survival model where the model is specified directly as a string in the programming language R (**?**).

```
Stan_exponential_survival_model<-"
data{
  int<lower=1> N_uncensored;
  int<lower=1> N_censored;
  int<lower=0> numCovariates;
  matrix[N_censored,numCovariates] X_censored;
  matrix[N_uncensored,numCovariates] X_uncensored;
  vector<lower=0>[N_censored] times_censored;
  vector<lower=0>[N_uncensored] times_uncensored;
}
parameters{
  vector[numCovariates] beta; // regression coefficients
  real alpha; // intercept
}
model{
  beta ~ normal(0,10); // prior on regression coefficients
  alpha ~ normal(0,10); // prior on intercept

  target += exponential_lpdf(times_uncensored | exp(alpha+X_uncensored*beta)); // log-likelihood part for uncensored times
  target += exponential_lccdf(times_censored | exp(alpha+X_censored*beta)); // log-likelihood for censored times
}
generated quantities{
  vector[N_uncensored] times_uncensored_sampled; // prediction of death
  for(i in 1:N_uncensored) {
    times_uncensored_sampled[i] = exponential_rng(exp(alpha+X_uncensored[i,]*beta));
  }
}
"
```

Listing 13.1: Exponential Survival Model in Stan

The Stan model code consists of three core blocks: The `data` block, the `parameters` and the `model` block. Also, the `generated quantities` block is added here, which is optional. The `data` block contains all data variables handed to the Stan model. Here, the number of uncensored and censored observations are defined form of the variables `N_uncensored` and `N_censored`. `numCovariates` is the number of covariates used, which will be one in the example below. Then, the design matrices `X_censored` and `X_uncensored` for the censored and uncensored observations are defined. The last data handed to Stan as input are the observations y_i, split into the censored and uncensored observations in form of the vectors `times_censored` and `times_uncensored`. The `parameters` block includes all parameters posterior MCMC draws are desired from. Interest lies in $\lambda = \exp(x_i'\beta)$, and more specific in the coefficients β, denoted as `beta`. Also, the intercept term is modelled directly via the parameter `alpha` instead of assuming that a design matrix with the first column consisting only of ones. The `model` block then proceeds by computing the likelihood for all observations. The first line beginning with `target +=` uses the log-exponential probability density function for the uncensored observations. Note that $\lambda = \exp(x_i'\beta)$ so that `exp(alpha+X_uncensored*beta)` is the parameter of the log-exponential density. The following line proceeds by adding the log-complement cumulative distribution function for the censored times. The log-complement cumulative density function is defined as $1 - F(x)$, where $F(x)$ is the cumulative distribution function. This is exactly the survival function $S(t)$. Finally, in the `generated quantities` block failure (death) times for the uncensored observations of the input data are generated. This way, failure times for each uncensored observation y_i can be predicted. The priors for β and α are defined as $\mathcal{N}(0, 10)$ here, which is a weakly-informative prior.

The Listing below shows the R-Code to fit the Stan model to the `ovarian` dataset via the interface `rstan`, which is the official interface for R to Stan. The ovarian dataset contains survival times in a randomised clinical trial comparing two treatments for ovarian cancer and can be accessed by installing the `survival` package in R from CRAN.[1]. The predictor used in the example is the treatment used, where the first treatment is coded as 1 and the second treatment as 2. First, the data is prepared into the formats defined in the `data` block of the Stan model, and after that Stan is run.

```
# Prepare data
set.seed(42);
require(tidyverse);
N <- nrow(ovarian);
X <- as.matrix(pull(ovarian, rx));
is_censored <- pull(ovarian,fustat)==0;
times <- pull(ovarian,futime);
msk_censored <- is_censored == 1;
N_censored <- sum(msk_censored);

# Put data into a list for Stan
Stan_data <- list(N_uncensored=N-N_censored, N_censored=N_censored,
numCovariates=ncol(X), X_censored=as.matrix(X[msk_censored,]),
X_uncensored=as.matrix(X[!msk_censored ,]),
times_censored=times[msk_censored],
times_uncensored = times[!msk_censored])
Stan_data

# Fit Stan model
require(rStan)
exp_surv_model_fit <- Stan(model_code = Stan_exponential_survival_model,
 data=Stan_data)

# Print model fit
exp_surv_model_fit
           mean  se_mean   sd   2.5%   25%     50%     75%
beta[1]   -0.67  0.02    0.62 -1.90  -1.06   -0.67   -0.27
alpha     -6.25  0.03    0.92 -8.17  -6.84   -6.21   -5.64
```

Listing 13.2: Exponential Survival Model fit in R via Stan

The model fit shows that the treatment coefficient β_1 has a posterior mean of -0.67, indicating that the second treatment (coded as two) may increase the survival probability of patients. Still, as the 2.5% quantile is -1.90, the estimate is somewhat uncertain. However, as the 75% quantile is still below zero, the effect of treatment two is beneficial with at least 75% probability. Computing other quantiles like a 97.5% quantile is of course possible, too. A likelihood based exponential model would yield a MLE for the treatment coefficient of -0.596 with a Standard error of 0.587, indicating that there is a slightly beneficial effect in the second treatment on the survival time. Still, the p-value would be $p = 0.3$, indicating no significance. However, with only 26 patients the approximations used for computation of the standard error and p-value are highly questionable, making the obtained results questionable as well. The advantage of the Bayesian model is that uncertainty is embraced, and the increased flexibility via the

[1] See https://cran.r-project.org/web/packages/survival/index.html

prior modeling. Few researchers would accept a weakly-informative prior if previous studies indicated evidence of a positive effect for one of both treatments. By modifying the prior parameters, such prior knowledge can be incorporated into the analysis easily, while a frequentist analysis does not offer this option. A second advantage is that it is easy to construct survival functions $S(t|\lambda, x_i = j)$ for given covariate values $x_i = j$. In the above example, there is only one covariate x_1 which is either 1 if the first treatment is used, or 2 if the second treatment is used. Thus, one can compare the estimated posterior survival functions $S(t|\lambda, x_i) = \exp(-\lambda y_i) = \exp[-\exp(x_i'\beta) y_i]$ for different treatments now, where $S(t|\lambda, x_i = 1) = \exp[-\exp(\beta) y_i]$ and $S(t|\lambda, x_i = 2) = \exp[-\exp(2\beta) y_i]$. Figure 13.1 shows the posterior survival functions for the first and second treatment using the posterior mean of β_1, as well as the 2.5% and 97.5% quantiles (lower and upper dotted lines). Overlayed are various survival functions using a range of credible posterior values of β_1. It is clear that while the survival function of the first treatment group decreases much faster, the credible ranges of survival functions for both groups overlap widely. Thus, while a traditional survival analysis using the Cox proportional hazards model would yields a single p-value and at best a point-estimator with confidence intervals, the Bayesian parametric exponential model embraces the uncertainty in the very small dataset of just 26 patients by providing a whole posterior distribution for the treatment coefficient β_1 which in turn leads to a range of credible survival curves, given the data, or given specific covariate values. With regard to hypothesis testing via the Bayes factor, numerical methods like the Savage-Dickey density ratio (Dickey and Lientz, 1970; Verdinelli and Wasserman, 1995; Wagenmakers et al., 2010; Kelter, 2020b) or bridge sampling (Gronau et al., 2017, 2019) allow for computation of the Bayes factor solely based on the prior and posterior densities. As Stan provides the posterior, computing Bayes factors in the above example is therefore also straightforward. For example, testing $H_0 : \beta_1 = 0$ against $H_0 : \beta_1 \neq 0$, the Bayes factor representation according to the Savage-Dickey density ratio is given as

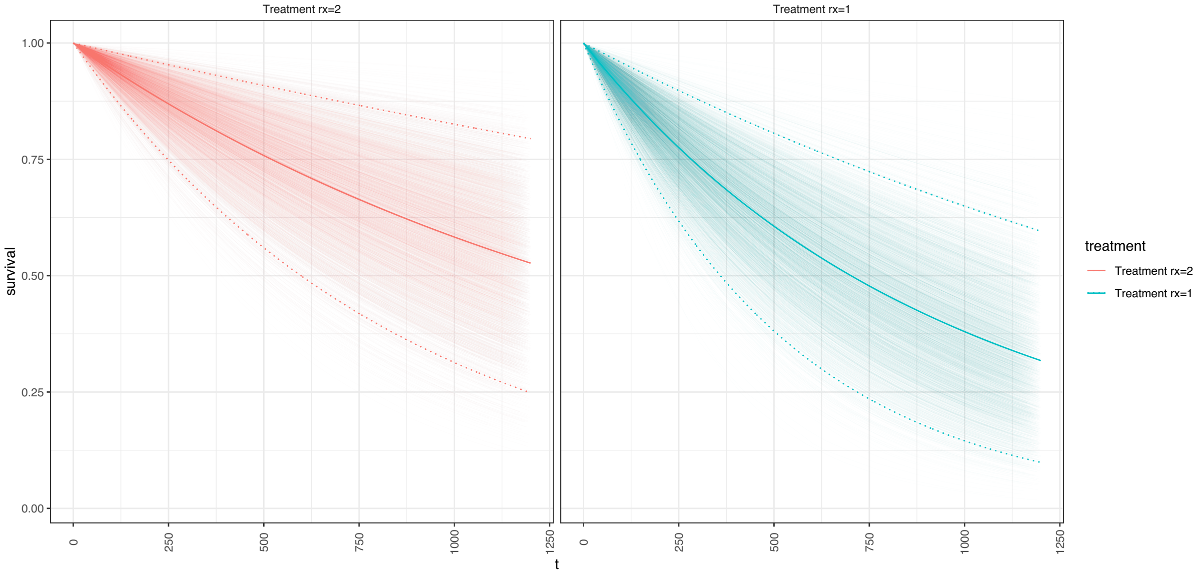

Figure 13.1: Posterior survival functions per treatment group for the `ovarian` dataset

$$\mathrm{BF}_{01}(x) = \frac{p(0|x)}{p(0)} \tag{13.7}$$

where the prior and posterior are computed under $H_1 : \beta_1 \neq 0$. In the example, the resulting Bayes factor based on the Savage-Dickey representation is $\text{BF}_{01} = 9.33$, which indicates moderate evidence for the null hypothesis $H_0 : \beta = 0$. Thus, the Bayes factor confirms what Figure 13.1 already visualised: The difference in risk between both groups is not convincing enough to accept the alternative $H_1 : \beta_1 \neq 0$, and the posterior survival functions per treatment group overlap considerably.

13.4 Conclusion

Stan offers some excellent features for Bayesian inference, which include a highly performative algorithm, interfaces to a wide range of programming languages and customizable program output. The learning curve of Stan is steep, which is in part due to its limitations in form of a highly technical documentation and the missing graphical user interface, but even more to the fact that both programming experience as well as solid theoretical knowledge of Bayesian inference are needed to (1) code the correct model in Stan's probabilistic programming language and (2) run the model via a programming language interface (like the *rstan* package used in the example above). However, as shown in the preceding section, the effort is only necessary for non-standard models which are not covered by JASP: The survival analysis example illustrated that not only is Bayesian inference possible for complex models and simplified by using Stan, also the uncertainty of parameter estimates is embraced, gauging the reliability of the results better than via traditional maximum likelihood based point-estimates. Also, measurement can be seen as associated with the probability model generating the quantities measured, and here Stan plays out another major strength: The probabilistic programming language offers to code complex and highly customizable models of a real phenomenon, therefore enabling researchers to build and subsequently analyse otherwise untreatable probability models in a reasonable amount of time. Only by this added layer of complexity the processing of the results in form of a credible range of survival curves for each treatment in the ovarian example could be achieved. Also, obtaining a Bayes factor without the Hamiltonian-Monte-Carlo algorithm and making use of the Savage-Dickey-density ratio would be difficult. This in turn allows for a precise measuring of the predicted survival time for each treatment condition, and the model could be extended to predict survival time for each patient, each gender or combinations thereof. Also, it allows for a simple application of a Bayes factor test without the need to resort to analytical calculations for each model under consideration.

Chapter 14

Analysis of Bayesian Posterior Significance and Effect Size Indices for the two-sample t-test

> The evidence concerning the possibility of an event occurring usually divides into a part about which statistics are available, or some mathematical method can be applied, and a less definite part about which one can only use one's judgement.
>
> Alan M. Turing
> The Applications of Probability to Cryptography

Chapter 12 showed that robust Bayesian analysis is possible for most standard statistical models in the biomedical and cognitive sciences and Chapter 13 even complex models can be fitted by using Hamiltonian Monte Carlo algorithms. Bayesian hypothesis testing via the Bayes factor can be carried out easily in both cases. However, although the previous sections showed that robust Bayesian analysis is possible for a variety of statistical models, all of the previous discussions focussed on the Bayes factor as the measure which quantifies the evidence about a hypothesis.

In NHST, testing for the significance of an effect is the standard approach, but the significance of an effect does not imply that the discovered relationship is also scientifically meaningful. It only means that the observed effect is unlikely to be observed under the assumption of the null hypothesis, no matter how large or small it is, compare Part I. Also, a non-significant result does not indicate that the null hypothesis is correct, and together these drawbacks of NHST can be seen as the reason why multiple measures of significance and magnitude of an effect based on the posterior distribution have been proposed recently in the Bayesian literature. In practice, drawing conclusions from the posterior distribution is achieved by using different posterior indices or evidence measures. There are measures which state the significance of an effect, and measures which also gauge the size of it. Among them is the Bayes factor introduced by Jeffreys (1961), the region of practical equivalence (ROPE) championed by Kruschke and Liddell (2018b), the probability of direction (PD) as detailed in Makowski et al.

(2019b), the MAP-based p-value proposed by Mills (2018), and the Full Bayesian Significance Test (FBST) featuring the *e*-value, which was introduced by Pereira and Stern (1999) and Pereira et al. (2008). The appropriateness of these indices is still debated in the literature, which makes it challenging to choose among them because by now there is no explicit agreement on which measure researchers should use to report the results of a robust Bayesian analysis (Robert, 2016; Ly et al., 2016a,b; Kruschke, 2018; Kelter, 2020b,a).

What is missing are investigations which of the available measures of significance and effect size are appropriate for a specific Bayesian hypothesis test. The results of such studies can guide researchers in the selection of an appropriate index to assess the results of the Bayesian hypothesis test. In order to provide such guidance, this section investigates the behaviour of common Bayesian posterior indices for the presence and size of an effect in the setting of the two-sample Student's and Welch's t-test, which is among the most widely used parametric two-sample tests in the biomedical and cognitive sciences: Nuijten et al. (2016) showed in a meta-analysis that of 258105 p-values reported in journals between 1985 and 2013, 26% belonged to a t-statistic, see also Wetzels et al. (2011).

14.1 Bayesian Posterior Significance and Effect Size Indices

In this subsection, the existing Bayesian indices of significance and magnitude of an observed effect are briefly outlined which are compared subsequently.

14.1.1 The Bayes factor (BF)

The oldest and still widely used index is the Bayes factor BF_{01}, the evolution of which has been analyzed in Part II, and which measures the change in relative beliefs about both hypotheses H_0 and H_1 given the data x:

$$\underbrace{\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)}}_{\text{Posterior odds}} = \underbrace{\frac{f(x|H_0)}{f(x|H_1)}}_{BF_{01}(x)} \cdot \underbrace{\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}}_{\text{Prior odds}} \tag{14.1}$$

The Bayes factor BF_{01} can be rewritten as the ratio of the two marginal likelihoods of both models, which is calculated by integrating out the respective model parameters according to the prior distribution of the parameters. Generally, the calculation of these marginals can be complex for non-trivial models. In the setting of the two-sample Student's t-test, the Bayes factor is used for testing a null hypothesis $H_0 : \delta = 0$ of no effect against a one- or two-sided alternative $H_1 : \delta > 0$, $H_1 : \delta < 0$ or $H_1 : \delta \neq 0$, where $\delta = (\mu_1 - \mu_2)/\sigma$ is the effect size according to Cohen (1988, p. 20), under the assumption of two independent samples and identical standard deviation σ in each group. An often lamented problem with Bayes factors as detailed in Kamary et al. (2014) and Robert (2016) is the dependence on the prior distributions assigned to the model parameters. Nevertheless, the Bayes factor has deep roots in Bayesian thinking as detailed in Part II and is one of the most widely used Bayesian evidence measures for hypothesis testing. Over the years, several authors including Jeffreys (1961), Kass and Raftery (1995), Goodman (1999), Lee and Wagenmakers (2013), Held and Ott (2018) or van Doorn et al. (2021) have offered thresholds for interpreting different values of it. As detailed in Chapter 12, according to van Doorn et al. (2021), a Bayes factor $BF_{10} > 3$ can be interpreted as moderate evidence for the alternative H_1 relative to the null hypothesis H_0, and a Bayes factor $BF_{10} > 10$ can be interpreted as strong evidence. The Bayes factor BF_{10} can be obtained by inverting BF_{01} in Equation (14.1), that is: $BF_{10} = p(x|H_1)/p(x|H_0) = 1/BF_{01}$. So, if for example $BF_{01} = 4$ states moderate evidence for the null hypothesis $H_0 : \delta = 0$, then $BF_{10} = 1/BF_{01}$ is obtained as $1/4$ for the alternative hypothesis $H_1 : \delta \neq 0$.

14.1.2 The region of practical equivalence (ROPE)

The region of practical equivalence was championed by Kruschke (2015), who stressed that such a region is often observed in different scientific domains under different names "such as indifference zone, range of equivalence, equivalence margin, margin of noninferiority, smallest effect size of interest, and good-enough belt" (Kruschke, 2018, p. 272). The essential idea is that in applied research, parameter values can often be termed practically equivalent if they lie in a given range. Starting from the posterior distribution of the parameter of interest, researchers should interpret values inside the region of practical equivalence (ROPE) as equivalent. For example, when conducting a clinical trial which compares the weight in kilograms of patients in two groups, one could define that the difference of means $\mu_2 - \mu_1$ is practically equivalent to zero if it lies inside the ROPE $[-1, 1]$. That means a difference of only one kilogram is interpreted as practically equivalent to zero. If the posterior distribution of $\mu_2 - \mu_1$ now is entirely located inside the ROPE, the difference $\mu_2 - \mu_1$ is interpreted as practically equivalent to zero a posteriori. On the other hand, if the total probability mass of the posterior distribution $\mu_2 - \mu_1$ is located outside the ROPE, the null hypothesis $\mu_2 = \mu_1$ of no difference can be rejected. The same procedure can be applied to any parameter, θ of interest. If the probability mass of the posterior lies partially inside and outside the ROPE, the situation is inconclusive.

There are two versions of the ROPE, one in which the 95% Highest-Posterior-Density-Interval (HPD) is used for the analysis (95% ROPE), and one in which the full posterior distribution is used (full ROPE). For the effect size δ, Kruschke (2015) proposed to use $[-0.1, 0.1]$ as the ROPE for the null hypothesis $H_0 : \delta = 0$ of no effect, which is half of the effect size necessary for at least a small effect according to Cohen (1988) (a small effect is defined as $0.2 \leq \delta < 0.5$ or $-0.5 < \delta \leq -0.2$ according to Cohen (1988)).

The default ROPEs for effect sizes or regression coefficients are inspired both by mathematical arguments (Kruschke and Liddell, 2018a; Kruschke, 2015) and official guidelines from the U.S. Food and Drug Administration Center for Drug Evaluation and Research (2001), the U.S. Food and Drug Administration Center for Veterinary Medicine (2016) and the U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research (2016). Also, the ROPE itself was independently proposed in a variety of scientific areas, see Carlin and Louis (2009); Hobbs and Carlin (2007); Schuirmann (1987); Lakens (2017); Westlake (1976); Kirkwood (1981).

14.1.3 The probability of direction (PD)

The probability of direction is detailed in Makowski et al. (2019b) and varies between 50% and 100%. It is defined as the proportion of the posterior distribution of the parameter that is of the posterior median's sign:

$$PD := \int_A p(\theta|x)d\theta \tag{14.2}$$

In the above, $A := \{\theta \in \Theta : \text{sign}(\theta) = \text{sign}(\theta_{\text{MED}})\}$, θ_{MED} is the posterior median and sign denotes the sign function. As a consequence, if for example the posterior distribution assigns probability mass to both positive and negative parameter values, and the median is positive, it is the percentage of the posterior distributions probability mass located on the positive real numbers $(0, \infty)$ (analogue for dimensions larger than one).

14.1.4 The MAP-based p-value

The MAP-based p-value was proposed by Mills (2018), and can be related to the odds that a parameter has against the null hypothesis: It is defined as the ratio of the posterior density at the null value and the value of the posterior density at the maximum a posteriori (MAP) value, which is the equivalent of the mode for continuous probability distributions:

$$p_{\text{MAP}} := \frac{p(\theta_0|x)}{p(\theta_{\text{MAP}}|x)} \tag{14.3}$$

The rationale behind the MAP-based p-value is that whenever the value $p(\theta_0|x)$ is small compared to $p(\theta_{\text{MAP}}|x)$, the null hypothesis value θ_0 has a low posterior density value compared to the MAP-value, and thus $H_0 : \theta = \theta_0$ should be rejected.

14.1.5 The *e*-value and the Full Bayesian Significance Test (FBST)

The Full Bayesian Significance Test (FBST) was originally developed by Pereira and Stern Pereira and Stern (1999) and created under the assumption that a significance test of a sharp hypothesis had to be conducted. A sharp hypothesis refers to any submanifold of the parameter space of interest, see Pereira et al. (2008), which includes for example point hypotheses like $H_0 : \delta = 0$. Considering a standard parametric statistical model, where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a (vector) parameter of interest, $f(x|\theta)$ is the likelihood function associated to the observed data x, and $p(\theta)$ is the prior distribution of θ, the posterior distribution $p(\theta|x)$ is proportional to the product of the likelihood and prior density:

$$p(\theta|x) \propto f(x|\theta)p(\theta)$$

A hypothesis H makes the statement that the parameter θ lies in the corresponding null set Θ_H then. Following Pereira and Stern (2020) in notation, the Full Bayesian Significance Test (FBST) then defines two quantities: $\text{ev}(H)$, which is the e-value supporting (or in favour of) the hypothesis H, and $\overline{\text{ev}}(H)$, the e-value against H, also called the *Bayesian evidence value against* H, see Pereira and Stern Pereira and Stern (1999). First, the posterior *surprise function* $s(\theta)$ and its maximum s^* restricted to the null set Θ_H are denoted as

$$s(\theta) := \frac{p(\theta|x)}{r(\theta)}, \qquad s^* := s(\theta^*) = \sup_{\theta \in \Theta_H} s(\theta)$$

In the definition of the posterior surprise function $s(\theta)$, the denominator $r(\theta)$ is a reference density. If the improper flat prior $r(\theta) \propto 1$ is used, the surprise function becomes the posterior distribution $p(\theta|x)$. Otherwise, a noninformative prior distribution can be used as a reference density, see Pereira and Stern (2020). The next step towards the e-value is to define

$$T(\nu) := \{\theta \in \Theta | s(\theta) \leq \nu\}, \quad \overline{T}(\nu) := \Theta \setminus T(\nu)$$

and $\overline{T}(s^*)$ is then called the *tangential set to the hypothesis* H, which contains the points of the parameter space with higher surprise (relative to the reference density $r(\theta)$) than any point in the null set Θ_H. Integrating the posterior $p(\theta|x)$ over this set can be interpreted as the Bayesian evidence against H, the e-value $\overline{\text{ev}}(H)$:

$$\overline{\text{ev}}(H) := \overline{W}(s^*), \quad W(\nu) := \int_{T(\nu)} p(\theta|x) d\theta$$

In the above, $W(\nu)$ is called the cumulative surprise function, and $\overline{W}(\nu) := 1 - W(\nu)$. The e-value $\text{ev}(H)$ supporting H is obtained as $\text{ev}(H) := 1 - \overline{\text{ev}}(H)$. Therefore, large values of $\overline{\text{ev}}(H)$ indicate that the hypothesis H traverses low-density regions (or equivalently, that the alternative hypothesis traverses high-density regions) so that the *evidence against H is large*. The theoretical properties of the FBST and the e-value(s) have been detailed in Madruga et al. (2001, 2003), Stern (2003), Borges and Stern (2007), Pereira et al. (2008) and Pereira and Stern (2020). In this chapter, the focus is on the behaviour of the e-value $\overline{\text{ev}}(H)$ against $H : \delta = 0$ in the context of the Bayesian two-sample t-test. While one can use $\text{ev}(H)$ to reject H if $\text{ev}(H)$ is sufficiently small (or when $\overline{\text{ev}}(H)$ is large), it is not to confirm H via $\text{ev}(H)$, which may be seen as a drawback of the FBST. The reason is that $\text{ev}(H)$ is the posterior probability of parameters which attain a smaller or equal surprise than the null hypothesis value, and in the alternative hypothesis there may very well exist a parameter value θ' which even attains higher surprise. Thus, $\text{ev}(H)$ is no evidence against the alternative. There also exist asymptotic arguments via the distribution of $\text{ev}(H)$ which make it possible to obtain critical values based on this distribution to reject a hypothesis H, similar to p-values in NHST. Details are provided in Kelter and Stern (2020). However, in the simulation study below, no asymptotic arguments are used and solely the e-value $\overline{\text{ev}}(H)$ against H is reported.

Figures 14.1 and 14.2 show the different posterior Bayesian indices for significance and size of an effect for a Bayesian two-sample t-test. Group one was simulated as $\mathcal{N}(0.5, 1)$ and group two as $\mathcal{N}(2, 1)$ each with $n = 10$ samples and the true effect size is $\delta = -1.5$. The FBST is visualised in Figure 14.1, where the left plot shows a Cauchy prior $C(0, 1)$ (dashed line) and the resulting posterior $p(\delta|x)$ (solid black line), which is obtained by the Bayesian two-sample t-test of Rouder et al. (2009). s^* is computed as $s(0) = 0.1103$ (indicated by the blue point) and the integral $W(0)$ over the set $T(0)$ is shown as the red area under the posterior. This area is $\text{ev}(H)$, which is 0.0418 in this case. The blue area corresponds to the integral $\overline{W}(0)$ over the set $\overline{T}(0)$, which consists

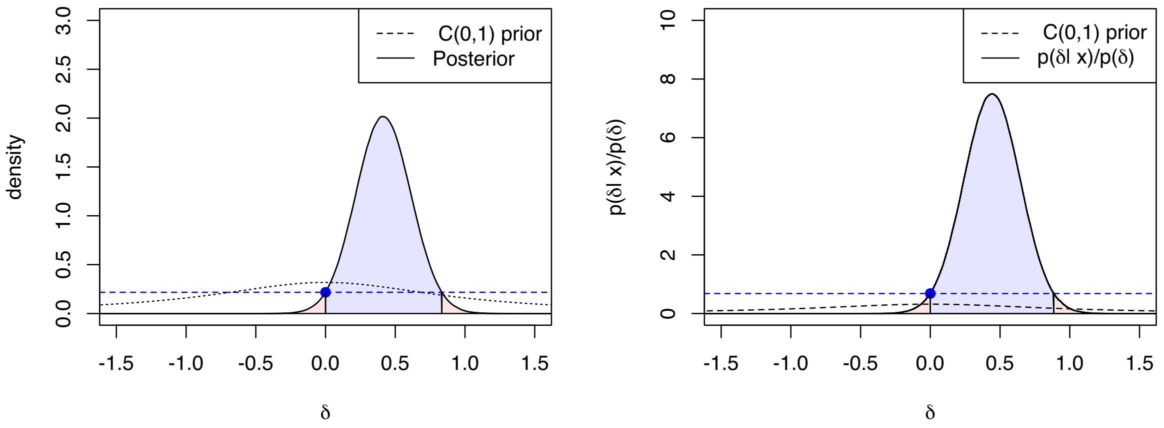

Figure 14.1: The e-value and FBST using a flat reference prior $r(\delta) \propto 1$ (left) and wide Cauchy reference prior $C(0,1)$ (right) against H_0 for the Bayesian two-sample t-test; the blue area indicates the integral over the tangential set $\overline{T}(0)$ against $H_0 : \delta = 0$, which is the e-value $\overline{\text{ev}}$ against H_0; the red area is the integral over $T(0)$, which is the e-value $\text{ev}(H)$ in favour of $H_0 : \delta = 0$

of all parameter values δ attaining a posterior density $p(\delta|x)$ larger than $p(0) = 0.1103$, indicated by the horizontal dashed blue line. The value of this integral is the evidence against $H_0 : \delta = 0$, $\overline{\text{ev}}(H) = 0.9582$, which advises the researcher to reject $H_0 : \delta = 0$ if a threshold of $\overline{\text{ev}}(H) > 0.95$ is used for making a decision in light of the obtained evidence. The right plot in Figure 14.1 shows the same situation, but now the reference function $r(\delta)$ used in the surprise function has been changed from the improper flat prior $r(\delta) \propto 1$ to the wide Cauchy prior $C(0,1)$ which is also used on the effect size model parameter in the Bayesian two-sample t-test of Rouder et al. (2009). Therefore, the surprise function values differ (see the scaling of the y-axis) and values of $p(\delta|x)/p(\delta) > 1$ indicate that the posterior $p(\delta|x)$ assigns a larger probability to a given parameter value than the prior $p(\delta)$. This can be interpreted as the data having increased this parameters probability.

The Bayes factor BF_{10} of $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ is shown in the upper left plot of Figure 14.2 and can be interpreted as the ratio of the prior density at the point-null value $\delta_0 = 0$ visualised as the grey lollipop and the posterior density at the point-null value $\delta_0 = 0$ visualised as the red lollipop.[1] After observing the data, H_0 becomes less probable, which is reflected in the Bayes factor of $BF_{10} = 3.38$. This magnitude indicates only moderate evidence for H_1, which is due to the small sample size of $n = 10$.

The MAP-based p-value is shown in the upper right plot and is defined as the ratio of the height of the posterior density at the null value $\delta_0 = 0$ and the MAP-value δ_{MAP}, the maximum a posteriori parameter. As can be seen, the MAP estimate is near $\delta = -1$,

[1] This general relationship was first discovered and proven by Dickey and Lientz (1970) and subsequently titled the Savage-Dickey density method (or ratio). Details are provided in Verdinelli and Wasserman (1995) and Wagenmakers et al. (2010) (for a simple proof of the representation of the Bayes factor see the appendix in Wagenmakers et al. (2010)), and the Savage-Dickey density ratio essentially makes it possible to obtain the Bayes factor as long as the posterior distribution can be obtained via some MCMC or HMC algorithm as shown in Chapter 13 for the parametric exponential survival model. An accessible proof of this relationship is given in the appendix of Wagenmakers et al. (2010).

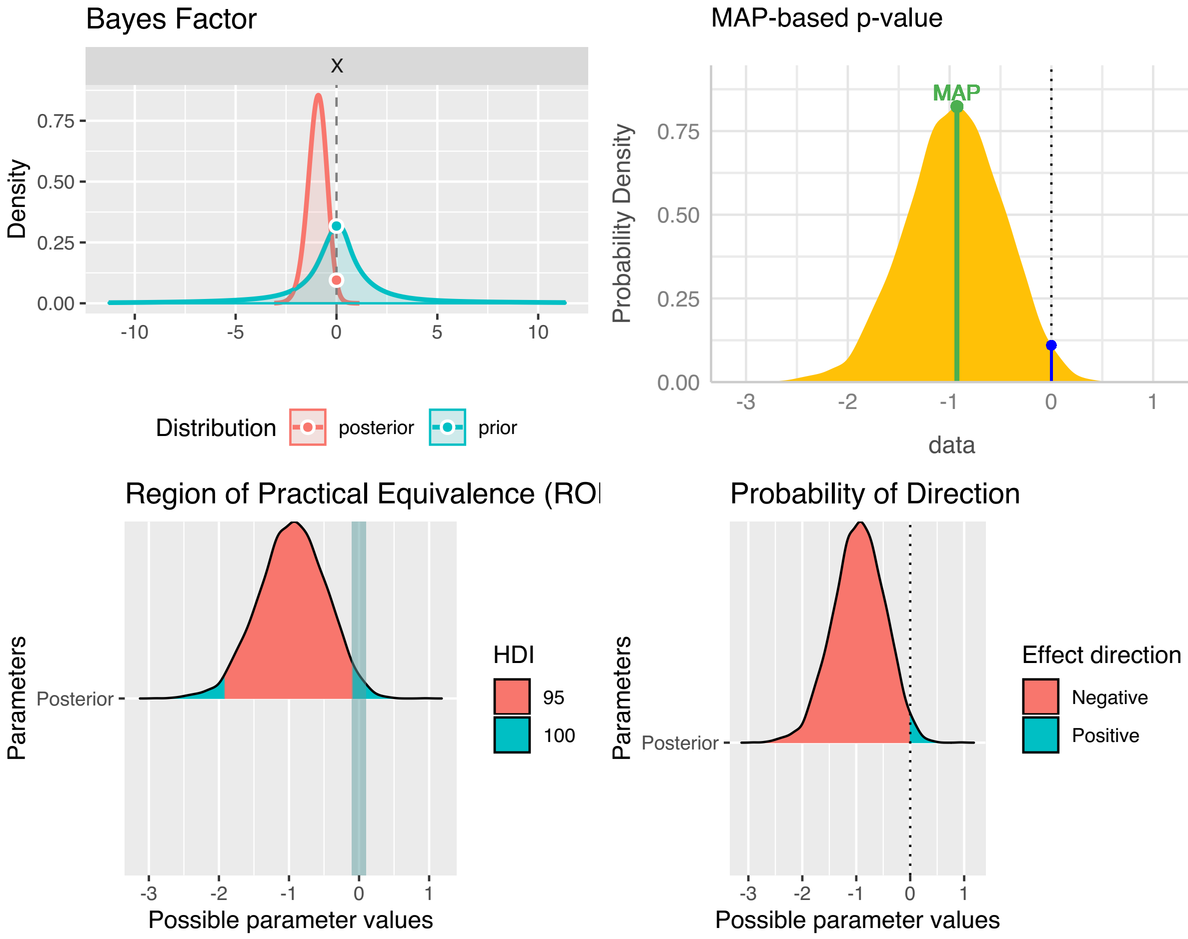

Figure 14.2: Different Bayesian posterior indices for significance and size of an effect for a Bayesian two-sample t-test

indicating a clear shift away from the null hypothesis. Still, the MAP-based p-value is given as $p_{\text{MAP}} = 0.203$, which is not significant when a threshold like 0.05 is used to declare significance.[2]

The lower left plot visualises the 95% and full ROPE, where the ROPE is defined as $[-0.1, 0.1]$, following the recommendations of Kruschke (2013). 2.38% probability mass of the posterior distribution is located inside the ROPE when using the 95% ROPE and 3.00% is located inside the ROPE when using the full ROPE. In a test of practical equivalence, where the null is only rejected if the posterior is located entirely outside the ROPE, the null hypothesis H_0 cannot be rejected based on the ROPE. Still, if an estimation-oriented perspective is used, avoiding the classical testing stance, the ROPE-analysis shows evidence for the alternative H_1 for both the 95% and full ROPE.

The lower right plot in Figure 14.2 shows the probability of direction (PD). It enjoys some desirable properties: First, it clearly shows that the effect is more likely to be of negative than positive sign, as 97.70% of the posterior is located on the negative real numbers. Also, the PD embraces estimation under uncertainty instead of hypothesis testing, in the same way as the ROPE does when avoiding an explicit testing stance.

[2] Note that the MAP-based p-value has no connection to traditional test levels like the ones used in the Neyman-Pearson theory, compare Chapter 4. As a consequence, no decision-theoretic optimality can be associated with using a significance threshold for the MAP-based p-value.

The posterior distribution can then be used in a second step to obtain, for example, the mean and standard deviation as estimates for the parameter. Still, hypothesis testing is also possible via rejecting the null $H_0 : \delta \geq 0$ if at least 95% of the posterior of δ is located on the negative real axis.

14.2 Methods

A simulation study was performed to analyse the behaviour of the different Bayesian evidence measures for hypothesis testing in the setting of Welch's two-sample t-test (Rüschendorf, 2014). Pairs of data were simulated, consisting of two samples, one for each group, each normally distributed. Four settings were selected: In the first, no effect is present, and both groups are identically distributed as standard normal $\mathcal{N}(0, 1)$. In the second, a small effect is present, and the first group is simulated as $\mathcal{N}(2.89, 1.84)$ and the second as $\mathcal{N}(3.5, 1.56)$, resulting in a true effect size of

$$\delta = \frac{(2.89 - 3.5)}{\sqrt{((1.84^2 + 1.56^2)/2)}} \approx -0.357 \tag{14.4}$$

In the third simulation setting, a medium effect is present. The first group is simulated as $\mathcal{N}(254.08, 2.36)$ and the second as $\mathcal{N}(255.84, 3.04)$, resulting in a true effect size of

$$\delta = \frac{(254.08 - 255.84)}{\sqrt{((2.36^2 + 3.04^2)/2)}} \approx -0.646 \tag{14.5}$$

The last setting uses $\mathcal{N}(15.01, 3.4)$ and $\mathcal{N}(19.91, 5.8)$ distributions for the first and second group, yielding a true effect size of

$$\delta = \frac{(15.01 - 19.91)}{\sqrt{((3.4^2 + 5.8^2)/2)}} \approx -1.03 \tag{14.6}$$

For each of the four effect size settings, 10000 datasets following the corresponding group distributions as detailed above were simulated. This procedure was repeated for different samples sizes n, ranging from $n = 10$ to $n = 100$ in steps of size 10 to investigate the influence of sample size on the indices. In each case, the traditional p-value, the Bayes factor BF_{10}, the ROPE 95%, the full ROPE, the probability of direction, the MAP-based p-value and the e-value $\overline{\text{ev}}(H_0)$, that is the evidence against $H_0 : \delta = 0$ were computed. The Bayes factor was calculated as the Jeffreys-Zellner-Siow Bayes factor for the null hypothesis $H_0 : \delta = 0$ of no effect against the alternative $H_1 : \delta \neq 0$, see Rouder et al. (2009) and Gronau et al. (2020). More precisely, the calculated quantities are (1) the Bayes factor, a single number that quantifies the evidence for the presence or absence of an effect and (2) the posterior distribution, which quantifies the uncertainty about the size of the effect under the assumption $H_1 : \delta \neq 0$ that it exists. This posterior distribution (2) of the effect size δ was then used to compute the 95% ROPE, the full ROPE, the PD and the MAP-based p-value as well as the e-value $\overline{\text{ev}}(H_0)$. The traditional p-value was obtained via a two-sample Welch's t-test with test level $\alpha = 0.05$.

The above procedure was conducted three times with the prior on the effect size δ set to three different hyperparameters to investigate the influence of the prior modelling: A noninformative Jeffrey's prior was always put on the standard deviation of the normal population, while a Cauchy prior was placed on the standardised effect size. The Cauchy prior $C(0, \sqrt{2}/2)$ was used in the first setting, $C(0, 1)$ in the second and $C(0, \sqrt{2})$ in the third, corresponding to a medium, wide and ultrawide prior on the effect size δ. This way, the influence of the prior modelling on the resulting indices can be measured. To get more insights about the e-value $\overline{\text{ev}}(H_0)$, for each prior setting $\overline{\text{ev}}(H_0)$ was once computed using a flat improper reference function $r(\delta) \propto 1$ (that is, the surprise function equals the posterior distribution), and once using the Cauchy prior assigned to δ as a reference density in the surprise function $s(\delta)$.

Finally, the above procedure was repeated for fixed sample size to investigate the influence of noise. Thus, $n = 30$ samples were simulated in each group to control for the influence of sample size and Gaussian noise $\mathcal{N}(0, \varepsilon)$ was added to the group data x and y, where ε varies from $\varepsilon = 0.5$ to $\varepsilon = 5$ in steps of 0.5.

The percentage of significant results was computed for samples of increasing size n as the number of significant results divided by 10000. This number is a Monte Carlo estimate for the type I error probabilities of the indices, a crucial quantity for reproducible research (McElreath and Smaldino, 2015). Significant is defined here as a Bayes factor $BF_{10} \geq 3$. A posterior distribution using the 95% ROPE or full ROPE is significant when it is located completely outside the corresponding ROPE $[-0.1, 0.1]$ around $\delta = 0$. The MAP-based p-value is significant when $p_{MAP} < 0.05$. The p-value is significant when $p < 0.05$. The PD is significant when $PD = 1$ or $PD = 0$, and the e-value is significant when $\overline{\text{ev}}(H) > 0.95$ (no matter whether a flat reference density or the Cauchy reference density was used).

The statistical programming language R was used (**?**) for the simulations. The Bayes factor was computed via Gaussian quadrature in the `BayesFactor` R package (Morey and Rouder, 2018), which was also used to obtain the posterior distribution of δ under the alternative H_1 of an existing effect. The package `bayestestR` (Makowski et al., 2019a) was used to compute the 95% ROPE, full ROPE, PD and MAP-based p-value. The evidence $\overline{\text{ev}}$ against $H_0 : \delta = 0$ in the FBST was computed with the posterior Markov-Chain-Monte-Carlo draws of the posterior distribution of δ provided by the `BayesFactor` package (Morey and Rouder, 2018). These posterior draws were interpolated to construct a posterior density of δ, which was then integrated numerically over the tangential set to H_0 as required for $\overline{\text{ev}}(H_0)$.

14.3 Results

14.3.1 Influence of sample size and prior modelling

Figure 14.3 shows the dependence of the Bayesian indices on sample size for four different effect sizes using the ultrawide prior $C(0, \sqrt{2})$. The four plots in each row show the succession of the results for no effect, a small effect, a medium effect and finally a large effect, while the x-axis shows increasing sample size $n = 10$ to $n = 100$ in each group in steps of 10. The left plot of the first row shows that the p-value is distributed uniformly under the null hypothesis $H_0 : \delta = 0$. If the alternative $H_1 : \delta \neq 0$ is true, the three plots right beneath show that for increasing sample size n, the p-value becomes significant, where the necessary sample size for stating significance decreases with increasing actual effect size δ. The second row shows the succession for the Bayes factor BF_{10}. The left plot indicates that under the null hypothesis $H_0 : \delta = 0$ the Bayes

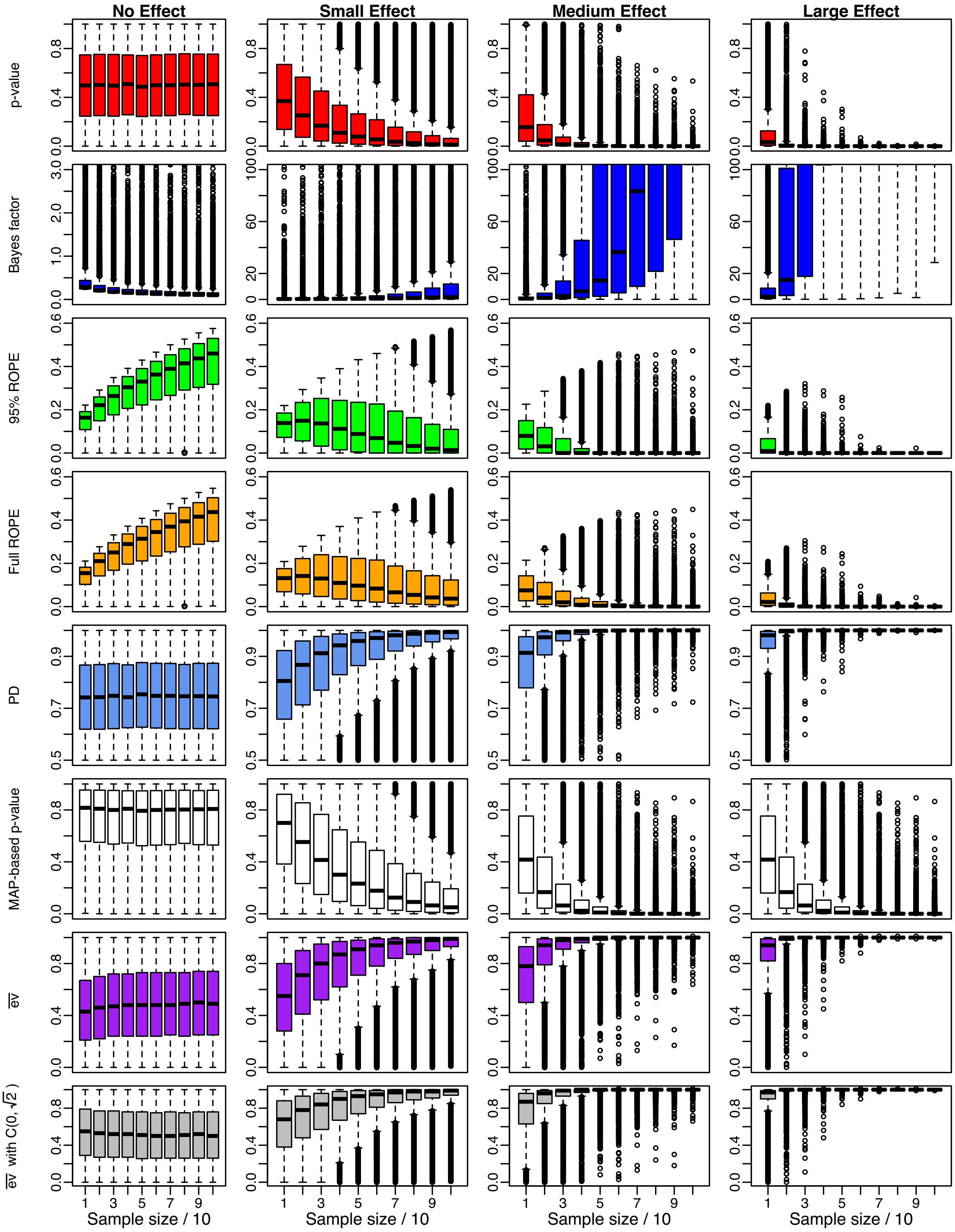

Figure 14.3: Influence of the sample size n on Bayesian effect significance and size indices for small, medium, large and no existing effect using an ultrawide prior $C(0, \sqrt{2})$ on the effect size δ

factor correctly converges to zero (in contrast to the p-value). This property opens the possibility of confirming the null hypothesis, which is not possible via a p-value. The three figures right beneath this plot show the progression of the Bayes factor BF_{10} for increasing effect size: The Bayes factor accumulates more and more evidence for the alternative $H_1 : \delta \neq 0$ for small, medium and large effect sizes. For more substantial effect sizes, the Bayes factor requires a much smaller sample size to state evidence for the alternative. The plots are limited to a y-range of $[0, 100]$ (except for the first plot) for better visibility, as BF_{10} becomes very large quickly.

The third and fourth row shows the results for the 95% and full ROPE $[-0.1, 0.1]$ around the effect size $\delta = 0$. Under the null, in both cases, the percentage of the posterior's probability mass inside the ROPE increases. As $\delta = 0$ under the null, for $n \to \infty$, the posterior will eventually concentrate completely inside the ROPE, but the necessary sample size can be substantial. For $n = 100$, about 50% of the probability mass of the posterior is located inside the ROPE $[-0.1, 0.1]$ around $\delta = 0$. For increasing sample size n, this percentage will eventually attain 100%. Considering the 95% and full ROPE, even for small sample sizes like $n = 10$ the majority of values shows that at least 10% of the posterior is located inside the ROPE so that hardly any false-positive statements are produced.

Under the alternative $H_1 : \delta \neq 0$, both the 95% and full ROPE show that the percentage of the posterior located inside the ROPE $[-0.1, 0.1]$ of no effect converges to zero for increasing sample size n. For increasing effect size δ, the necessary sample size n needed to reject the null hypothesis H_0 becomes smaller.

The fifth row shows the results for the probability of direction (PD). Under the null hypothesis $H_0 : \delta = 0$, the PD is not uniformly distributed as was the case for p-values. The PD concentrates at about 70% here (see the scaling of the y-axis), which does not reflect the true effect size of $\delta = 0$, which should yield a PD near 50%. Still, under the alternative $H_1 : \delta \neq 0$, the PD converges to 100% if sample sizes grow. The speed of convergence is faster for larger effect sizes $\delta \neq 0$.

The MAP-based p-value shown in the sixth row shows a behaviour similar to the classic p-value. One difference is that under the null hypothesis H_0, it is much larger on average than the traditional p-value. Still, this behaviour is robust to increasing sample size n and as correct interpretation of the MAP-based p-value only allows to state significance when p_{MAP} is smaller than a significance threshold. Interpreting large p_{MAP} as evidence for H_0 is not allowed at all. Under the alternative H_1, the behaviour is quite similar to the classic p-value: For increasing sample size n, the MAP-based p-value becomes significant, where the necessary sample size n for stating significance decreases with increasing effect size δ.

The evidence $\overline{\text{ev}}(H_0)$ (in the following denoted as $\overline{\text{ev}}$) under the flat improper reference density $r(\delta) \propto 1$ is shown in the seventh row and concentrates around $\delta = 0.5$ under the null hypothesis $H_0 : \delta = 0$. The reason for this can be seen in the fact that the posterior of δ concentrates for $n \to \infty$ around $\delta = 0$ if $H_0 : \delta = 0$ is true, and the posterior density $p(\delta|x)$ also concentrates around $\delta = 0$ with slight fluctuations happening due to the randomness in simulation. However, as the FBST measures the ratio of posterior mass inside and outside the tangential set, this ratio jitters for increasing sample size between zero and one. The only thing that changes when increasing sample size n is thus the concentration of the posterior $p(\delta|x)$ around the null value, so that $\overline{\text{ev}}$ is not influenced into either direction by increasing sample size. From a measure-theoretic perspective, any point null value has zero prior probability mass under a prior which is absolutely continuous with respect to the Lebesgue measure. As a consequence, the posterior probability of any point null value θ_0 will be zero, too (Schervish, 1995; Robert, 2007). Thus, the FBST cannot accept a point null hypothesis primarily due to measure-theoretic reasons, and the Bayes factor achieves confirmation of a point null value only via the use of the mixture prior structure which was detailed in Chapter 7. For the FBST,

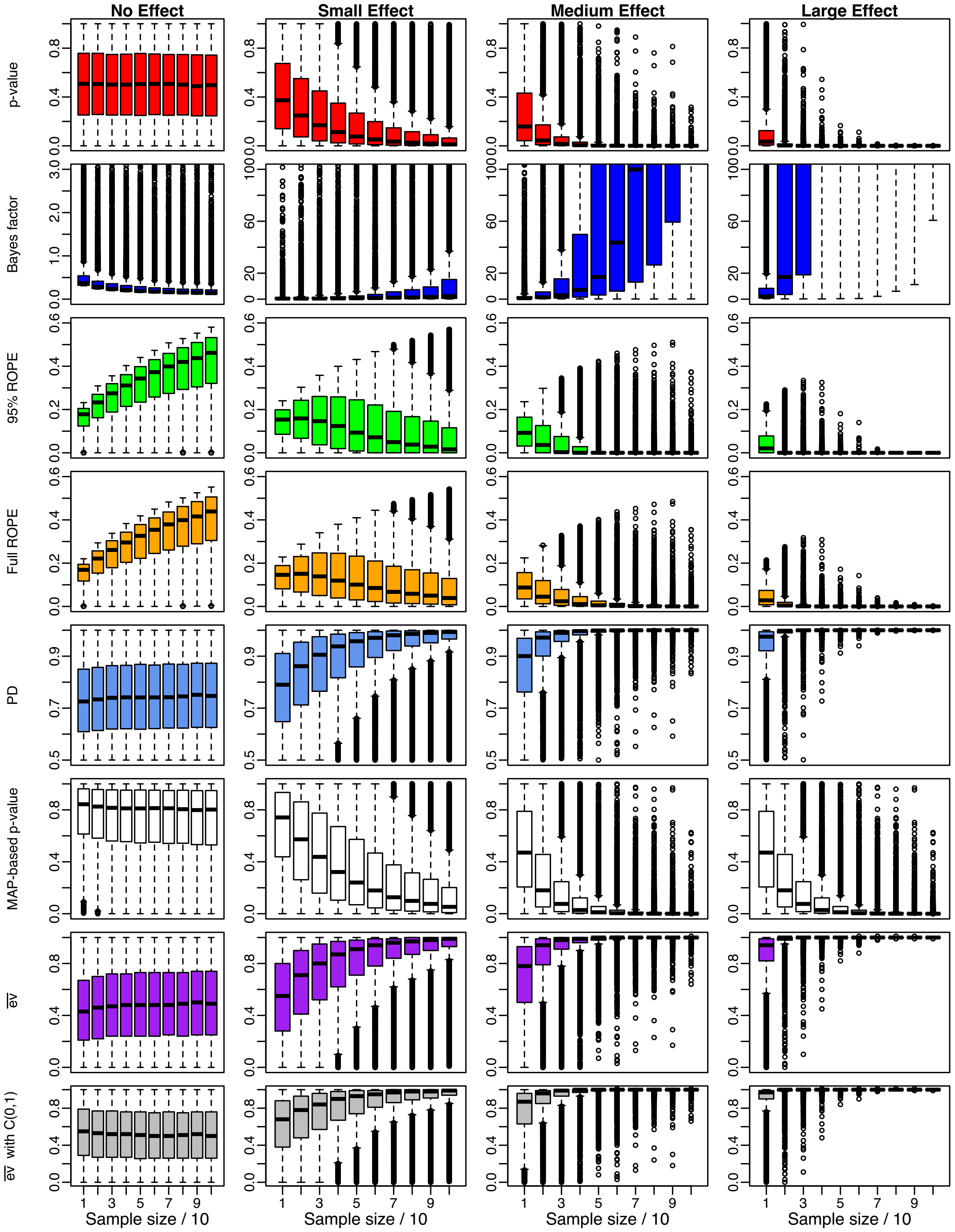

Figure 14.4: Influence of the sample size n on Bayesian effect significance and size indices for small, medium, large and no existing effect using a wide prior $C(0,1)$ on the effect size δ

the support for H_0 can easily be obtained by calculating $\text{ev}(H_0) = 1 - \overline{\text{ev}}(H_0)$, which in this case also concentrates around 0.5, instead of concentrating around 1. If on the other hand $H_1 : \delta \neq 0$ is true, $\overline{\text{ev}}$ quickly signals evidence against H_0 for increasing sample size n and increasing effect size δ, as shown by the three right-hand plots in the seventh row. When using the medium Cauchy prior $C(0, \sqrt{2}/2)$ instead of the improper reference density $r(\delta) \propto 1$, the situation is similar, but the plots in the last row in Figure 14.5 show that the evidence $\overline{\text{ev}}$ against H_0 accumulate faster then if H_1 is true.

Figure 14.4 shows the results of the simulation when using a wide prior $C(0, 1)$ instead of the ultrawide prior $C(0, \sqrt{2})$. The classic p-value is of course not affected at all from this prior change. The BF_{10} shown in the second row is slightly larger under the alternative $H_1 : \delta \neq 0$, as the wide prior $C(0, 1)$ becomes more informative compared to the ultrawide prior $C(0, \sqrt{2})$. The probability mass located around $\delta = 0$ becomes more concentrated when using the wide $C(0, 1)$ prior instead of the ultrawide $C(0, \sqrt{2})$ prior, and therefore BF_{10} for small and medium effects is increased (compare the boxplots in the second and third column in Figures 14.3 and 14.4), while for large effects the influence is less apparent (see the fourth column in Figures 14.3 and 14.4).

For the same reasons, the percentage of probability mass inside the 95% and full ROPE increases under the null $H_0 : \delta = 0$, as shown by the third and fourth row in Figure 14.4. More prior mass around $\delta = 0$ due to the narrower $C(0, 1)$ prior on δ leads to more posterior mass inside the ROPE $[-0.1, 0.1]$ around $\delta = 0$. Under the alternative H_1, the 95% and full ROPE suffer from this change, as shown in the boxplots for small, medium and large effects in rows three and four, which are shifted up slightly. The increase of probability mass near $\delta = 0$ draws the posterior towards $\delta = 0$, and it becomes harder for the posterior to concentrate outside of the ROPE. Nevertheless, for increasing sample size, the ROPEs finally reveal evidence for the alternative H_1. Note that due to the concentration of probability mass around zero when using the $C(0, 1)$ prior, the boxplots of the ROPEs are shifted slightly up under the null hypothesis of no effect.

The same holds for the PD, which also needs a larger sample size now to achieve the same evidence for the alternative when an effect is present. No matter whether a small, medium or large effect size is present, all boxplots shift down slightly, indicating that less probability mass is strictly positive in the posteriors produced. The narrower prior distribution shrinks the complete posterior distribution towards smaller values, leading in turn to a smaller PD.

The MAP-based p-value is also influenced by the narrower prior: Due to the increased probability mass near $\delta = 0$, the MAP-estimate of δ shrinks towards $\delta = 0$. The ratio of the posterior density value $p(\delta_0|x)$ at the point-null value $\delta_0 = 0$ and the posterior density value $p(\delta_{\text{MAP}}|x)$ at the MAP-value thus gets closer to one compared to the ultrawide prior setting. This leads to a larger MAP-based p-values and slightly upshifted boxplots under the alternative H_1.

The last two rows show $\overline{\text{ev}}$ under the improper reference density $r(\delta) \propto 1$. Barely any change can be observed compared to the setting using the ultrawide prior $C(0, \sqrt{2})$, which is confirmed in the seventh row. Under the wide Cauchy prior reference density $r(\delta) = C(0, 1)$, the evidence against $H_0 : \delta = 0$ again concentrates around $\overline{\text{ev}} = 0.5$, indicating neither strong evidence against H_0 nor support for H_0. Compared to the ultrawide prior used in Figure 14.3, under the alternative $H_1 : \delta \neq 0$ the evidence $\overline{\text{ev}}$ against $H_0 : \delta = 0$ also barely changes. These results show that the e-value is quite robust against variations both in the reference density and prior selection.

Figure 14.5 shows the results when using a medium prior instead of a wide one. The classic p-value is again not affected from this prior, so the results are identical. In contrast to Figures 14.3 and 14.4, the Bayes factor now accumulates evidence even faster,

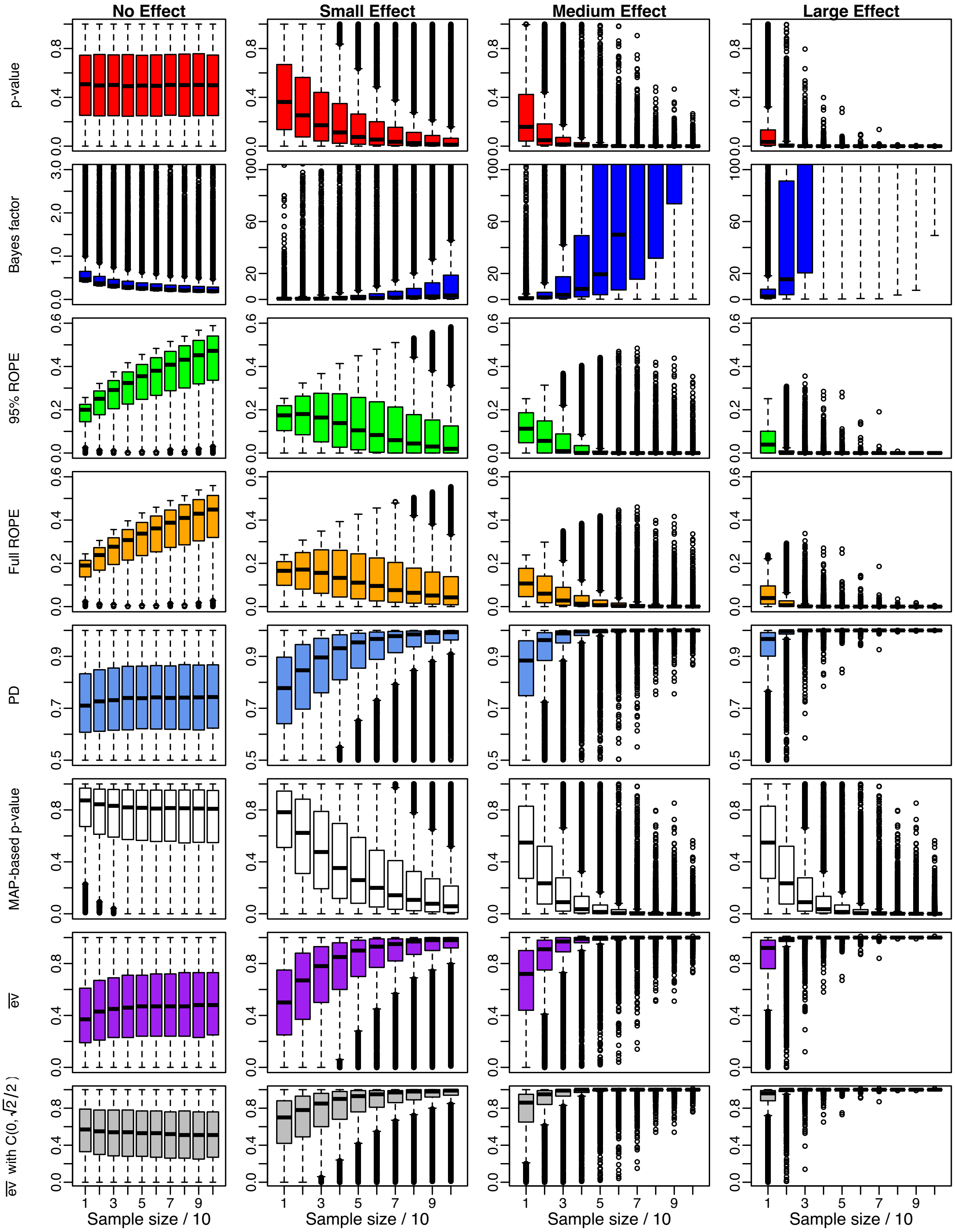

Figure 14.5: Influence of the sample size n on Bayesian effect significance and size indices for small, medium, large and no existing effect using a medium prior $C(0, \sqrt{2}/2)$ on the effect size δ

because the medium prior is even more informative than the wide and ultrawide one.

The 95% and full ROPE boxplots are shifted up even higher therefore under H_0, showing that switching from the noninformative ultrawide and weakly informative wide prior to the medium prior yields larger percentages of the posterior distributions probability mass inside the ROPE under the null hypothesis H_0 as even more probability mass concentrates around $\delta_0 = 0$ now. From a Bayesian perspective, the null hypothesis is thus faster confirmed. Under the alternative $H_1 : \delta \neq 0$, the medium prior makes it now even harder for the 95% and full ROPE to reject the null hypothesis. This is again due to the fact that under the medium prior $C(0, \sqrt{2}/2)$ the prior allocates even more probability mass to values near $\delta_0 = 0$ than under the wide $C(0, 1)$ or ultrawide Cauchy prior $C(0, \sqrt{2})$. Therefore, the posterior shifts more slowly away from the ROPE $[-0.1, 0.1]$ of no effect, and for the same sample size n, the probability mass located inside the ROPE is larger when using the medium prior on δ. Still, for increasing sample size, this effect vanishes and even under the medium prior, the concentration of posterior mass inside the ROPE converges to zero.

The same phenomenon holds for the PD and the MAP-based p-value. Here too, under the alternative the narrower prior on δ around zero makes it harder for the PD and MAP-based p-value to accumulate evidence for the alternative H_1. For increasing sample size n, both the PD and the MAP-based p-value eventually reject the null hypothesis. For a fixed sample size n, the same is achieved faster under the wide and ultrawide prior, which distribute less prior probability mass near $\delta_0 = 0$.

Considering $\overline{\text{ev}}$ in the last two rows, under the improper reference density $r(\delta) \propto 1$ again barely any changes can be observed compared to the setting in which the wide $C(0, 1)$ or ultrawide $C(0, \sqrt{2})$ prior were used, which is confirmed in the seventh row of Figure 14.5. Under the medium Cauchy prior reference density $r(\delta) = C(0, \sqrt{2}/2)$, the evidence against $H_0 : \delta = 0$ again concentrates around $\overline{\text{ev}} = 0.5$, indicating neither strong evidence against H_0 nor support for H_0. Compared to the ultrawide and wide priors used in Figures 14.3 and 14.4, under the alternative $H_1 : \delta \neq 0$ the evidence $\overline{\text{ev}}$ against $H_0 : \delta = 0$ again is barely influenced by shifting to the medium Cauchy prior, showing strong robustness of the e-value against the prior modelling.

At this point, the results show that both the MAP-based p-value, the classic p-value and the e-value $\overline{\text{ev}}$ cannot state evidence for the null hypothesis. These Bayesian evidence measures can only reject the null hypothesis H_0 and offer no possibility to confirm it. For practical research, this is limiting. Also, the PD stabilises around 75%, which is the middle of its possible extremes, 50% and 100%. It would be desirable that the PD converges to 50% under the null $H_0 : \delta = 0$, to show that both a positive and negative effect are equally possible. Given the behaviour of the PD under the null, it seems that the PD favours the directed alternative $\delta > 0$ although the null $H_0 : \delta = 0$ is true. Under the alternative, $H_1 : \delta \neq 0$, the PD as well as the p-value and MAP-based p-value behave as expected. Note that Pereira and Stern (1999) created the e-value to test a sharp hypothesis H_0, and rejection of H_0 was the intended goal of the procedure. In fact, the FBST does not make the (unrealistic) prior assumption of assigning a mixture prior in Jeffreys' sense to the parameter. Thus, the inability of the FBST to confirm a point null hypothesis can be seen as the price for this measure-theoretic coherence. In contrast to the p-value and MAP-based p-value, the e-value enjoys a multitude of highly desirable properties like compliance with the likelihood principle, being a probability value derived from the posterior distribution, and being invariant to alternative parameterisations, see also Pereira et al. (2008). Therefore, the e-value is preferable over the standard p-value and MAP-based p-value, also because of its robustness to the prior selection.

The Bayes factor BF_{10}, the 95% and full ROPE have two desirable properties: Under the null, all three measures indicate evidence for $H_0 : \delta = 0$ while under the alternative $H_1 : \delta \neq 0$, they indicate evidence for H_1.[3] It is somehow problematic while not astonishing that both constructs accumulate evidence faster under the null H_0 using a medium prior, than when using a wide or ultrawide prior. Under the alternative, evidence for H_1 accumulates faster when using a wide or ultrawide prior instead of a medium one. Thus, when using a medium prior, finding evidence for H_0 is easier than finding evidence for H_1 both with the BF and the ROPEs Using a wide or ultrawide prior, finding evidence for H_1 is easier (when H_1 postulates a sufficiently large effect) than finding evidence for H_0 with the BF and the ROPEs. Therefore, it is recommended to use the wide prior $C(0, 1)$, which places itself in the middle between these two extremes. Using a medium or ultrawide prior needs further justification, because otherwise, some kind of cherry-picking could happen by combining Bayes factors or ROPEs with a medium, wide or ultrawide prior depending on the goal of rejection or confirmation of the null hypothesis after the data have been observed. The e-value showed strong robustness to the prior selection. Therefore, if the rejection of a research hypothesis is the formulated goal of the scientific enterprise, the e-value based on the FBST procedure with the corresponding Cauchy prior as reference density in the surprise function may prevent such cherry-picking.

In summary, the combination of prior and significance and effect size measure together can make it easier to find evidence for some hypotheses, which underlines the importance of robustness checks as shown in Chapter 12. Also, taking into account that the focus of a variety of research is to reveal relevant differences (clinically, in biomedical research for example), it is recommended to use at least $n = 100$ participants in each group to ensure that also small effects can be detected reliably.

14.3.2 Influence of noise

Figure 14.6 shows the results for the influence of noise on Bayesian indices of significance and effect size. As expected and shown in the first row, the influence of noise on the classic p-value under the null H_0 is negligible. Under the alternative, the p-value gets disturbed more and more with increasing noise ε. The number of significant p-values reduces for increasing noise as shown by the boxplots, which are shifted upwards more and more when noise ε increases.

The BF_{10} has the same problems: When the null hypothesis $H_0 : \delta = 0$ is true, the Bayes factor is not influenced much by noise. When on the other hand $H_1 : \delta \neq 0$ is true, adding noise to the observations makes it more difficult for the Bayes factor to state evidence for the alternative $H_1 : \delta \neq 0$. This behaviour is also revealed when comparing Figure 14.3 and Figure 14.6: The boxplots in the fourth plot of the second row in Figure 14.3 show that the Bayes factor achieves higher values compared to the situation where noise is present, as shown in the fourth plot of the second row in Figure 14.6.

The 95% ROPE and full ROPE also suffer from increasing noise. Under the null hypothesis, the noise the noise does not influence the percentage of posterior mass inside the ROPE, but under the alternative H_1 increasing noise ε causes increasing amounts

[3] As noted above, the price paid by the Bayes factor for this ability is the introduction of a mixture prior as introduced first by Jeffreys and Haldane, see Part II. The assignment of positive probability to a Lebesgue-null-set is not without problems, as discussed in further detail in Rao and Lovric (2016), Sawilowsky (2016) and Zumbo and Kroc (2016), the arguments of which go back at least to Hodges and Lehmann (1954).

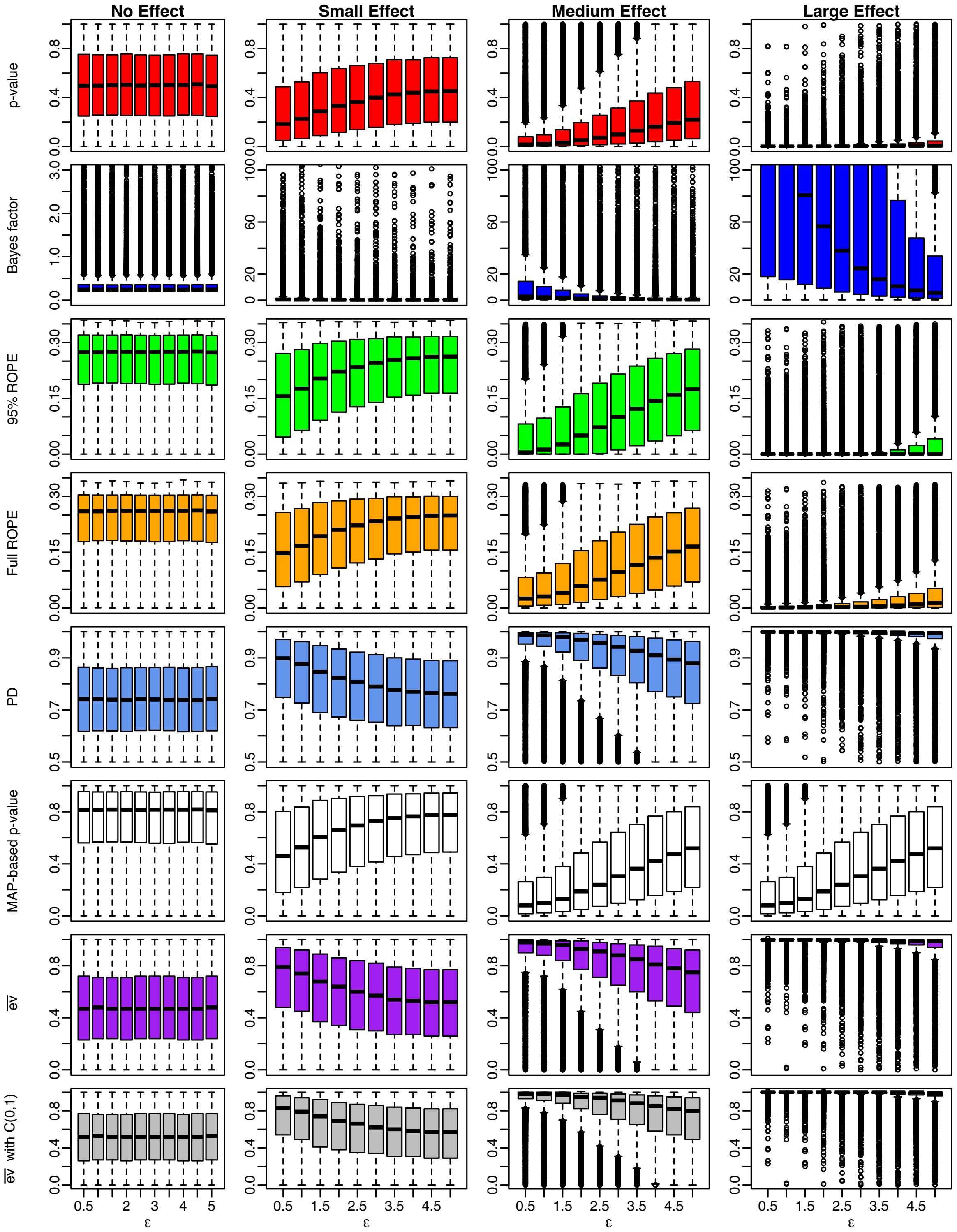

Figure 14.6: Influence of noise ε on Bayesian significance and effect size indices for small, medium, large and no existing effects using an ultrawide prior $C(0, \sqrt{2})$ on the effect size δ and sample size $n = 30$ in each groups

of posterior mass to be located inside the ROPE. This behaviour makes it harder for the ROPE to signal evidence for the alternative $H_1 : \delta \neq 0$.

The PD suffers from the same problem, as increasing noise causes the posterior to be more and more symmetric around $\delta_0 = 0$, indicated by the boxplots successively shifted down for increasing noise under H_1.

The MAP-based p-value is also not influenced by noise under the null hypothesis H_0, but the boxplots are shifted up under the alternative, indicating that increasing noise leads to larger p-values and less significant ones, which makes it harder for the MAP-based p-value to reject the null hypothesis in the presence of noise.

The *e*-value $\overline{\text{ev}}$ is also barely influenced by noise under the null hypothesis H_0 both when used in combination with the flat reference density $r(\delta) \propto 1$ and the wide Cauchy reference density $r(\delta) = C(0,1)$. Under the alternative, increasing noise makes it harder for $\overline{\text{ev}}$ to state evidence against H_0 as shown in the last two rows of Figure 14.6.

14.3.3 Sensitivity and type I error rates

Type I error rates and sensitivity of Bayesian posterior indices

Index	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	$n = 60$	$n = 70$	$n = 80$	$n = 90$	$n = 100$
No Effect										
p-value	0.0483	0.0500	0.0552	0.0508	0.0507	0.0500	0.0491	0.0499	0.0520	0.0529
BF_{10}	0.0221	0.0175	0.0192	0.0124	0.0137	0.0120	0.0104	0.0100	0.0100	0.0094
95% ROPE	0.0145	0.0159	0.0172	0.0127	0.0130	0.0107	0.0088	0.0083	0.0085	0.0069
Full ROPE	0.0002	0.0001	0.0000	0.0001	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000
PD	0.0003	0.0003	0.0000	0.0004	0.0004	0.0006	0.0003	0.0003	0.0004	0.0002
MAP-p-value	0.0060	0.0075	0.0118	0.0096	0.0120	0.0111	0.0107	0.0107	0.0121	0.0117
$\overline{\text{ev}}$	0.0225	0.0273	0.0311	0.0342	0.0362	0.0383	0.0404	0.0386	0.0393	0.0391
$\overline{\text{ev}}$ with $C(0,1)$	0.0490	0.0470	0.0477	0.0459	0.0480	0.0471	0.0487	0.0458	0.0481	0.0474
Small Effect										
p-value	0.1081	0.1990	0.2807	0.3457	0.4224	0.4890	0.5534	0.6149	0.6655	0.7092
BF_{10}	0.0559	0.1045	0.1490	0.1835	0.2319	0.2682	0.3221	0.3648	0.4150	0.4562
95% ROPE	0.0433	0.0945	0.1423	0.1752	0.2238	0.2526	0.3014	0.3374	0.3831	0.4165
Full ROPE	0.0005	0.0012	0.0024	0.0047	0.0061	0.0107	0.0139	0.0186	0.0235	0.0289
PD	0.0010	0.0034	0.0090	0.0144	0.0265	0.0333	0.0538	0.0747	0.0953	0.1175
MAP-p-value	0.0222	0.0590	0.1082	0.1539	0.2137	0.2593	0.3219	0.3746	0.4369	0.4878
$\overline{\text{ev}}$	0.0671	0.1417	0.2252	0.2976	0.3720	0.4415	0.5171	0.5659	0.6175	0.6755
$\overline{\text{ev}}$ with $C(0,1)$	0.1164	0.1972	0.2763	0.3436	0.4180	0.4835	0.5527	0.5976	0.6459	0.7018
Medium Effect										
p-value	0.2762	0.5149	0.6930	0.8193	0.8899	0.9417	0.9717	0.9831	0.9907	0.9951
BF_{10}	0.1709	0.3443	0.5013	0.6519	0.7439	0.8342	0.8928	0.9269	0.9561	0.9741
95% ROPE	0.1392	0.3247	0.4850	0.6389	0.7303	0.8197	0.8779	0.9165	0.9464	0.9685
Full ROPE	0.0017	0.0170	0.0382	0.0752	0.1282	0.1944	0.2769	0.3504	0.4386	0.5050
PD	0.0044	0.0320	0.0801	0.1635	0.2620	0.3830	0.4986	0.6010	0.6878	0.7606
MAP-p-value	0.0694	0.2431	0.4249	0.6039	0.7196	0.8256	0.8930	0.9317	0.9605	0.9779
$\overline{\text{ev}}$	0.1779	0.4373	0.6244	0.7698	0.8714	0.9256	0.9584	0.9752	0.9882	0.9951
$\overline{\text{ev}}$ with $C(0,1)$	0.2773	0.5227	0.6880	0.8083	0.8953	0.9376	0.9663	0.9807	0.9908	0.9960
Large Effect										
p-value	0.5824	0.8814	0.9746	0.9955	0.9987	1.0000	0.9999	1.0000	1.0000	1.0000
BF_{10}	0.4438	0.7776	0.9254	0.9801	0.9937	0.9986	0.9999	0.9999	1.0000	1.0000
95% ROPE	0.3844	0.7584	0.9185	0.9787	0.9928	0.9984	0.9997	0.9999	1.0000	1.0000
Full ROPE	0.0182	0.1252	0.3133	0.5407	0.7192	0.8535	0.9259	0.9664	0.9851	0.9929
PD	0.0268	0.2052	0.4704	0.7217	0.8597	0.9450	0.9795	0.9933	0.9969	0.9997
MAP-p-value	0.0694	0.2431	0.4249	0.6039	0.7196	0.8256	0.8930	0.9317	0.9605	0.9779
$\overline{\text{ev}}$	0.4486	0.8367	0.9597	0.9927	0.9990	0.9996	1.0000	1.0000	1.0000	1.0000
$\overline{\text{ev}}$ with $C(0,1)$	0.5800	0.8862	0.9743	0.9945	0.9992	0.9998	1.0000	1.0000	1.0000	1.0000

Table 14.1: Percentage of significant Bayesian indices of effect significance and magnitude for varying sample size

Table 14.1 shows Monte Carlo estimates for the type I error rates and the percentage of significant indices based on the results of the simulations. For increasing sample size

n, the type I error rates were estimated as the number of significant indices divided by 10000 when no effect was present.

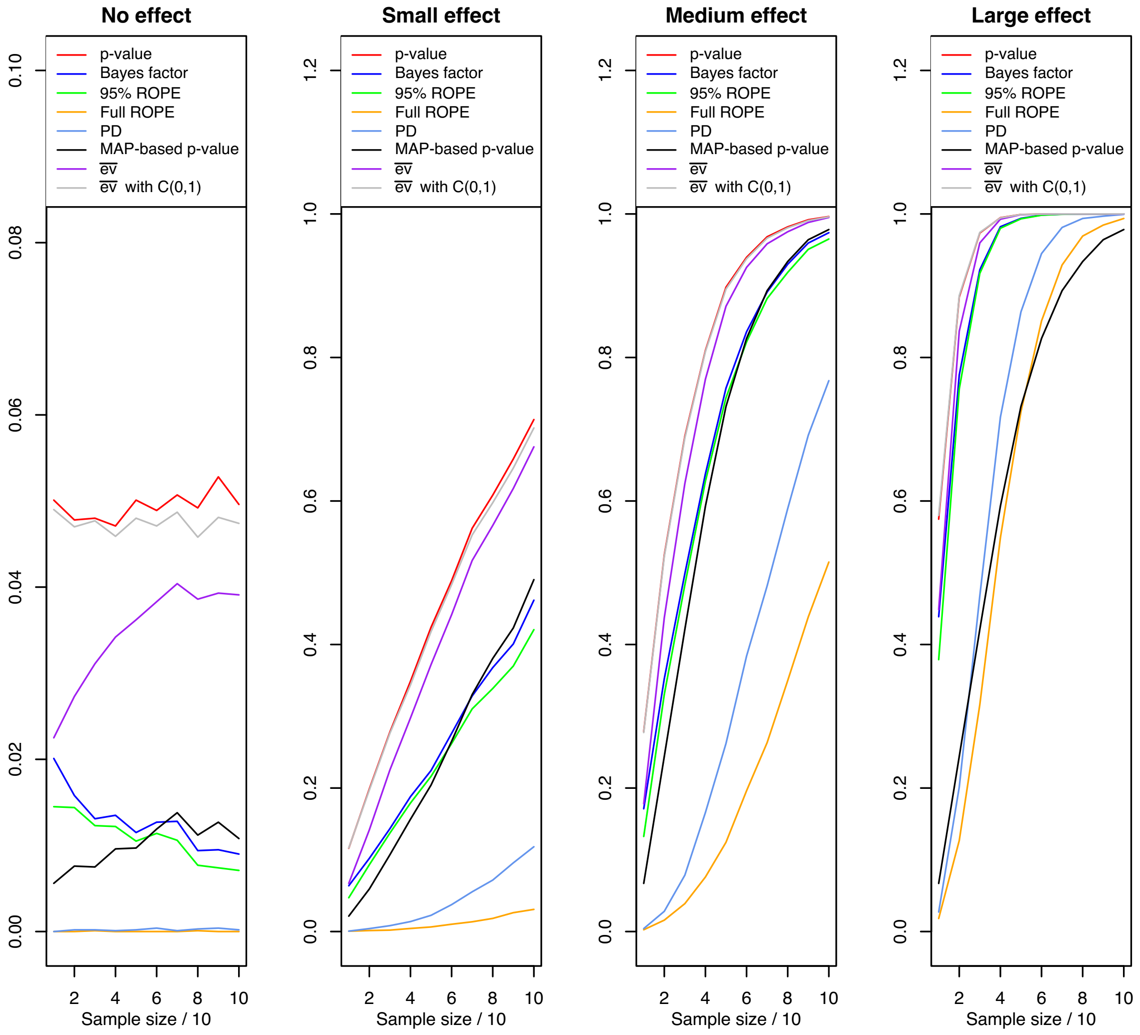

Figure 14.7: Sensitivity of Bayesian significance and effect size indices for small, medium, large and no existing effects using a wide prior $C(0,1)$ on the effect size δ and varying sample size n

Figure 14.7 visualises the results: The left plot corresponds to the table row of no effect and shows the type I error rates of the indices. As shown in the figure, the classic p-value fluctuates around its nominal significance level of $\alpha = .05$, although there is no effect present. In contrast, most Bayesian indices have lower type I error rates about half the size as the classic p-value. A comparison of the Bayesian posterior indices reveals three groups: The first group consists of the Bayes factor BF_{10}, the 95% ROPE and the MAP-based p-value. These indices concentrate around a false-positive rate of about 1% for increasing sample size. Still, the Bayes factor and ROPE make more type I errors for small sample size, while the MAP-based p-value errs more often for large sample sizes. The second group consists of the PD and the full ROPE, both of which make practically no type I error independent of the sample size n. This fact can be attributed to the quite conservative behaviour of both indices compared to the indices in group one. The third group consists of the e-value with improper or wide Cauchy prior, which achieves type

I error rates slightly smaller than the traditional p-value, but more massive than the other Bayesian indices.

The second plot corresponds to the small effect part of Table 12.10. Now the desired behaviour is that the indices detect the existing effect for the smallest possible sample size n. The classic p-value has the most liberate behaviour in stating that an effect is present, which reflects the often criticised fact that p-values overstate the significance of an effect compared to other indices of effect size and significance, see Wasserstein and Lazar (2016) and compare Chapter 1 and the Berger-Sellke upper bound in Chapter 11. The Bayesian indices signal evidence for the alternative more slowly than their frequentist counterparts, and the three groups already discovered in the first plot reveal themselves again: The BF_{10}, the 95% ROPE and the MAP-based p-value detect the small effect more often than the indices of the second group, which again includes the full ROPE and the PD. The third group consisting of the two versions of the e-value shows similar behaviour as the p-value: They signal the existence of an effect more quickly than their Bayesian competitors, which comes at the cost of increased type I errors as shown in the left plot previously.

The third and fourth plot correspond to the medium and large effect part of Table 14.1 and confirm the previous analysis. The p-value and e-value(s) state significance more often than every other index, but BF_{10}, the 95% ROPE and the MAP-based p-value yield a similar behaviour for increasing effect size δ now. Also, from the succession of the PD and full ROPE, it becomes clear that the PD more often states the presence of an effect in contrast to the full ROPE, which is more conservative, even for increasing effect size. Still, for increasing sample size, these "slow" indices eventually state the presence of the effect, too. Interestingly, the MAP-based p-value has a similar behaviour for large effect sizes as the full ROPE and PD, as shown in the right plot of Figure 14.7. The behaviour of the e-value again shows substantial similarity to the behaviour of the p-value under the medium and large effect setting.

14.4 Discussion

This chapter studied the behaviour of common Bayesian evidence measures for hypothesis testing in the setting of two-sample Welch's t-test, which is often applied in the biomedical and cognitive sciences. To guide researchers in choosing an appropriate evidence measure when the Bayesian counterpart to Welch's two-sample t-test as proposed by Rouder et al. (2009) is used instead, an extensive simulation study was conducted to analyse the influence of sample size n, the prior modelling and noise ε. Also, the type I error rates and sensitivities to detect an existing effect were studied.

The results show that one can split Bayesian significance and effect indices into two categories: Indices which can state evidence for the null hypothesis $H_0 : \delta = 0$ *and* the alternative $H_1 : \delta \neq 0$, and indices which can only state evidence for the alternative. The first group consists of the Bayes factor, the 95% and full ROPE. The MAP-based p-value, the PD and the e-value belong to the second group, the MAP-based p-value and the e-value showing a similar behaviour as the classic p-value. On the other hand, the e-value showed the best performance compared to all other indices when H_1 was true, and based on its other properties – for a review see Pereira et al. (2008) and Kelter and Stern (2020) – it is preferable over the MAP-based p-value, PD and classic p-value. The PD suffers from the fact that under H_0 it stabilizes at about 0.7, which is unintuitive and has to be interpreted as a tendency to favour evidence for the alternative when in fact the null hypothesis H_0 is true, see Figures 14.3, 14.4 and 14.5. Thus, when rejection of a null hypothesis is the goal, it is recommended to use the FBST and report the *e*-value based on the corresponding Cauchy prior as reference density in the surprise function. Also, the *e*-value is coherent with the likelihood principle and is very robust against the prior modelling. Importantly, it requires very little change in methodology when transitioning from frequentist p-values to Bayesian hypothesis tests and it is widely applicable as long as the posterior distribution can be obtained via MCMC, which is nearly always the case through the advent of modern Hamiltonian-Monte-Carlo algorithms, compare Part III. Thus, the e-value may be an attractive option to improve the reproducibility of research. However, a clear disadvantage is that the FBST is not able to confirm a hypothesis.

If the goal of the scientific enterprise is to confirm a research hypothesis, based on the results, the Bayes factor, the 95% ROPE or the full ROPE should be considered. All three indices show similar behaviour regarding increasing sample size n, and state both evidence for H_0 and H_1 depending on the presence of an effect.

The prior modelling showed that both the ultrawide and medium prior on δ could possibly lead to cherry-picking by combining a selected index like a ROPE or BF with the prior: For example, choosing a medium prior when the goal is to confirm H_0, evidence for H_0 accumulates faster than when using a wide or ultrawide prior. If the goal is to find evidence for the alternative, evidence for H_1 accumulates faster when using a wide or ultrawide prior instead of a medium one.

Therefore, to safeguard an analysis, it is recommended to use the wide prior $C(0, 1)$ when the goal is to confirm a hypothesis, as this choice places itself in the middle between the two other extremes and prevents cherry-picking in the case where no prior information is available. Also, robustness analyses are recommended which analyse how results change under different prior assumptions as discussed in Chapter 12.

The analysis of the influence of noise showed that all Bayesian indices suffered from increasing noise under H_1 with no apparent patterns or regularities, or one of the indices being more robust to noise than the others.

The type I error rates, and the sensitivity to detect an existing effect revealed that all Bayesian indices should be preferred to the classic p-value, although the *e*-value showed only slightly reduced type I error rates compared to the traditional *p*-value. This result is essential, as the control of type I error rates is one of the most critical aspects in clinical trials, see McElreath and Smaldino (2015) and Ioannidis (2016). The results showed further that the full ROPE and the PD achieve the best control of type I errors. As the PD cannot transparently state evidence for the null as shown previously, the full ROPE may be the better choice to control type I errors in clinical trials where often the goal is confirmation of a hypothesis (U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, 2019).

While the Bayes factor, the MAP-based p-value, the *e*-value and the 95% ROPE are more sensitive and detect more effects when using the same sample size n, their type I error control is weaker.

14.5 Conclusion

To guide researchers in the selection of an appropriate index for the biomedical and cognitive sciences, this section provided various new results. Based on these results, the following guidelines can be provided: Whenever type I error control has priority, it is recommended to use the full ROPE. Like the Bayes factor and 95% ROPE, the full ROPE can state evidence for both the null and the alternative hypothesis. The influence of sample size n, noise ε and prior modelling is similar for all three indices, but the type I error rate control is better for the full ROPE. The slightly weaker sensitivity to existing effects can be overcome by increasing the study sample size n, as shown in Figure 14.7: For sample sizes of $n = 100$, the sensitivity is equal to the sensitivity of the Bayes factor and 95% ROPE when a large effect is present. When medium or small effects are present, larger sample sizes are required, but as often multiple hundreds of patients participate in clinical trials, the benefits of type I error control overshadow the higher costs incurred by increased sample size. However, in situations where it is difficult or costly to recruit enough study participants (e.g. the study of rare diseases) it is recommended to opt for the Bayes factor because the Bayes factor achieves slightly better power than the full ROPE.

When rejection of a hypothesis is the goal, the e-value is the recommended choice, as it has the best sensitivity to detect an existing effect of all indices, and is an attractive a Bayesian replacement of the traditional p-value. For more details see also Kelter and Stern (2020), who provide computational details.

Chapter 15

A new Bayesian two-sample t-test for the Effect Size based on the Hodges-Lehmann Paradigm

> When testing statistical hypotheses, we usually do not wish to take the action of rejection unless the hypothesis (...) is false to an extent sufficient to matter.
>
> Joseph Lawson Hodges & Erich Leo Lehmann
> Testing the Approximate Validity of Statistical Hypotheses

15.1 Introduction

The last chapter outlined how the error rates and power of Bayesian indices for significance and size of an effect can be quantified in practice. Simulation studies provide insights similar to traditional power analyses or theoretical error guarantees which are used in combination with frequentist hypothesis tests. This ensures that even for complicated trial designs simulation studies can reveal the long-term properties, in particular, the resulting error rates of Bayesian hypothesis tests although the Bayesian approach formally has no concept of a type I and II error. Such results can help in improving the reliability of Bayesian hypothesis tests and the acceptance of Bayesian adaptive designs in clinical trials, compare (U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, 2019).

In this chapter, a conceptually different approach to Bayesian hypothesis testing is pursued which tackles one of the most important issues of statistical hypothesis testing: The validity of point null hypothesis for scientific research. The approach proposed in this chapter builds on the Hodges-Lehmann paradigm which was first proposed by Hodges and Lehmann (1954), and advocates replacing the test of a precise null hypothesis with the test of a small interval hypothesis.

One of the most important criticisms of hypothesis testing includes the "vexing issue of the relevance of point null hypotheses" (Robert, 2016, p. 5). The criticism that point null hypotheses are not realistic goes back at least to Savage (1954), and has evolved as the result of an ongoing debate between statisticians and philosophers of science over the last decades. Savage (1954, p. 332-333) already noted that "null hypotheses of no difference are usually known to be false before the data are collected" and that "their rejection ... is not a contribution to science". Also, Good (1950, p. 90) argued when testing the fairness of a die that "From one point of view it is unnecessary to look at the statistics since it is obvious that no die could be absolutely symmetrical.". In a footnote, he added: "It would be no contradiction (...) to say that the hypothesis that the die is absolutely symmetrical is almost impossible. In fact, this hypothesis is an idealised proposition rather than an empirical one." (Good, 1950, p. 90). Meehl (1967, p. 108) argued similarly, and stressed the "universal agreement that the old point-null hypothesis (...) is [quasi-] always false in biological and social science.". The same argument was brought forward by Cohen (1990, p. 1308), who pointed out that the null hypothesis "taken literally (...) is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false).". Also, in the discussion of Berger and Delampady (1987), Kadane (1987, p. 347) commented that for the "last 15 or so years I have been looking for applied cases in which I might have some serious belief in a null hypothesis. (...) I do not expect to test a precise hypothesis as a serious statistical calculation.".

On the other hand, Good (1994, p. 241) argued that there is at least one example of a precise hypothesis, which states that there is no extrasensory perception. However, in Good (1950, p. 90) he already admitted that his earlier remark for the case of a throw of a die "applies to all experiment – even the ESP experiment, since there may be no way of designing it so that the probabilities are *exactly* equal to $\frac{1}{2}$.".[1] Another example of a true null hypothesis was presented by Berger and Delampady (1987) as the hypothesis that talking to plants has no effect on their growth. However, like Good (1950), they admitted that minor biases in the experimental design (e.g. in randomization) may result in statistical significance again so that ultimately the hypothesis becomes false a priori. One approach to save point null testing was also presented by Berger and Sellke (1987) who showed that for reasonably small interval hypotheses, point null hypotheses are at least useful approximations of such small interval hypotheses (Berger and Delampady, 1987, Theorem 2). Good (1994) argued similarly and pointed out that the precise null hypothesis is simpler and "often a good enough approximation" (Good, 1994, p. 241). However, Bernado (1999) showed that this approximation breaks down for sufficiently large sample size, and Rousseau (2007) showed that for such large sample sizes, also the Bayes factor for a point null hypothesis is no reasonable approximation of the Bayes factor for an interval hypothesis anymore, unless the interval sizes are extremely small. This is problematic, because today large amounts of data are observed, and the times of small to moderate samples which were collected during the early days of statistics when Fisher or Neyman and Pearson proposed their theories of statistical hypothesis testing have long gone by. Thus, the argument that point null hypotheses are reasonable approximations of small interval hypotheses does not hold anymore. In summary, there is near consensus in the literature that "sharp null hypotheses are seldom exactly true." (Good, 1994, p. 241) and a "null hypothesis can usually be made more realis-

[1] In his conclusion, Good (1950, p. 94-95) remarked that "if n is very large, the test will probably give a significant result, because the chances $p_1, p_2, ..., p_6$ can hardly be exactly equal.", which is an implicit assertion that the null hypothesis is always false.

tic in principle by "spreading" the hypothesis over a small region in parameter space." (Good, 1994, p. 241).[2]

In this chapter, Good's proposal is followed and a new Bayesian hypothesis test for the Behrens-Fisher problem is proposed which replaces the test of a point null hypothesis with the test of a (small) interval hypothesis. The performance of the procedure is compared to the traditional frequentist solution to this problem, Welch's two-sample t-test, and theoretical results show that the proposed Bayesian test enjoys desirable properties. As the principal approach of testing small interval hypotheses was first proposed by Hodges and Lehmann (1954), the resulting test is called a Hodges-Lehmann test.

In medical research, the t-test is one of the most popular statistical procedures conducted. In randomized controlled trials (RCT), the goal often is to test the efficacy of new treatments or drugs and find out the size of an effect. Usually, a treatment and control group are used, and differences in a response variable like blood pressure or cholesterol level between both groups are observed. The gold standard for deciding if the new treatment or drug is more effective than the status quo treatment or drug is the p-value, which is the probability, under the null hypothesis H_0, of obtaining a difference equal to or more extreme than what was actually observed. The dominance of p-values when comparing two groups in medical (and other) research is overwhelming Nuijten et al. (2016).

The original two-sample t-test belongs to the class of frequentist solutions. These are based on sampling statistics, which allow to reject the null hypothesis via the use of p-values. The misuse and drawbacks of p-values in medical research have been detailed in Chapter 1. On the other side, Bayesian versions of the two-sample t-test have become more popular recently. Examples include the proposals in Gönen et al. (2005), Rouder et al. (2009), Wetzels et al. (2011), Wang and Liu (2016) and Gronau et al. (2020). All of these focus on the Bayes factor (BF) for testing a null hypothesis $H_0 : \delta = 0$ of no effect against a one- or two-sided alternative $H_1 : \delta > 0$, $H_1 : \delta < 0$ or $H_1 : \delta \neq 0$. Bayes factors themselves are also not without problems: (1) Bayes factors are sensible to prior modeling Kamary et al. (2014); (2) Bayes factors require the researcher to calculate marginal likelihoods, the calculations of which can be complex except when conjugate distributions exist; (3) In the setting of the two-sample t-test, Bayes factors weight the evidence for $H_0 : \delta = 0$ against the evidence for $H_1 : \delta \neq 0$ (or $H_1 : \delta < 0$, or $H_1 : \delta > 0$) given the data x. In the case when $BF_{10} = 20$, H_1 is 20 times more likely after observing the data than H_0. The natural question following in such cases is: How large is δ? A Bayes factor cannot answer this question and was not designed to answer such questions, but often this is of most relevance in applied biomedical research. Last, in most applied research, estimation of the effect size δ is more desirable than a mere rejection or acceptance of a point or composite hypothesis (Kruschke and Liddell, 2018b).

Of course, Bayes factors can be computed alongside posterior estimates, so testing and estimation do not mutually exclude each other. However, the mixture prior which is required to calculate a Bayes factor which can confirm a research hypothesis as detailed in Chapter 7 is unreasonable from a parameter estimation perspective: Assigning a prior probability to a point value θ_0 contradicts the usual prior beliefs about the pa-

[2] Earlier, Good (1993) proposed to change the current terminology: "when a statistician says that a hypothesis is rejected it would usually be better to say that the hypothesis is probably inexact (...). For the above reason (...) the words should be replaced by some other word, such as *inexactify*." (Good, 1993, p. 91).

rameter when the goal is parameter estimation. However, to be able to test a point null hypothesis via a Bayes factor one is forced to adopt Jeffreys' and Haldane's mixture prior representation.

In this chapter, this problem is bypassed by replacing the point null with an interval hypothesis, and reformulating the statistical model of the test as a two-component Gaussian mixture with known allocations. Also, the region of practical equivalence (ROPE) is employed as a criterion in the resulting test. Instead of focussing on rejection or confirmation of hypotheses, the proposed method's focus lies on estimation of the effect size under uncertainty. Together, the approach is an implementation of the Hodges-Lehmann paradigm as proposed by Hodges and Lehmann (1954) and Rao and Lovric (2016).

15.2 Methods

15.2.1 Modeling the Bayesian t-test as a mixture model with known allocations

In this section, the two-sample t-test is modelled as a two-component Gaussian mixture with known allocations. It is helpful to recite the idea of a mixture distribution:

> "Consider a population made up of K subgroups, mixed at random in proportion to the relative group sizes $\eta_1, ..., \eta_K$. Assume interest lies in some random feature Y which is heterogeneous across and homogeneous within the subgroups. Due to heterogeneity, Y has a different probability distribution in each group, usually assumed to arise from the same parametric family $p(y|\theta)$ however, with the parameter θ differing across the groups. The groups may be labeled through a discrete indicator variable S taking values in the set $\{1, ..., K\}$.
> When sampling randomly from such a population, we may record not only Y, but also the group indicator S. The probability of sampling from the group labeled S is equal to η_S, whereas conditional on knowing S, Y is a random variable following the distribution $p(y|\theta_S)$ with θ_S being the parameter in group S. (...) The marginal density $p(y)$ is obviously given by the following mixture density
>
> $$p(y) = \sum_{S=1}^{K} p(y, S) = \eta_1 p(y|\theta_1) + ... + \eta_K p(y|\theta_K)$$
>
> " Frühwirth-Schnatter (2006, p. 1)

Clearly, this resembles the situation of the two-sample t-test, in which the allocations S are known. While traditionally mixtures are treated with missing allocations, in the setting of the two-sample t-test these are known, leading to a "degenerate" mixture[3]. While this assumption does not only remove computational difficulties (these include problems like label switching, see Frühwirth-Schnatter (2006)), it also makes sense from a practical perspective: the inherent assumption of a researcher is that the population is indeed made up of $K = 2$ subgroups, which differ in a random feature Y which

[3] The mixture is called degenerate here, because when allocations are known, the likelihood is not mixed in the classical sense.

is heterogeneous across groups and homogeneous within each group. The group indicator S of course is recorded. When conducting a randomized controlled trial (RCT), the clinician will choose the patients according to a sampling plan, which could be set to achieve equally sized groups, that is, $\eta_1 = \eta_2$. Therefore, when sampling the population with the goal of equally sized groups, the researcher takes samples with equal probability from the population. For example, when a treatment group is compared to a control group, the block-randomization design ensures that the clinician prescribes the drug to a prespecified percentage η_1 of participants, and in balanced designs $\eta_1 = \eta_2 = \frac{1}{2}$. After the RCT is conducted, the resulting histogram of observed Y values will take the form of the mixture density $p(y)$ above and express bimodality due to the mixture model of the data-generating process.[4] After fixing the mixture weights, the family of distributions for the single groups needs to be chosen. The above considerations lead to consider finite mixtures of normal distributions, as these "occur frequently in many areas of applied statistics such as [...] medicine" (Frühwirth-Schnatter, 2006, p. 169). The components $p(y|\theta_i)$ become $f_N(y; \mu_i, \sigma_i^2)$ for $i = 1, ..., K$ in this case, where $f_N(y; \mu_i, \sigma_i^2)$ is the density of the univariate normal distribution. Parameter estimation in finite mixtures of normal distributions consists of estimation of the component parameters (μ_i, σ_i^2), the allocations $S_i, i = 1, ..., n$ and the weight distribution $(\eta_1, ..., \eta_K)$ based on the available data $y_i, i = 1, ..., n$. In the case of the two-sample Bayesian t-test, the allocations S_i (where $S_i = 0$ if y_i belongs to the first component and else $S_i = 1$) are known for all observations $y_i, i = 1, ..., n$. Also, the weights η_1, η_2 are known. Therefore, inference is concerned only with the component parameters μ_k, σ_k^2 given the complete data S, y.

Definition 15.1 (Bayesian two-sample t-test model)**.** Let S, Y be random variables with S taking values in the set $\{1, 2\}$ and Y in $\mathbb{R}$. If $Y|S = i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$, so conditional on S the component densities of Y are Gaussian with unknown parameters μ_i and σ_i^2, and if the marginal density is a two-component Gaussian mixture with known allocations, with marginal density

$$p(y) = \eta_1 f_N(y; \mu_1, \sigma_1^2) + \eta_2 f_N(y; \mu_2, \sigma_2^2)$$

where $\eta_2 := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{S_i=1}(y_i, S_i)$ and $\eta_1 = 1 - \eta_2$, the complete data S, Y are said to follow the Bayesian two-sample t-test model.

15.2.2 Inference via Gibbs Sampling

From the above line of thought it is clear that due to the representation via a mixture model with known allocations, no prior is placed directly on the effect size $\delta := \frac{\mu_1 - \mu_2}{s}$ itself, where

$$s := \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

and s_1^2 and s_2^2 are the empirical variances of the two groups, see also Cohen (1988). This is the common approach in existing Bayesian t-tests (Gronau et al., 2020). Instead,

[4] If unbalanced groups are the goal, the weights could be adjusted accordingly. As in most cases equally sized groups are considered, $\eta_1 = \eta_2 = 0.5$ is a justified assumption regarding the sampling process in the study or experiment.

in the proposed mixture model, priors are assigned to the parameters of the Gaussian mixture components μ_1, μ_2 and σ_1^2, σ_2^2. This has several benefits: Incorporation of available prior knowledge is easier achieved for the mixture component parameters than for the effect size, which is an aggregate of these component parameters. Consider a drug where from biochemical properties it can safely be assumed that the mean in the treatment group will become larger, but the variance will increase, too. Incorporating such knowledge on μ_i and σ_i is much easier than incorporating it in the prior of the effect size δ. This situation holds in particular, when group sizes n_1, n_2 are not balanced. These practical gains of translating prior knowledge into prior parameters comes at a cost: In contrast to existing solutions the model implies that no closed form expression for the posterior of δ (or the Bayes factor) is available anymore. Therefore, MCMC sampling is used here, to first construct the joint posterior $p(\mu_1, \mu_2, \sigma_1, \sigma_2 | S, y)$ and subsequently use a sample

$$\left((\mu_1^{(1)}, \mu_2^{(1)}, \sigma_1^{(1)}, \sigma_2^{(1)}), ..., (\mu_1^{(m)}, \mu_2^{(m)}, \sigma_1^{(m)}, \sigma_2^{(m)})\right)$$

of size m, to produce a sample $(\delta^{(1)}, \delta^{(2)}, ..., \delta^{(m)})$ of δ, where $\delta^{(i)} := \frac{\mu_1^{(i)} - \mu_2^{(i)}}{s^{(i)}}$ and

$$s^{(i)} = \sqrt{\frac{(n_1 - 1)(s_1^{(i)})^2 + (n_2 - 1)(s_2^{(i)})^2}{n_1 + n_2 - 2}}$$

In summary, via Gibbs sampling (compare Chapter 8 and Robert and Casella (2004)), the posterior of δ can be simulated via Markov-Chain-Monte-Carlo. In order to apply Gibbs sampling, the conditional distributions need to be derived.

15.2.3 Derivation of the full conditionals using the independence prior

To derive the full conditionals, the prior distributions for the mixture component parameters need to be selected. There are multiple priors available, the most prominent among them the conditionally conjugate prior and the independence prior (Escobar and West, 1995; Frühwirth-Schnatter, 2006). While the conditionally conjugate prior has the advantage of leading to a closed-form posterior $p(\mu, \sigma^2 | S, y)$, the main difficulty in the setting of the Bayesian two-sample t-test is that while a priori the component parameters $\theta_k = (\mu_k, \sigma_k^2)$ are pairwise independent across both groups, inside each group the mean μ_k and variance σ_k^2 are dependent. This is in contrast to the assumption in the setting of the Bayesian two-sample t-test, and therefore the independence prior is chosen, which is used in Escobar and West (1995) and Richardson and Green (1997). The independence prior assumes the mean μ_k and the variance σ_k^2 are a priori independent, that is $p(\mu, \sigma^2) = \prod_{k=1}^K p(\mu_k) \prod_{k=1}^K p(\sigma_k^2)$, with $\mu_k \sim \mathcal{N}(b_0, B_0)$ and $\sigma_k^2 \sim IG(c_0, C_0)$, where $IG(\cdot)$ denotes the inverse Gamma distribution. The normal prior on the means μ_k seems reasonable as the parameters b_0 and B_0 can be chosen to keep the influence of the prior only weakly informative.[5] The inverse Gamma prior is chosen because for a two-component Gaussian mixture to show any signs of bimodality – in which case one would assume differences between two subgroups in the whole sample – the variance should not be huge, because otherwise the modes (or the bell-shape) of the two

[5] Another option would be a t_n prior, but this would also imply another free hyperparameter to be estimated simultaneously, which would be the degrees of freedom n of the t_n-distribution.

normal-components of the mixture will flatten out more and more, until unimodality is reached. Thus, the inverse Gamma prior connects this model aspect by giving more probability mass to smaller values of σ_k^2, while extremely large values get much less prior probability mass.[6] The hyperparameters c_0 and C_0 then offer control over this kind of shrinkage on σ_k^2 towards zero. In the simulation study below the prior sensitivity will also be studied briefly. The independence prior is therefore used and leads to the following full conditionals:

Theorem 15.2. For the Bayesian two-sample t-test model, the full conditional distributions under the independence prior

$$p(\mu, \sigma^2) := \prod_{k=1}^{K} p(\mu_k) \prod_{k=1}^{K} p(\sigma_k^2)$$

with $\mu_k \sim \mathcal{N}(b_0, B_0)$ and $\sigma_k^2 \sim IG(c_0, C_0)$ (where $IG(\cdot)$ denotes the inverse Gamma distribution) are given as:

$$\begin{aligned}
p(\mu_1|\mu_2, \sigma_1^2, \sigma_2^2, S, y) &= p(\mu_1|\sigma_1^2, S, y) \sim \mathcal{N}(b_1(S), B_1(S)) \\
p(\mu_2|\mu_1, \sigma_1^2, \sigma_2^2, S, y) &= p(\mu_2|\sigma_2^2, S, y) \sim \mathcal{N}(b_2(S), B_2(S)) \\
p(\sigma_1^2|\mu_1, \mu_2, \sigma_2^2, S, y) &= p(\sigma_1^2|\mu_1, S, y) \sim IG(c_1(S), C_1(S)) \\
p(\sigma_2^2|\mu_1, \mu_2, \sigma_1^2, S, y) &= p(\sigma_2^2|\mu_2, S, y) \sim IG(c_2(S), C_2(S))
\end{aligned}$$

with $B_1(S), b_1(S), B_2(S), b_2(S)$ as defined in Equations (A.12) and (A.13), and $c_1(S)$, $c_2(S)$, $C_1(S)$ and $C_2(S)$ as defined in Equations (A.14) and (A.15) in Appendix A.1.

Proof. See Appendix A.2, which builds on the derivations in Appendix A.1. □

Note that when $\eta_1 \neq \eta_2$, $N_1(S)$ and $N_2(S)$ in the Appendix just need to be changed accordingly. For example, if the first group consists of 30 observations, and the second group of 70, setting $N_1(S) = 30$ and $N_2(S) = 70$ implies $\eta_1 = 0.3$ and $\eta_2 = 0.7$, handling the case of unequal group sizes easily.

15.2.4 Derivation of the single-block Gibbs sampler

Based on the full conditionals derived in the last section, this section now derives a single-block Gibbs sampler to obtain the joint posterior distribution

$$p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y)$$

given the complete data (S, y). The resulting Gibbs sampler is given as follows:

Corollary 15.3 (Single-block Gibbs sampler for the Bayesian two-sample t-test)**.** The joint posterior distribution

$$p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y)$$

in the Bayesian two-sample t-test model can be simulated under the independence prior as follows:
Conditional on the classification $S = (S_1, ..., S_N)$:

[6] Another option would be an exponential prior with parameter λ, but as the exponential distribution is just a special case of the gamma distribution, and the inverse gamma distribution is directly related to the gamma distribution, the more general inverse gamma prior is selected here.

1. *Sample σ_k^2 in each group k, $k = 1, 2$ from an inverse Gamma distribution $IG(c_k(S), C_k(S))$*
2. *Sample μ_k in each group k, $k = 1, 2$, from a normal distribution $\mathcal{N}(b_k(S), B_k(S))$*

where $B_k(S), b_k(S)$ and $c_k(S), C_k(S)$ are given by equations $(A.12)$, $(A.13)$, $(A.14)$ and $(A.15)$ in Appendix A.1.

Proof. See Appendix A.3. □

15.2.5 The Hodges-Lehmann paradigm and the region of practical equivalence (ROPE)

Hodges and Lehmann (1954) discussed the validity of statistical hypotheses more than half a century ago, and concluded that testing small interval hypotheses is more realistic:

> "When testing statistical hypotheses, we usually do not wish to take the action of rejection unless the hypothesis (...) is false to an extent sufficient to matter. For example, we may formulate the hypothesis that a population is normally distributed, but we realize that no natural population is ever exactly normal."
> Hodges and Lehmann (1954, p. 261)

They proposed to introduce "into the space of parameters a measure, say $\Delta(\theta)$ of the distance of θ from H_0 on a scale reflecting at least roughly the materiality of departures from H_0, and then define H_1 as the set of those θ for which $\Delta(\theta)$ does not exceed a specified value Δ_0." (Hodges and Lehmann, 1954, p. 262). Their approach can be formalized as testing

$$H_0 : \theta \in [\theta_0 - \Delta_0, \theta_0 + \Delta_0] \text{ versus } H_1 : \theta \notin [\theta_0 - \Delta_0, \theta_0 + \Delta_0] \tag{15.1}$$

instead of

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0 \tag{15.2}$$

As discussed in the introduction section, a Hodges-Lehmann test is more realistic in the majority of situations faced in the biomedical, social and natural sciences than a hypothesis test for a precise point null hypothesis.

One proposal which is very similar to the approach of Hodges and Lehmann is the one of Kruschke and Liddell (2018b), which advocates the *region of practical equivalence* (*ROPE*) that was already studied in Chapter 14. As they note: "ROPE's go by different names in the literature, including "interval of clinical equivalence", "range of equivalence", "equivalence interval", "indifference zone", "smallest effect size of interest," and "good-enough belt" ..." (Kruschke and Liddell, 2018b, p. 185), where these terms come from a wide spectrum of scientific domains, see Carlin and Louis (2009), Freedman et al. (1983), Hobbs and Carlin (2007), Lakens (2014) and Schuirmann (1987). The uniting idea is to establish a region of practical equivalence around the null value of the hypothesis, which expresses "the range of parameter values that are equivalent to the null value for current practical purposes." (Kruschke and Liddell, 2018b, p. 185). With a caution not to slip back into dichotomic black-and-white thinking, the following decision rule was proposed by Kruschke and Liddell (2018b): Reject the null value, if the 95% highest posterior density interval (HPD) falls entirely outside the ROPE. Accept the null value, if the 95% HPD falls entirely inside the ROPE. In the first case, with more than 95% probability the parameter value is not inside the ROPE, and therefore not practically equivalent to the null value. A rejection of the null value then seems legitimate. In the second case, the parameter value is inside the ROPE with at least 95% posterior probability, and therefore practically equivalent to the null value. It seems legitimate to accept the null value. Of course, it would also be possible to accept the null value iff the whole posterior is located inside the ROPE, leading to an even stricter decision rule. The ROPE is thus a direct implementation of the Hodges-Lehmann approach when it is interpreted as an interval hypothesis. Below, the ROPE is formalized this way:

Definition 15.4 (ROPE)**.** The region of practical equivalence (ROPE) R for (or around) a hypothesis $H \subset \Theta$ is a subset of the parameter space Θ with $H \subset R$.

Given the above definition, a statistical hypothesis H is now described via a region of practical equivalence R. For example, the hypothesis $H : \delta = \delta_0$ can be described as $R := [\delta_0 - \varepsilon, \delta_0 + \varepsilon]$ for $\varepsilon > 0$. Thus, the ROPE is precisely a hypothesis in the Hodges-Lehmann paradigm when formalized as in the above definition. Also, by definition, any set $R \subset \Theta$ with $H \subset R$ is allowed to describe H, and should be selected depending on how precise the measuring process of the experiment or study is assumed to be. Next, two options for the ROPE are defined:

Definition 15.5 (Correctness)**.** Let $R \subset \Theta$ a ROPE around a hypothesis $H \subset \Theta$, that is $H \subset R$, where H makes a statement about the unknown model parameter θ. If the true parameter value θ_0 lies in R, that is $\theta_0 \in R$, then R is called correct, otherwise incorrect.

A correct ROPE therefore contains the true parameter value θ_0, while an incorrect one does not.

15.2.6 Boundary elicitation for the interval hypothesis

While the Hodges-Lehmann approach is conceptually appealing, a major challenge is the selection of the interval hypothesis boundaries (which are ROPE boundaries). However, there is a vast range of options to determine these boundaries:

1. Lakens et al. (2018) proposed to base the selection on resource availability: According to them, researchers often know better which sample sizes are attainable in their field of work than which effect sizes can expected to be observed in a study. As the amount of available data limits the effect size that can be detected, researchers can derive the smallest effect size which they can detect after selecting a test level α and their sample size n and use this smallest detectable effect size as the equivalence boundary. Note that although it seems that this method primarily applies to frequentist tests because the Bayesian paradigm contains no concept of a type I error, the simulation study conducted in Chapter 14 showed that it is straightforward to study the smallest detectable effect size and the resulting error rates also for Bayesian tests, see also Kelter (2020a,e) and Makowski et al. (2019b).

2. The U.S. Food and Drug Administration has set equivalence bounds for establishing bioequivalence (U.S. Food and Drug Administration Center for Drug Evaluation and Research, 2001), which can be interpreted as the interval hypothesis bounds. For a thorough discussion of current challenges see also Senn (2001).

3. Cook et al. (2014, 2018) proposed three methods for determining the bounds: First, the anchor method for determining the minimally clinical important difference (MCID), where the judgement of relevant stakeholders is used, see also Jaeschke et al. (1989). Second, the distribution method, where both the standard error of a measurement and the smallest detectable difference of a statistical test is employed. Third, the health economic method which aims at optimising the cost of a "unit of health" for the amount of money spent: Termed differently, which effect is necessary in "health units" to justify the amount of money spent for the treatment or therapy?

4. Weber and Popova (2012) recommended to incorporate subject-domain knowledge from meta-analyses to determine the boundaries in a more principled way.

5. Simonsohn (2015) proposed to set the boundary at the effect size that a previous study would have had $\approx 33\%$ power to detect. For details on the motivation and justification of this so-called *small-telescopes* approach see also Lakens et al. (2018).

6. Ferguson (2009), Beribisky et al. (2019) and Rusticus and Eva (2016) have argued for incorporating pilot studies for boundary selection.

7. Other approaches and examples which base the equivalence bound selection on previous research are given by Perugini et al. (2014) and Kordsmeyer and Penke (2017).

8. In case none of the other justifications of interval hypothesis boundaries is possible, Maxwell et al. (2015) recommended to use a trivially small value like an effect size of $\delta = .10$ according to Cohen (1988) as the boundary of the equivalence region. Lakens et al. (2018) underlined that this is the weakest possible justification of such a boundary.

9. Kruschke (2018) provided an in-depth discussion of selecting the boundaries for the ROPE in the Bayesian approach. Kelter (2020a,e) and Makowski et al. (2019b) provided simulation-based insights which relate the ROPE width to statistical quantities like type I and II errors and the sensitivity and specificity of a test in the context of two-sample tests and regression settings. This allows for selecting the ROPE boundaries based on objective criteria like a desired maximum type I or type II error rate.

10. Ultimately, "the ideal decision about a specific meaningful effect should be made through a multi-faceted decision-making process, standardized context-free effect sizes provide helpful additional information when there are no other viable alternatives" (Beribisky et al., 2019, p. 5), see also Rogers et al. (1993).

In summary, there is a vast range of techniques that allow transitioning from precise to interval hypothesis testing according to the Hodges-Lehmann paradigm. The selection of the interval hypothesis boundaries may feel subjective. However, the selection of a precise point null hypothesis is equally subjective. The fact that current practice primarily consists of using strawman null-effect or null-difference hypotheses (compare the development of the inconsistent hybrid approach in Chapter 5) only masquerades the problem that in most situations, these precise null hypotheses are meaningless: One would not entertain the effort to study the efficacy of a new drug or the outcome of an educational intervention if the a priori belief would be that there will be no effect after all. In practice, researchers often invest time and effort because they believe that there has to be at least some non-negligible effect, and the null hypothesis is not what most scientists are interested in. The no-effect hypothesis $H_0 : \delta = 0$ is only put up to reject it and conclude that there is indeed a non-zero effect. However, it could be argued that this is evident a priori as there will always be *any* effect (although maybe a negligibly small one).[7] In contrast, a reasonable null hypothesis would postulate an effect θ_0 and interest lies in investigating if an effect is at least as large or at least as small as a pre-specified value θ_0. That is, then the one-sided hypothesis testing setting $H_0 : \theta \leq \theta_0$ (or $H_0 : \theta \geq \theta_0$) is recovered, and selection of the value θ_0 is equally arbitrary as the selection of the boundaries for the ROPE. In summary, although the boundary selection presents a challenge for practical research, it is a challenge to face, not to evade.[8]

15.2.7 Implementing the Hodges-Lehmann paradigm

The two major drawbacks of the proposal of Kruschke (2018) are that the ROPE still facilitates hypothesis testing, enforcing a binary decision of rejection or acceptance, while it is also unclear what to do when the 95%-HPD lies partly inside and partly outside the ROPE. Therefore, a different proposal is made here, which is estimation of the *mean probable effect size* (*MPE*) in the proposed two-sample t-test. The method can be generalized for other tests analogue. First, the acceptance or rejection of a hypothesis H can be formalized as follows:

Definition 15.6 (α-accepted / α-rejected)**.** Let θ the unknown parameter (or vector of unknown parameters) in an experiment $E := \{X, \theta, \{f_\theta\}\}$, where the random variable X taking values in $\mathbb{R}$ and having density f_θ for some $\theta \subset \Theta$, is observed. Let $p(\theta|x)$ the posterior distribution of θ (under any prior $p(\theta)$), and let C_α the corresponding α% highest density interval of $p(\theta|x)$. Let $R \subset \Theta$ a ROPE around the hypothesis H of interest, which makes a statement about θ. Then, if $C_\alpha \subset R$, the hypothesis H is called α-accepted, else α-rejected. If $\alpha = 1$, then H is simply called accepted, else rejected.

Thus, if C_α lies completely inside the ROPE R and if $\alpha = 1$, the entire posterior probability mass indicates that θ is practically equivalent to the values described by the ROPE R. Thus, H can be accepted. If $\alpha < 1$, the strength of this statement becomes less with decreasing α. For example, if H is 0.75-accepted for a given ROPE R, 25% of the posterior indicate that θ may take values different than the ones included in the ROPE R. It is clear that the use is limited if the value of α is small or close to zero when speaking of α-acceptance. Therefore, instead of forcing an acceptance or rejection (which only makes sense for substantial values of α, a perspective focussing on continuous estimation is preferred:

[7] This also weakens Bayes factor tests for precise point null hypotheses, because confirmation of a hypothesis which is a priori known to be false is of little use. However, there are modifications of Bayes factor tests like interval Bayes factors that bypass this problem, compare Morey and Rouder (2011).

[8] This is in close analogy to the prior selection in the Bayesian paradigm: Selecting so-called uninformative or flat priors leads to paradoxical behaviour in Bayesian hypothesis testing like the Lindley-paradox Lindley (1957), see Kelter (2020b). Thus, it is recommended to elicit a prior in a meaningful way instead of resorting to reference or default solutions. As Jeffreys once stated: “There are practical difficulties in assessing the prior probability in many cases as they actually arise. This is not a situation to evade, but one to face.” (Jeffreys, 1931, p. 34)

Definition 15.7 (Posterior mass percentage)**.** Let $p(\theta|x)$ the posterior density for θ and $\delta_{MPE} := \mathbb{E}[\theta|x]$ the mean posterior effect size (where the expectation $\mathbb{E}$ is taken with respect to the posterior probability measure). Let $R_1, ..., R_m$ be a partition of the support of the posterior $p(\theta|x)$ into different ROPEs corresponding to different hypotheses $H_1, ..., H_m$, which make statements about the unknown parameter (vector) θ. Without loss of generality, let R_j the ROPE for which $\delta_{MPE} \subset R_j$, $j \in \{1, ..., m\}$. The posterior mass percentage $PMP_{R_j}(\delta_{MPE})$ of δ_{MPE} is given as

$$PMP_{R_j}(\delta_{MPE}) := \int_{R_j} p(\theta|x)d\theta$$

that is, the percentage of the posterior distribution's probability mass inside the ROPE R_j around δ_{MPE}.

For simplicity of notation, the subscript R_j is omitted whenever it is clear which ROPEs R_j are used for partitioning the support of $p(\theta|x)$. Now, in contrast to strict α-acceptance or α-rejection rules based on the ROPE R_j, it is proposed to use δ_{MPE} and $PMP(\delta_{MPE})$ together to estimate the effect size δ under under uncertainty, and to quantify this uncertainty via $PMP(\delta_{MPE})$. If δ_{MPE} is non-zero, the t-test found a difference between both groups. The size of this difference is quantified by δ_{MPE} itself. The uncertainty in this statement is quantified by $PMP(\delta_{MPE})$. For the developed two-sample t-test, the following procedure is proposed:

1. For a fixed credible level α, the *effect size range* (ESR) should be reported. That is, which effect sizes δ are assigned positive probability mass by the α% HPD interval, $0 \leq \alpha \leq 1$. The ESR is a first estimate of credible effect sizes a posteriori.

2. The support of the posterior distribution $p(\delta|S, Y)$ in the Bayesian t-test model is partitioned into the standardized ROPEs of the effect size δ of Cohen (1988), leading to a partition $\mathcal{P}$ of the support as given in the definition of $PMP(\delta_{MPE})$.

3. The *mean posterior effect size* δ_{MPE} is calculated as an estimate of the true effect size δ. The surrounding ROPE R_j with $\delta_{MPE} \subset R_j$ of the partition $\mathcal{P}$ is selected, and the exact percentage inside R_j is reported as the *posterior mass percentage* $PMP(\delta_{MPE})$.

The above procedure leads to a simultaneous estimation of the effect size δ under uncertainty in combination with an interval hypothesis test in the Hodges-Lehmann paradigm. Additionally, next to the posterior mean δ_{MPE}, the posterior mass percentage $PMP(\delta_{MPE})$ gives a *continuous* measure of the trustworthiness of the estimate ranging from 0% to 100% (actually from zero to one, but for better interpretability the percentage of posterior probability mass which is located in the ROPE R_j will be used henceforth). δ_{MPE} estimates with $PMP(\delta_{MPE}) > 0.5$ (or 50%) could be interpreted as decisive, but do not need to. $PMP(\delta_{MPE})$ can be treated as a continuous measure of support for the effect size estimated by δ_{MPE}. There are multiple advantages of using a ROPE R_j and combining it with δ_{MPE} and $PMP(\delta_{MPE})$, the most important of which may be expressed in the following result:

Theorem 15.8. Let $R_j \subset \Theta$ a ROPE around δ_{MPE}, that is $\delta_{MPE} \subset R_j$. If R_j is correct, that is, the true parameter $\theta_0 \subseteq R_j$, then $PMP(\delta_{MPE}) \to 1$ for $n \to \infty$ almost surely, and if R_j is incorrect, then $PMP(\delta_{MPE}) \to 0$ for $n \to \infty$ π-almost surely, except possibly on a set of π-measure zero for any prior π on θ.

Proof. See Appendix A.4. □

Using δ_{MPE} together with $PMP(\delta_{MPE})$ therefore will eventually lead to the correct estimation of δ in the sense that when a correct ROPE is chosen, the posterior mass percentage will converge to one, and if an incorrect ROPE is chosen, the posterior mass percentage will converge to zero. Thus, the procedure indicates whether δ is practically equivalent to the values given by the ROPE or not. If necessary, explicit hypothesis testing can be performed via α-rejection. The advantages compared to p-values and Bayes factors which favour an explicit hypothesis testing perspective are:

1. As Greenland et al. (2016, p. 338) stressed with regard to the dichotomy induced by hypothesis testing, "estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference and sound judgment than any such classification."

2. In contrast to the Bayes factor (BF) the ROPE and δ_{MPE} have important advantages: they do not encourage the same automatic calculation routines as Bayes factors. For example, Gigerenzer and Marewski (2015) warned explicitly against Bayes factors becoming the new p-values due to the same automatic calculation routines, and the approach via the ROPE fosters estimation and judging the evidence based on the *continuous* support for δ_{MPE} provided by $PMP(\delta_{MPE})$, instead of using thresholds. Also, the explicit assignment of positive probability mass to a point null value which has Lebesgue measure zero as is required for a Bayes factor calculation is avoided in the proposed procedure. The Hodges-Lehmann approach is thus uniting parameter estimation and hypothesis testing from the perspective of prior elicitation.

3. In practice, measuring is always done with finite precision (like blood pressure, or the heart rate), and therefore the goal rarely is to show (or reject) that the effect size δ is exactly *equal* to zero, but much more that δ is *negligibly small* to deny the existence of any existing, (clinically) relevant effect. Therefore, invariances like $\delta = 0$ can be interpreted as not existing, at least not exactly, and the search for approximate invariances, as described by a ROPE $R = (-.2, .2)$ around $\delta = 0$ is (more) compelling. A clinician will be satisfied by the statement that the true effect size is not exactly zero, but with 95% probability negligibly small.

15.2.8 Illustrative example

To clarify the above line of thought, the following example illustrates the use of the developed Bayesian t-test in the Hodges-Lehmann paradigm. The Gibbs sampler, the ROPE, δ_{MPE} and $PMP(\delta_{MPE})$ are used to test an interval hypothesis in the two-sample setting. The illustrative example uses data from Wagenmakers et al. (2015), who conducted a randomized controlled trial in which participants had to fill out a personality questionnaire while rolling a kitchen roll clockwise or counter-clockwise. The mean score of both groups was compared afterwards. A traditional two-sided two-sample Welch's t-test indicates that there is no significant difference between both groups, yielding a p-value of 0.4542. What is missing is the effect size, which is of much more interest. Note that computing the effect size from the raw study data does not quantify the uncertainty in the data, which is undesirable. From the p-value, a clinician can only judge that the results are unlikely to be observed under the null hypothesis. However, if the effect is clinically relevant or negligible remains unknown (or at best only based on the raw sample effect size). In this case, as the p-value is quite large, and the null hypothesis of no effect can not be rejected. Also, it is not possible to state with certainty that there is indeed no effect in the sense of confirming the null hypothesis. Thus, the only option is to collect more data before drawing a conclusion. In contrast, Figure 15.1 shows an analysis of the posterior of δ that is produced by the Gibbs sampler for the Bayesian two-sample t-test model. In it, the ROPE, δ_{MPE} and $PMP(\delta_{MPE})$ are used. The posterior distribution of δ is given in the upper plot and shows that the

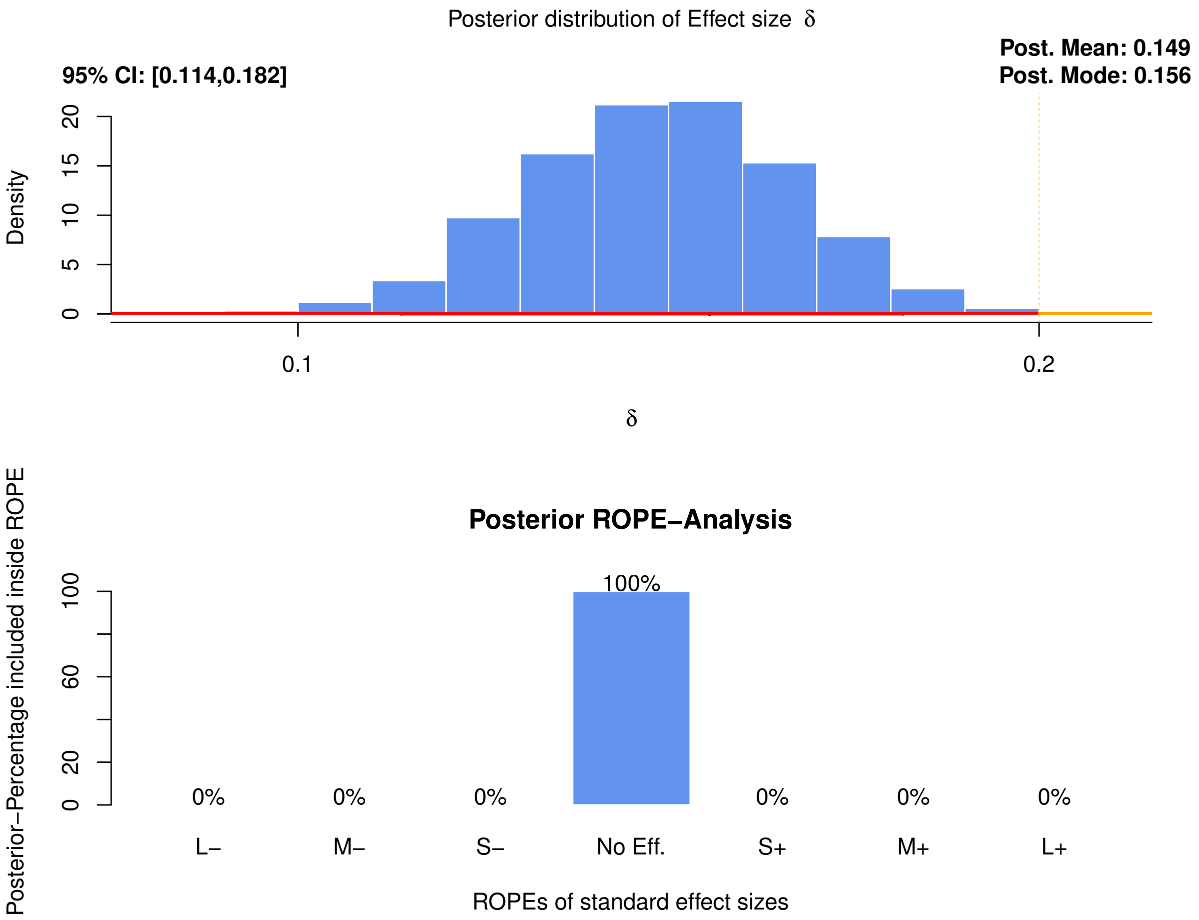

Figure 15.1: Posterior distribution of the effect size δ and analysis for the kitchen roll RCT of Wagenmakers et al. (2015) via δ_{MPE} and $PMP(\delta_{MPE})$

posterior mean is 0.149 and the posterior mode 0.156, that is, the mean posterior effect size given the data is 0.149, no effect discernible from zero. The 95% highest density interval ranges from 0.114 to 0.182, showing that with 95% probability, there is no effect discernible from $\delta = 0$, given the data. Even when taking the 100% highest posterior density interval (HPD), this situation does not change as indicated by the upper plot. The coloured horizontal lines and vertical dotted lines represent the boundaries of the different effect size categories according to Cohen (1988). The lower plot shows the results of partitioning the posterior mass of δ into the ROPEs that correspond to the different effect sizes, which are standardized as small, if $\delta \in (-0.5, -0.2] \cup [0.2, 0.5)$, medium, if $\delta \in (-0.8, -0.5] \cup [0.5, 0.8)$ and large, if $\delta \in (-\infty, -0.8] \cup [0.8, \infty)$ (Cohen, 1988). Now, 100% of this posterior probability mass is located inside the $ROPE$ $(-0.2, 0.2)$ of no effect. So $\delta_{MPE} = 0.149$ indicates that no effect discernible from zero (in the sense that effect sizes $||delta| < 0.2$ are interpreted as scientifically not relevant) is apparent, and the posterior mass percentage $PMP(\delta_{MPE}) = 1$ (or 100%) shows that the estimate δ_{MPE} is trustworthy, as the entire posterior probability mass is located inside the ROPE $(-0.2, 0.2)$ of no effect (also indicated by the upper plot). Based on this analysis, one can conclude that given the data, it is highly probable, that there exists no effect. The method provides more insight than the information a p-value is giving: In the example, the p-value cannot reject the null hypothesis $H_0 : \delta = 0$ and neither can the null hypothesis be confirmed. Even if the p-value would have been significant, this means only that the result will unlikely have happened by chance under the null hypothesis. The non-significant p-value of 0.4542 in this case does not allow to accept the null hypothesis $H_0 : \delta = 0$ of no effect. The proposed procedure in contrast does. Importantly, it allows to accept the interval hypothesis $H_0 : \delta \in (-0.2, 0.2)$ which is much more meaningful for practical research. The value $\delta_{MPE} = 0.149$ indicates that there is a non-zero effect, but for the situation at hand it is too small to be considered relevant.

Note also that a Bayes factor would have to be combined with estimation to yield the same information, and a Bayes factor alone of course would not have provided this information. In the example, the Bayes factor BF_{01} of $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ is $BF_{01} = 5.015$ when using the recommended wide Cauchy $C(0, 1)$ prior of Rouder et al. (2009), which indicates only moderate evidence for the null hypothesis $H_0 : \delta = 0$ according to van Doorn et al. (2021). This is in sharp contrast to the $PMP(\delta_{MPE})$ value of 100%, which strongly suggests that the null hypothesis $H_0 : \delta = 0$ is confirmed. The posterior in Figure 15.1 is obtained by the Gibbs sampler given in Corollary 1.

15.3 Simulation study

Primary interest now lies in the ability to correctly estimate different sizes of effects via the combination of the derived t-test, δ_{MPE} and $PMP(\delta_{MPE})$. The effect size ROPEs are oriented at the standard effect sizes of Cohen (1988), where an effect is categorized as *small*, if $\delta \in [0.2, 0.5)$ or $\delta \in (-0.5, -0.2]$, *medium*, if $\delta \in [0.5, 0.8]$ or $\delta \in (-0.8, -0.5]$ and *large*, if $\delta \geq 0.8$ or $\delta \leq -0.8$. Secondary interest lies in analysing if the Gibbs sampler achieves better performance regarding the type I and II error compared with Welch's t-test, the standard NHST solution. The plan of the simulation study is therefore as follows: If there is indeed an effect, the Gibbs sampler should lead to a posterior distribution of δ which lies outside the ROPE $(-0.2, 0.2)$, which is equivalent to the rejection of the interval null hypothesis $H_0 : \delta \in (-0.2, 0.2)$. The precise estimation of the size of an effect is a second task, one which is more demanding than the sole rejection of $H_0 : \delta \in (-0.2, 0.2)$. If the sampler correctly rejects the null hypothesis, because the posteriors concentrate in the set $(-\infty, -0.2] \cup [0.2, \infty)$, this indicates that it makes no type II error and subsequently achieves a power of nearly 100%. Of course, this will depend on the sample sizes in both groups. If additionally, the 95%-credible intervals of the posteriors concentrate in the set $\{(-0.5, -0.2] \cup [0.2, 0.5)\}$, then the Gibbs sampler is also consistent for small effect sizes, again depending on the sample size. The same rationale applies for medium effect sizes and the interval hypothesis $H_0 : \delta \in (-0.8, 0.5] \cup [0.5, 0.8)$ and for large effect sizes and the interval hypothesis $H_0 : \delta \in (-\infty, 0.8] \cup [0.8, \infty)$. Therefore, three two-component Gaussian mixtures have been fixed in advance, each representing one of the three effect sizes. For the small effect, the first component is $\mathcal{N}(2.89, 1.84)$ and the second component $\mathcal{N}(3.5, 1.56)$, resulting in a effect size of $\delta = (2.89 - 3.5)/\sqrt{((1.56^2 + 1.84^2)/2)} = -0.35$. For a medium effect, the first and second group are simulated as $\mathcal{N}(254.08, 2.36)$ and $\mathcal{N}(255.84, 3.04)$, yielding a true effect size of

$$\delta = \frac{(255.84 - 254.08)}{\sqrt{((3.04^2 + 2.36^2)/2)}} = 0.6467 \approx 0.65$$

For the large effect, the first and second group are simulated from normal distributions $\mathcal{N}(15.01, 3.4^2)$ and $\mathcal{N}(19.91, 5.8^2)$, yielding a true effect size of

$$\delta = \frac{(19.91 - 15.01)}{\sqrt{((5.8^2 + 3.4^2)/2)}} = 1.03$$

In each of the three effect size scenarios, 100 datasets of the corresponding two-component Gaussian mixture were simulated for different sample sizes and the Gibbs sampler was run for each of the 100 datasets for 10000 iterations, using a burnin of 5000 posterior draws which are discarded. Based on the resulting posterior, δ_{MPE}, the ESR and the ROPE criterion together with α-acceptance are applied, that is, the hypothesis H stating a small, medium or large effect size is α-accepted if the 95%-HPD lies completely inside the corresponding ROPE $\{(-0.5, -0.2] \cup [0.2, 0.5)\}$, $\{(-0.8, -0.5] \cup [0.5, 0.8)\}$ or $\{(-\infty, -0.8] \cup [0.8, \infty)\}$. This implies that the interval null hypotheses H_0 is specified as $H_0 : \delta \in (-0.5, -0.2] \cup [0.2, 0.5)$, $H_0 : \delta \in (-\infty, 0.8] \cup [0.8, \infty)$ or $H_0 : \delta \in (-\infty, 0.8] \cup [0.8, \infty)$. The recommended wide prior was used for all simulations, which is detailed later in the prior sensitivity analysis.

In total, the Gibbs sampler should stabilize around the true effect size δ.[9]

15.3.1 Results

The upper row of Figure 15.2 shows the results for small effect sizes. The two left plots show the results for $n = 100$ and $n = 200$ observations in each group. It is clear that the 95%-HPDs in both cases fluctuate strongly, indicating that anything from no effect to a medium effect is possible. The two right plots of the upper row show the results when increasing to $n = 300$ and $n = 700$ observations per group. The 95%-HPDs get narrower and stabilize inside the ROPE. While for $n = 300$ there are still some outliers, for $n = 700$ all HPDs have concentrated inside the ROPE of a small effect – that is $PMP(\delta_{MPE}) = 1$ (100%) for all iterations – and the estimates δ_{MPE} (blue points) have already converged closely to the true effect size indicated by the solid black line. The necessary sample size for this precision is not small, but the setting of a small effect requires a large sample size to be detected. Also, the applied criterion is very strict in the sense that it requires $PMP(\delta_{MPE}) = 1$, which means the entire posterior distribution needs to be located inside the ROPE. Less strict requirements like $PMP(\delta_{MPE}) = 0.95$ will require smaller sample sizes.

The middle row of Figure 15.2 shows the results for medium effect sizes. The two left plots show the result for $n = 100$ and $n = 200$ observations in each group for a medium effect size. Increasing the sample size to $n = 400$ and $n = 600$ leads to the

[9] Here, balanced groups are used, but unbalanced groups could also be treated easily by setting $N_1(S)$ and $N_2(S)$ accordingly, as described above.

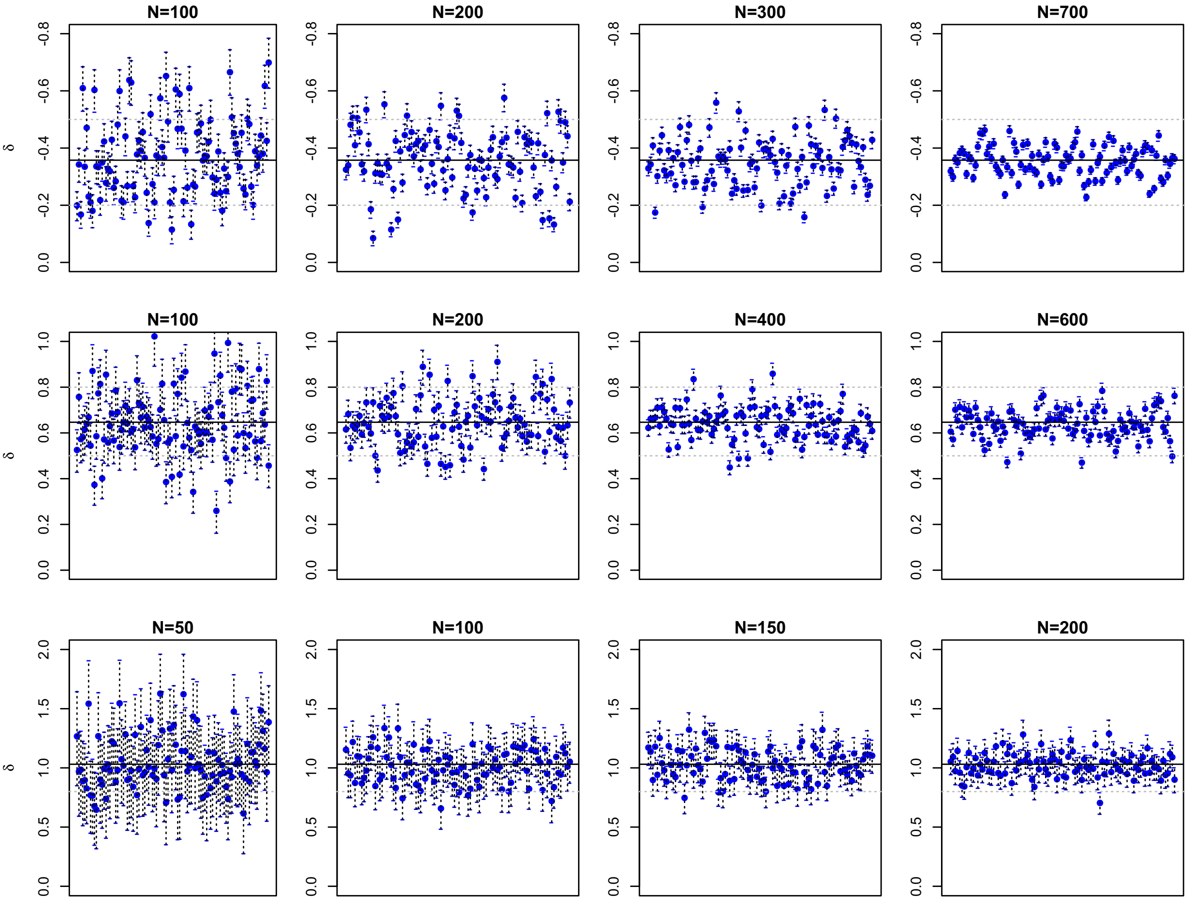

Figure 15.2: Posterior Means δ_{MPE} and 95% credible-intervals for δ_{MPE} for 100 datasets consisting of sample sizes $2n$, with n observations in each group; dotted lines represent the ROPE boundaries; *upper row*: Small effect size; *middle row*: Medium effect size; *lower row*: Large effect size

results shown in the two right plots. These figures show that even for sample sizes of $n = 100$ in both groups, no 95% HPD lies completely inside the ROPE $(-0.2, 0.2)$ around $\delta_0 = 0$ of no effect, indicating that while the size of the effect may still not be estimated accurately, a null hypothesis of no effect $\delta_0 = 0$ could always be rejected when using sample sizes of at least $n = 100$ in each group and the underlying effect has medium size. When it comes to precisely estimating the size of the effect, larger sample sizes similar to those needed to detect small effect sizes are necessary, as shown by the right plots of the second row.

The lower row of Figure 15.2 shows the results for large effect sizes. About $n = 50$ observations in each group suffice to produce δ_{MPE} and $PMP(\delta_{MPE})$ which estimate small to large effects, and thereby reject a null hypothesis of no effect, while about $n =$ 150 to $n = 200$ seem reasonable to precisely estimate a large effect size. When using δ_{MPE} (blue points) as an estimator for δ, sample sizes of $n = 200$ produce an estimate close to the true effect size, which in this case was $\delta_0 = 1.030723$.

15.3.2 Controlling the type I error rate

In frequentist NHST, the Neyman-Pearson theory aims at controlling the type I error rate α, which is the probability to reject the null hypothesis H_0 falsely, when indeed it is correct. In the setting of the two-sample Bayesian t-test this equals the rejection of $H_0 : \delta \in (-0.2, 0.2)$ although the true effect size $\delta_0 \in (-0.2, 0.2)$. Following Cohen (1988), an effect is considered small if the effect size is at least $|\delta| \geq 0.2$, so effect sizes in the interval $(-0.2, 0.2)$ can be considered as noise, or *practically equivalent to zero*. Therefore, a ROPE of $(-0.2, 0.2)$ is set around the null value $\delta_0 = 0$ to compare the type I error rate of the proposed method against the standard frequentist NHST solution, Welch's t-test. That is, the interval hypothesis $H_0 : \delta \in (-0.2, 0.2)$ is tested against its alternative $H_0 : \delta \notin (-0.2, 0.2)$. Again, 100 datasets of different sample sizes are simulated where the true effect size δ_0 is set to zero. The Gibbs sampler should produce a posterior distribution of δ which concentrates inside the ROPE, so that the null hypothesis $H : \delta_0 = 0$ is α-accepted. To facilitate comparison with the frequentist solution, α is set to $\alpha = 0.95$, see Definition 15.6. Thus, if the 95%-HPD interval lies (entirely) outside the ROPE $R := (-0.2, 0.2)$, this equals α-rejection for $\alpha = 0.95$, or in frequentist terms the rejection of the null hypothesis $H_0 : \delta_0 = 0$ of no effect and therefore the commitment of a type I error. From the perspective of the proposed Bayesian two-sample t-test this implies that the interval hypothesis $H_0 : \delta \in (-0.2, 0.2)$ is falsely rejected. The following two definitions formalize the type I and II error building on the concept of α-rejection for the proposed Hodges-Lehmann test:

Definition 15.9 (α type I error). An α type I error happens if the true parameter value $\delta_0 \in H$, with $H \subset R$ for a ROPE $R \subset \Theta$, but H is α-rejected for α.

Definition 15.10 (α type II error). An α type II error happens if the true parameter value $\delta_0 \notin H$ and $\delta_0 \notin R$, with $H \subset R$ for a ROPE $R \subset \Theta$, but H is α-accepted for α.

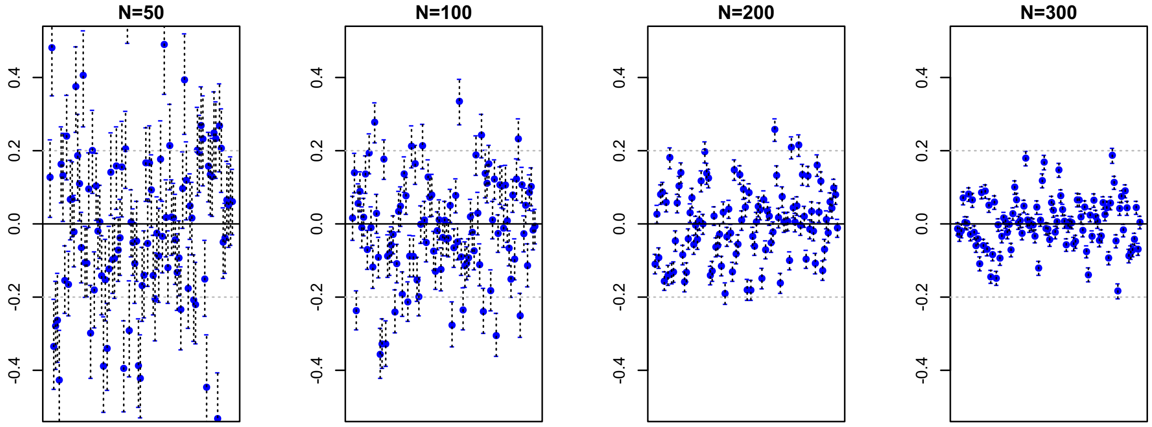

Figure 15.3: Posterior means δ_{MPE} and 95% credible-intervals for δ for 100 datasets consisting of different sample sizes $2n$, with n observations from a $\mathcal{N}(148.3, 1.34)$ distribution and n observations from a $\mathcal{N}(148.3, 2.04)$ distribution; dotted lines represent the ROPE $(-0.2, 0.2)$ of no effect size around $\delta = 0$; posterior distributions are based on 10000 iterations of the Gibbs sampler with a burnin of 5000 iterations

The left plot in Figure 15.3 shows the results of 100 datasets of size $n = 50$ in each group. The first group was simulated as $\mathcal{N}(148.3, 1.34)$, and the second group was simulated as $\mathcal{N}(148.3, 2.03)$. The true effect size is

$$\delta_0 = \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} = 0$$

The blue points represent δ_{MPE} and the blue dotted lines the 95%-HPDs of the posterior of δ. While the estimates fluctuate strongly for $n = 50$, increasing sample size in each group successively to $n = 200$ as shown by the progression of the plots from left to right shows that false-positive results – α type I errors with $\alpha = 0.95$ – get completely eliminated for sufficiently large sample size. The right plot with sample size $n = 300$ shows that no α type I error with $\alpha = 0.95$ occurs anymore. Also, δ_{MPE} stabilizes around the true value $\delta_0 = 0$ of no effect, indicating its convergence to the true effect size δ.

The simulations show that the α type I error rate converges to zero when the sample size is increased. The number of credible intervals which lie partly inside and partly outside the ROPE $R = (-0.2, 0.2)$ (or equivalently, outside the interval hypothesis $H_0 : \delta \in (-0.2, 0.2)$) decreases to zero. In contrast, p-values are uniformly distributed under the null hypothesis, so that no matter what size the samples in both groups are, in the long-run one will still obtain α% (most often 5%) type I errors. Conducting Welch's t-tests will thus inevitably lead to a type I error rate of 5% for the corresponding test level. If the sample size is at least $n = 200$ in each group, the proposed Bayesian t-test together with δ_{MPE} and $PMP(\delta_{MPE})$ performs better with respect to control the α type I error rate. From a theoretical perspective, it is of course of interest for which values of α this fact does hold, and indeed, using the two generalized types of type I and II errors, it can also be shown that the number of type I (type II) errors converges to zero for any $\alpha \neq 0$, when a correct (incorrect) ROPE is chosen:

Theorem 15.11. For the Bayesian two-sample t-test model, the probability of making an α type I error for any $\alpha \neq 0$ converges to zero for $n \to \infty$ for any correct ROPE R around the hypothesis H which makes a statement about the unknown parameter δ, where n is the sample size. Also, the probability of making an α type II error for any $\alpha \neq 0$ converges to zero for $n \to \infty$ for any incorrect ROPE R.

Proof. See Appendix A.5. □

The implications of Theorem 3 are that if a correct ROPE R is chosen (that is, the selected interval hypothesis contains the true parameter value), then the probability of making a α type I error will eventually become zero for large enough sample size n. As a special case, this implies that when the ROPE R includes the true parameter δ_0, eventually the hypothesis H will be α-accepted for $\alpha = 1$, that is, accepted, because the entire posterior concentrates inside the interval hypothesis. If on the other hand an incorrect ROPE is selected, which does not include the true parameter δ_0, then eventually the probability of making a α type II error – that is, accepting H although $\delta_0 \notin H$ – will converge to zero for growing sample size n.

15.3.3 Prior sensitivity analysis

Section 15.2.3 detailed the independence prior used in the model, and of specific interest is of course the influence of this prior on the results produced by the proposed Bayesian t-test which implements the Hodges-Lehmann paradigm. Therefore, three different hyperparameter settings were selected to resemble a wide, medium and narrow prior, where the shrinkage effect on the standard deviations $\sigma_k^2, k = 1,2$ caused by the inverse Gamma prior $IG(c_0, C_0)$ on σ_k^2 increases with the prior getting narrower (that is, σ_i^2 is shrunken towards zero). The same applies for the normal prior $\mathcal{N}(b_0, B_0)$ on the means $\mu_k, k = 1,2$. The following hyperparameters were chosen for the three different settings: For the wide prior, $b_0 := \bar{x}$ and $B_0 := 10 \cdot s^2(x)$ where $\bar{x}$ and $s^2(x)$ are the complete sample mean and variance. c_0 and C_0 were selected as both 0.01 for the wide prior, implying fatter tails of the inverse Gamma prior than in the medium or narrow prior. For the medium prior, B_0 was decreased to $5 \cdot s^2(x)$, and c_0 and C_0 decreased to 0.1 both. For the narrow prior finally, $B_0 := s^2(x)$ and $c_0 = C_0 = 1$, which is the most informative of all three priors.

Subsequently, 100 datasets with $n = 100$ observations in each group were simulated, where the first group was generated as $\mathcal{N}(0,1)$ and the second as $\mathcal{N}(1,1)$. The Gibbs sampler was run for 10000 iterations with a burn-in of 5000 once for each prior on each dataset. Figure 15.4 shows the results of the simulations. Here, the resulting posterior densities of δ_{MPE} are overlayed in Figure 15.4, and it becomes clear that the wide and medium prior do result in barely differing posteriors. When using the narrow prior the shrinkage moves the posterior slightly towards smaller values of δ. The lower plot in Figure 15.4 also shows the posterior distributions of differences between means obtained from the three priors: The left hand plot shows the posterior distribution of differences between δ_{MPE} obtained via a wide and a medium prior. The middle plot

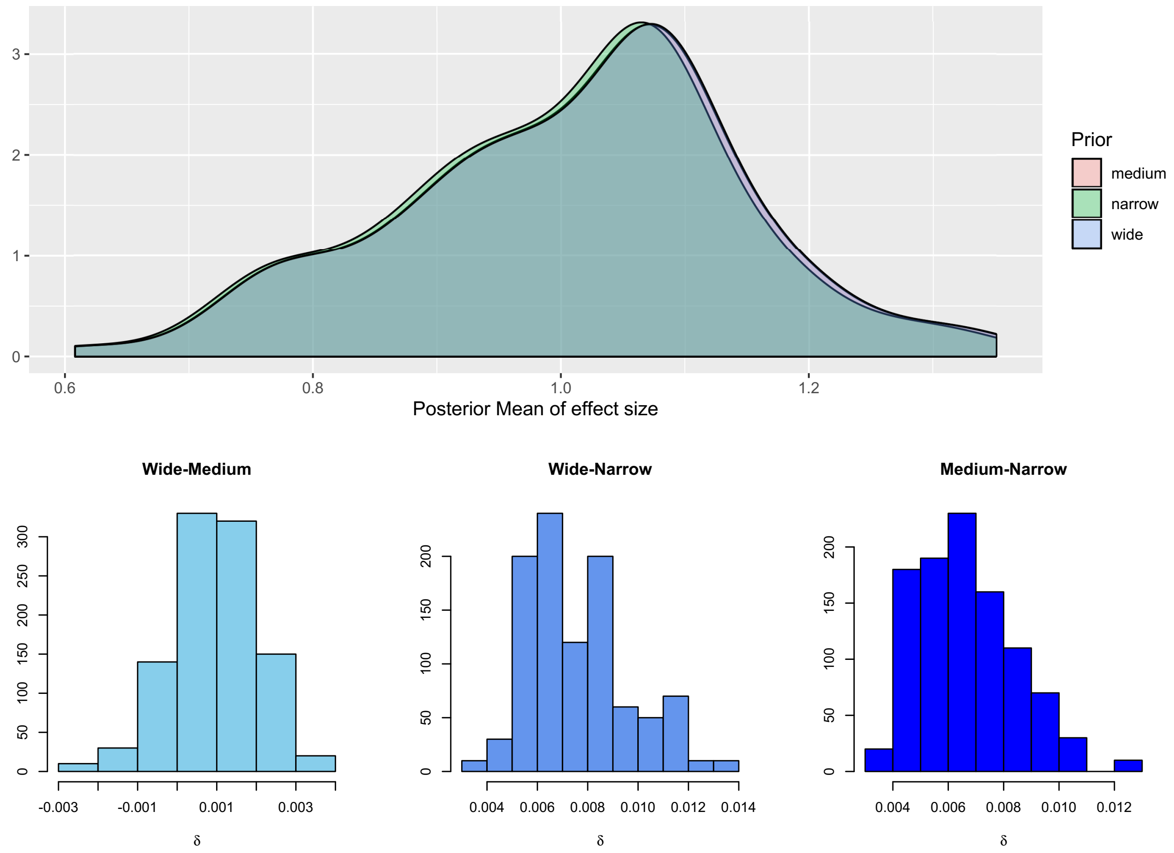

Figure 15.4: Prior sensitivity analysis for the $\mathcal{N}(b_0, B_0)$ prior on the means μ_k and inverse Gamma prior $IG(c_0, C_0)$ on the variances σ_k^2, $k = 1,2$ for 100 datasets with first group simulated as $\mathcal{N}(0,1)$ and the second as $\mathcal{N}(1,1)$

shows the posterior distribution of differences between δ_{MPE} obtained via a wide and a narrow prior, and the right hand plot the posterior distribution of differences between δ_{MPE} obtained via a medium and a narrow prior. The results show that in all cases the differences are of tiny magnitude, indicating that the proposed t-test is quite robust to the prior hyperparameter selection. Of course, it can happen that δ_{MPE} will be drawn towards smaller values when switching from the wide to the narrow prior, but the posterior mass percentage supporting a large effect will not vary much as shown by the nearly identical resulting posterior densities in Figure 15.4, which is a strength of the continuous quantification through $PMP(\delta_{MPE})$. Based on the sensitivity analysis all three hyperparameter settings differ only slightly, and therefore the wide prior seems suitable for most applications, as it places itself between the other two priors.

15.4 Discussion

The theory presented in this chapter proposed a new Bayesian two-sample t-test which focusses on effect size estimation and implements the Hodges-Lehmann approach by replacing a traditional point null hypothesis with an interval hypothesis. Following the proposal of a shift from hypothesis testing to estimation under uncertainty, a Gibbs sampler was constructed for the Gaussian mixture model that underlies the proposed test. Statistical inference is performed for the effect size δ, which is the quantity of interest in most biomedical research like clinical trials. Also, the dichotomy of the ROPE decision rule of Kruschke and Liddell (2018b) was resolved by introducing the mean probable effect size δ_{MPE} as an estimator of δ, combined with the posterior mass percentage $PMP(\delta_{MPE})$, a continuous measure which quantifies the support for the evidence suggested by δ_{MPE}. In summary, the proposal shows that a Hodges-Lehmann test for an interval hypothesis is more reasonable than a test of a point null hypothesis.

Theorems 15.8 and 15.11 showed that the use of the proposed test leads to a consistent estimation procedure which shifts from hypothesis testing to estimation under uncertainty in the spirit of the Hodges-Lehmann paradigm which was proposed first by Hodges and Lehmann (1954) and advocated later by Rao and Lovric (2016). Under any correct ROPE R, the number of introduced α type I errors converges to zero a.s. under any prior π on δ, while under any incorrect ROPE (which means the ROPE picked by the researcher does not cover the true parameter value δ_0), the number of introduced α type II errors converges to zero a.s. under any prior π on δ. The ROPE models an interval hypothesis in the Hodges-Lehmann approach here. Together, these properties and the introduced concept of α-rejection make the proposed method an attractive alternative to existing solutions via p-values or the Bayes factor for a precise point null hypothesis.

One important limitation of the approach is that the results depend on the chosen priors for μ_i and σ_i. Although quite robust, especially the choice of the inverse Gamma prior for the variances may be questioned. A sensitivity analysis using different priors for the variances could be useful, and one remedy which allows changing the priors would be to switch to different sampling techniques, for example Hamiltonian Monte Carlo in Stan (Carpenter et al., 2017). Also, the speed of convergence to an entire elimination of type I errors which depends on the sample size may be slow, although the simulation results are promising. However, while here a Gibbs sampler was derived, the availability of modern MCMC methods allows to apply the idea presented in this chapter to a wide variety of models, compare Chapter 9 and Chapter 13. Thus, one could implement a similar test for the regression coefficients in the survival model considered in Chapter 13 by using Stan without the need to derive a Gibbs sampler. The same holds for other tests like the ones considered in Chapter 12.

The proposed method is thus widely applicable and can easily be generalized to other statistical models. Also, in light of the theoretical results, it may be helpful in improving the reproducibility of biomedical research, especially by reducing the number of false-positive results, which is one of the biggest problems of the biomedical sciences, see McElreath and Smaldino (2015). Finally, it should be noted that the approach is different both from Bayes factors and p-values, as a precise hypothesis is replaced with an interval hypothesis. The proposed method has the benefit that hypothesis testing is performed with respect to a more realistic hypothesis, and effect size estimation under uncertainty is built into the procedure from the start. Therefore, the main advantage may be seen in the shift towards interval hypothesis testing in the spirit of Hodges and Lehmann (1954) and simultaneous effect size estimation.

Chapter 16

Revisiting the Replication Crisis

The quiet statisticians have changed our world – not by discovering new facts or technical developments, but by changing the ways we reason, experiment and form our opinions about it.

Ian Hacking
Trial by number

Statistical hypothesis testing has become the central method for the judgement of empirical findings in the biomedical, social and cognitive sciences. As discussed in Chapter 1, in recent years, the ongoing problems with null hypothesis significance testing and p-values have shown that the underlying paradigm for quantifying statistical evidence about a research hypothesis is highly problematic, and the situation has been termed a scientific replication crisis.

The Evolution of Statistical Hypothesis Testing

In this thesis, the evolution of statistical hypothesis testing was reconstructed and it was shown that major problems of the replication crisis are due to the misapplication or application of theories outside of their intended contexts.

In Part I it was shown that various recently observed problems with the reproducibility of research findings can be attributed to the underlying statistical theory: The inconsistent hybrid approach which emerged out of Fisher's theory of significance tests and the Neyman-Pearson theory that itself was created for applications like quality control can be seen as a primary reason of the unsatisfactory status quo of statistical hypothesis testing in science. The blend of p-values and test levels is not allowed and was neither intended by Fisher nor by Neyman and Pearson. Part II contrasted the development of these frequentist theories of statistical hypothesis testing with the evolution of Bayesian approaches, in particular, the Bayes factor. It was shown that these approaches are much more in the veins of a theory for testing statistical hypotheses in scientific research. However, computational obstacles have historically prevented a more widespread use of Bayesian hypothesis tests as shown in Part III, and the recent advent of Markov-Chain-Monte-Carlo and Hamiltonian-Monte-Carlo algorithms have solved this main hurdle effectively now. The philosophical considerations in Part IV showed that Bayes theorem can be interpreted from a philosophy of science perspective as a statistical implementation of probabilistic enumerative induction. The following axiomatic analysis in Part IV demonstrated that the currently experienced replication problems can be attributed in large parts to the axiomatic foundations of statistical inference. The violation of certain principles of statistical inference is highly undesirable from a scientific perspective, and the conflict of frequentist hypothesis testing theories with the likelihood principle presents a profound problem for application of these theories in scientific contexts. In contrast, Bayesian theories of hypothesis testing are more adequate for application in scientific research contexts due to their coherence with the likelihood principle, and the resulting stopping rule and censoring principle. Thus, Bayesian hypothesis tests present an attractive alternative to mitigate the replication problems in biomedical research from a purely axiomatic point of view.

Bayesian Statistical Solutions to the Replication Crisis

Based on the axiomatic foundations discussed in Chapter 11, it was shown that a shift towards robust Bayesian analysis is required to improve the reproducibility of research. The new results and solutions presented in Part V demonstrated that for the majority of statistical models in the biomedical and cognitive sciences robust Bayesian hypothesis tests are available, compare Chapter 12. Also, it was shown in Chapter 13 that even complex statistical models are tractable through the use of modern Hamiltonian-Monte-Carlo algorithms, and the Bayes factor can be obtained in such situations based only on the posterior MCMC sample. Also, new results in Chapter 14 demonstrated that the implicit error control of Bayesian hypothesis tests is comparable to frequentist tests based on p-values, and that a variety of Bayesian evidence measures attains reasonable type I error control and power in practice. This provides a further justification for Bayesian hypothesis tests and allows to select between competing Bayesian evidence measures for hypothesis testing. Furthermore, robust Bayesian hypothesis tests can be interpreted as slightly more conservative than frequentist tests based on p-values, and as false-positive results are among the biggest problems in the replication crisis (Smaldino and McElreath, 2016; McElreath and Smaldino, 2015), a shift towards these Bayesian hypothesis tests is an attractive solution, compare Chapter 14.

A Paradigm Change towards Hodges-Lehmann Tests

Next to shifting towards robust Bayesian hypothesis tests, an important step to improve the reproducibility of science is a shift towards what could be called the Hodges-Lehmann paradigm. Statistical hypothesis testing has become a dominating inferential procedure in a wide range of sciences (Howie, 2002), and since the foundational contributions of Fisher, Neyman and Pearson to frequentist hypothesis testing, and Wrinch, Jeffreys and Haldane to Bayesian hypothesis testing, the focus on precise point null hypotheses has barely changed. The frequentist hybrid approach has been adopted by more and more researchers (Halpin and Stam, 2006) to test what is today well-known as a point null hypothesis. While Bayesian hypothesis tests are becoming more popular, the majority of these also focusses on point null hypotheses, which is primarily due to the fact that the underlying mathematical models were easier to handle when

Jeffreys' and others developed first Bayesian hypothesis tests based on the Bayes factor, compare Chapter 7. However, the availability of modern MCMC algorithms allows to transition easily towards more realistic interval hypotheses as shown in Chapter 15. The application of point null hypothesis testing has been debated extensively since its investigation, both by statisticians and non-statisticians. Early critics include Buchanan-Wollaston (1935), and in the last decades the proposal of an entire ban of the method has grown in popularity (Lindley, 1972; Hunter, 1997; Gigerenzer, 2004). As shown in Part I, the root of all problems is often declared to be the p-value in Fisher's theory of significance testing. Still, one of the most important criticisms of statistical hypothesis testing includes the relevance and validity of a precise point null hypothesis for scientific research:

> "The decision whether or not to formulate an inference problem as one of testing a precise null hypothesis centers on assessing the plausibility of such an hypothesis. Sometimes this is easy, as in testing for the presence of extrasensory perception, or testing that a proposed law of physics holds. Often it is less clear. In medical testing scenarios, for instance, it is often argued that any treatment will have some effect, even if only a very small effect, and so exact equality of effects (between, say, a treatment and a placebo) will never occur."
> Berger et al. (1994, p. 145)

In a vast range of research in the biomedical and cognitive sciences the test of a precise point null hypothesis like $H_0 : \delta = 0$ is only of limited use. In exploratory research, there may be no precise point null hypothesis available. Also, in settings with limited measuring precision or moderate measurement error a precise null hypothesis will eventually be rejected because even if the null hypothesis is true data cannot be measured with infinite precision. Furthermore, researchers are often less interested in rejecting or accepting a precise point null hypothesis, but in the size of an observed effect and its relevance from a scientific perspective in the application context. Researchers often want to know whether a parameter value is located inside or outside some boundaries which separate relevant from negligible effects. Rouder et al. argued:

> "It is reasonable to ask whether hypothesis testing is always necessary. In many ways, hypothesis testing has been employed (...) too often and too hastily. (...) As a rule of thumb, hypothesis testing should be reserved for those cases in which the researcher will entertain the null as theoretically interesting and plausible, at least *approximately*."
> Rouder et al. (2009, p. 235)

Hodges and Lehmann (1954) discussed the validity of statistical hypotheses more than half a century ago, and their proposal was widely ignored over the course of time. Although the debate about the validity of point null hypotheses never came to an end, interest in it was only moderate which probably was also due to computational obstacles and the fact that for moderate amounts of data, a point null hypothesis provides a reasonable approximation of a small interval hypothesis (Berger and Delampady, 1987). However, in times where big data are becoming abundant, sample sizes grow larger and data sets are often high-dimensional, this argument does not hold anymore.

Thus, Hodges and Lehmann[1] identified one major flaw in the appropriateness of the status quo for scientific contexts when proposing the test of small interval hypotheses

[1] Erich Leo Lehmann was PhD student of Jerzy Neyman.

more than half a century ago: Point null hypotheses are of questionable use and were chosen primarily because they were mathematically easy to handle. Alan Birnbaum[2] provided the axiomatic analysis which ultimately demonstrated the severe problems of frequentist hypothesis tests for scientific research as discussed in Chapter 11. A shift toward robust Bayesian hypothesis test presents an attractive solution as shown in Chapter 11. However, most of these Bayesian tests also concentrate on testing precise hypotheses. Recently, Rao and Lovric (2016)[3] have argued for a shift towards the Hodges-Lehmann paradigm as a more realistic approach to statistical hypothesis testing. Section 15 provided a first step towards the Hodges-Lehmann paradigm, and the approach presented there can easily be generalized to a wide range of statistical models as long as the posterior distribution is obtainable via Markov-Chain-Monte-Carlo. However, more work is required to transition from testing precise hypotheses towards testing small interval hypotheses in the Bayesian approach. Shifting towards Bayesian Hodges-Lehmann tests can be seen as the next important step in the evolution of statistical hypothesis testing, and to increase the reproducibility of science. Solutions to the replication crisis should thus focus on establishing robust Bayesian hypothesis tests as the standard in scientific research and transitioning towards Bayesian Hodges-Lehmann tests. Although these Bayesian statistical solutions are only a piece in the puzzle of the solution to the replication crisis in the biomedical sciences, it may be worthwhile to keep in mind Alan Birnbaum's words when developing new statistical hypothesis tests from a Hodges-Lehmann perspective in the future:

> The only theories which are formally complete, and of adequate scope for treating statistical evidence and its interpretation in scientific research contexts, are Bayesian.
>
> Alan Birnbaum
> *The anomalous concept of statistical evidence*, 1964

[2] Alan Birnbaum was PhD student of Erich Leo Lehmann.

[3] Calyampudi Radhakrishna Rao was PhD student of Ronald Fisher.

Appendices

Appendix A

Proofs and Derivations for Chapter 15

A.1 Derivation of the single-block Gibbs sampler

This appendix provides the basis for the derivation of the joint posterior and full conditionals for the single-block Gibbs sampler in Appendix A.2.

Bayesian parameter estimation for known allocations

In this section, for the setting when the allocations S are known the posterior distribution of μ_k, σ_k^2 given the complete data S, y are derived. The weights $(\eta_1, ..., \eta_K)$ are known in this case because for every observation $y_i \in (y_1, ..., y_N)$ it is known to which group y_i belongs, that is, the quantities $S_i = k, k \in \{1, ...K\}$ are available for all $i \in \{1, ..., N\}$. In the setting of a two-sample t-test between two groups with equal sample sizes $n_1 = n_2$ and $n_1 + n_2 = N$, the belonging of an observation y_i to its group is known for all observations $i \in \{1, ..., N\}$. The underlying data generating process therefore can be assumed to consist of a mixture of $K = 2$ components with weights $\eta_1 = \eta_2 = 0.5$. This makes inference in the mixture model much easier compared to the case when both the weights $(\eta_1, ...\eta_K)$ as well as the component parameters $(\mu_1, ..., \mu_K)$ and $(\sigma_1^2, ..., \sigma_K^2)$ are unknown. To conduct inference about the unknown parameters, the necessary group-specific quantities are the number $N_k(S)$ of observations in group k, the within-group variance $s_{y,k}^2(S)$ and the group mean $\bar{y}_k(S)$:

$$\begin{aligned}
N_k(S) &:= |\{i : S_i = k\}| \\
\bar{y}_k(S) &:= \frac{1}{N_k(S)} \sum_{i:S_i=k} y_i \\
s_{y,k}^2(S) &:= \frac{1}{N_k(S)} \sum_{i:S_i=k} (y_i - \bar{y}_k(S))^2
\end{aligned}$$

where $|\cdot|$ denotes the cardinality of a set. These quantities depend on S, so the classification of the observation y_i to the component $S_i = k$ needs to be available. When $S_i = k$ for an observation y_i holds, then the observational model for observation y_i is $\mathcal{N}(\mu_k, \sigma_k^2)$ and y_i contributes to the complete-data likelihood $p(y|\mu, \sigma^2, S)$ by a factor of

$$\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(y_i - \mu_k)^2\right)$$

Taking into account all observations $y_1, ..., y_N$, the complete data likelihood function can be written as

$$p(y|\mu, \sigma^2, S) = \prod_{k=1}^{K} (\frac{1}{2\pi\sigma_k^2})^{N_k(S)/2} \cdot \exp(-\frac{1}{2} \sum_{i:S_i=k} \frac{(y_i - \mu_k)^2}{\sigma_k^2})$$

The complete-data likelihood is a product of K components, of which each summarizes the information about the i-th group, $i \in \{1, ..., K\}$. These K factors are then combined in a Bayesian analysis with a prior. Interest lies in the posterior of both μ_k, σ_k^2, and first two different cases are considered, which will eventually lead to the solution of the joint posterior for μ_k and σ_k^2. In the first case, when the variance σ_k^2 is fixed, the complete-data likelihood function as a function of μ is the kernel of a univariate normal distribution (Held and Sabanés Bové, 2014, p. 181). Choosing a $\mathcal{N}(b_0, B_0)$-distribution as a conjugate prior, the posterior density of μ_k given σ_k^2 and the $N_k(S)$ observations in group k can be derived as

$$\begin{aligned} p(\mu_k|\sigma_k^2, S, y) &\propto p(y|\mu_k, \sigma_k^2, S) \cdot p(\mu_k, \sigma_k^2, S) \\ &\overset{(1)}{=} p(y|\mu_k, \sigma_k^2, S) \cdot p(\mu_k) \\ &= (\frac{1}{2\pi\sigma_k^2})^{N_k(S)/2} \cdot \exp(-\frac{1}{2} \sum_{i:S_i=k} \frac{(y_i - \mu_k)^2}{\sigma_k^2}) \\ &\cdot \frac{1}{\sqrt{2\pi B_0}} \exp(-\frac{1}{2} \frac{(\mu_k - b_0)^2}{B_0}) \end{aligned} \tag{A.1}$$

where in (1) the fact that σ_k^2 is assumed to be given and the allocations S are known constants, too, was used. For a sample of size n from a $\mathcal{N}(\mu, \sigma)$ distribution with known variance σ^2, a standard Bayesian analysis yields, see e.g. (Held and Sabanés Bové, 2014, p. 181), that the likelihood

$$L(\mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right)$$

when combined with a prior $\mu \sim \mathcal{N}(\nu, \tau^2)$ leads to the posterior

$$\mu|x \sim \mathcal{N}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \cdot \left(\frac{n\bar{x}}{\sigma^2} + \frac{\nu}{\tau^2}\right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right) \tag{A.2}$$

Substituting $\nu = b_0$ and $\tau^2 = B_0$ for the prior of μ as well as μ_k for μ and σ_k^2 for σ^2 in the likelihood, the posterior $p(\mu_k|\sigma_k^2, S, y)$ in Equation (A.1) based on Equation (A.2) becomes

$$p(\mu_k|\sigma_k^2, S, y) \sim \mathcal{N}\left(\left(\frac{N_k(S)}{\sigma_k^2} + \frac{1}{B_0}\right)^{-1} \cdot \left(\frac{N_k(S)\bar{y}_k(S)}{\sigma_k^2} + \frac{b_0}{B_0}\right), \left(\frac{N_k(S)}{\sigma_k^2} + \frac{1}{B_0}\right)^{-1}\right) \tag{A.3}$$

By Equation (A.3), the posterior can be written as

$$\mu_k|\sigma_k^2, S, y \sim \mathcal{N}(b_k(S), B_k(S))$$

with

$$B_k(S)^{-1} = B_0^{-1} + \sigma_k^{-2} N_k(S) \tag{A.4}$$
$$b_k(S) = B_k(S)(\sigma_k^{-2} N_k(S)\bar{y}_k(S) + B_0^{-1} b_0) \tag{A.5}$$

where for an empty group k the term $N_k(S)\bar{y}_k(S)$ is defined as zero. In the second case, if the mean μ_k is regarded as fixed, the complete-data likelihood as a function of σ_k^2 is the kernel of an inverse Gamma density. Choosing the conjugate inverse Gamma prior $\sigma_k^2 \sim IG(c_0, C_0)$, a standard Bayesian analysis – for details, see Held and Sabanés Bové (2014, p. 181) – yields the posterior of $\sigma_k^2|\mu_k, S, y$ as

$$p(\sigma_k^2|\mu_k, S, y) \sim IG(c_k(S), C_k(S)) \tag{A.6}$$

with

$$c_k(S) = c_0 + \frac{1}{2} N_k(S) \tag{A.7}$$
$$C_k(S) = C_0 + \frac{1}{2} \sum_{i:S_i=k} (y_i - \mu_k)^2 \tag{A.8}$$

The case of interest here is when both μ_k and σ_k^2 are unknown, and in this case a closed-form solution for the joint posterior $p(\mu_k, \sigma_k^2|S, y)$ does exist only under specific conditions. That is, the prior variance of the mean $\mu_k$of group k must depend on σ_k^2 through the relation $B_{0,k} = \frac{\sigma_k^2}{N_0}$, where N_0 is a newly introduced hyperparameter in the prior of μ_k, that is, the prior $\mu_k \sim \mathcal{N}(b_0, B_0)$ becomes $\mu_k \sim \mathcal{N}(b_0, \sigma_k^2/N_0)$. The joint posterior then can be rewritten as

$$\begin{aligned} p(\mu, \sigma^2|S, y) = p(\mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2|S, y) &\overset{(1)}{=} \prod_{k=1}^K p(\mu_k, \sigma_k^2|S, y) \\ &\overset{(2)}{=} \prod_{k=1}^K \underbrace{p(\mu_k|\sigma_k^2, S, y)}_{=:(A)} \cdot \underbrace{p(\sigma_k^2|S, y)}_{:=(B)} \end{aligned} \tag{A.9}$$

where (1) follows from the fact that the group parameters μ_k, σ_k^2 are assumed to be independent across groups and (2) follows from factorising the joint posterior as

$$\begin{aligned} p(\mu_k, \sigma_k^2|S, y) &= \frac{p(\mu_k, \sigma_k^2, S, y)}{p(S, y)} = \frac{p(\mu_k, \sigma_k^2, S, y) p(\sigma_k^2, S, y)}{p(S, y) p(\sigma_k^2, S, y)} = \frac{p(\mu_k, \sigma_k^2, S, y)}{p(\sigma_k^2, S, y)} \frac{p(\sigma_k^2, S, y)}{p(S, y)} \\ &= p(\mu_k|\sigma_k^2, S, y) \cdot p(\sigma_k^2|S, y) \end{aligned}$$

As the factors (A) and (B) in Equation (A.9) were already derived in Equation (A.3) and Equation (A.6) (Equation (A.6) still needs to be marginalized with respect to μ_k to match factor (B)) for arbitrary k, and the factor (A) of the posterior in Equation (A.9) is normal-distributed $\mathcal{N}(b_k(S), B_k(S))$ with parameters

$$B_k(S) \overset{(1)}{=} \frac{1}{B_0^{-1} + \sigma_k^{-2} N_k(S)} \overset{(2)}{=} \frac{1}{\sigma_k^{-2} N_0 + \sigma_k^{-2} N_k(S)} = \frac{1}{N_0 + N_k(S)} \sigma_k^2 \tag{A.10}$$

and

$$\begin{aligned} b_k(S) &\overset{(3)}{=} B_k(S)(\sigma_k^{-2}N_k(S)\bar{y}_k(S) + B_0^{-1}b_0) \\ &\overset{(4)}{=} B_k(S)(\sigma_k^{-2}N_k(S)\bar{y}_k(S) + \frac{N_0}{\sigma_k^2}b_0) \\ &\overset{(5)}{=} \frac{1}{N_0 + N_k(S)}\sigma_k^2(\sigma_k^{-2}N_k(S)\bar{y}_k(S) + \frac{N_0}{\sigma_k^2}b_0) \\ &= \frac{N_k(S)\bar{y}_k(S) + N_0 b_0}{N_0 + N_k(S)} \\ &= \frac{N_0}{N_k(S) + N_0}b_0 + \frac{N_k(S)}{N_k(S) + N_0}\bar{y}_k(S) \end{aligned} \tag{A.11}$$

where in (1) $B_k(S)^{-1} = B_0^{-1} + \sigma_k^{-2}N_k(S)$ from Equation (A.4) was used and in (2) the relation $B_0 = B_{0,k} = \frac{\sigma_k^2}{N_0}$, where N_0 is the newly introduced hyperparameter. In (3), Equation (A.5) was used, in (4) again the relation $B_{0,k} = \frac{\sigma_k^2}{N_0}$, in (5) the right-hand side of Equation (A.10) was substituted for $B_k(S)$ in Equation (A.11). The remaining term (B) of Equation (A.9) is the marginal posterior of σ_k^2, that is

$$(B) := p(\sigma_k^2|S,y) = \int p(\sigma_k^2|\mu_k, S, y)d\mu_k$$

and by integrating out μ_k, a standard Bayesian analysis shows that the marginal posterior of σ_k^2 is distributed as inverse Gamma $IG(c_k(S), C_k(S))$, where $c_k(S)$ is already given in Equation (A.7), and the parameter $C_k(S)$ in Equation (A.8) changes to

$$C_k(S) = C_0 + \frac{1}{2}\left(N_k(S)s_{y,k}^2(S) + \frac{N_k(S)N_0}{N_k(S) + N_0}(\bar{y}_k(S) - b_0)^2\right)$$

This is, because by combining an inverse-gamma prior with the normal likelihood with known mean yields an inverse-gamma posterior as shown above and marginalising this posterior for the variance yields exactly another inverse-gamma distribution with different parameters. For details see Held and Sabanés Bové (2014).

Application to the two-sample t-test – Derivation of the marginal and joint posterior distributions

In the case of the two-sample t-test, the general derivations above can be specified more precisely. For two groups, the mixture can be interpreted as a data generating process consisting of $K = 2$ components. The weights η_1 and η_2 are both equal to $1/2$ for equally sized groups, that is, $N = n_1 + n_2$ with n_1 being the sample size of group one and n_2 the sample size of group two and $n_1 = n_2$. Taking into account all observations $y_1, ..., y_N$, the complete data likelihood function can be written as

$$\begin{aligned}
p(y|\mu,\sigma^2,S) &= \prod_{k=1}^{2}(\frac{1}{2\pi\sigma_k^2})^{N_k(S)/2}\cdot\exp(-\frac{1}{2}\sum_{i:S_i=k}\frac{(y_i-\mu_k)^2}{\sigma_k^2}) \\
&= (\frac{1}{2\pi\sigma_1^2})^{N_1(S)/2}\cdot\exp(-\frac{1}{2}\sum_{i:S_i=1}\frac{(y_i-\mu_1)^2}{\sigma_1^2}) \\
&\cdot(\frac{1}{2\pi\sigma_2^2})^{N_2(S)/2}\cdot\exp(-\frac{1}{2}\sum_{i:S_i=2}\frac{(y_i-\mu_2)^2}{\sigma_2^2})
\end{aligned}$$

where $N_1(S) = N_2(S) = N/2$. The posteriors $p(\mu_k|\sigma_k^2, S, y)$ for $k = 1, 2$ in Equation (A.1) are then $\mathcal{N}(b_k(S), B_k(S))$-distributed with

$$B_k(S) = \frac{1}{N_0 + N_k(S)}\sigma_k^2 \tag{A.12}$$

$$b_k(S) = \frac{N_0}{N_k(S) + N_0}b_0 + \frac{N_k(S)}{N_k(S) + N_0}\bar{y}_k(S) \tag{A.13}$$

where also $N_1(S) = N_2(S) = N/2$ are half of total sample size and $\bar{y}_k(S)$ is the mean of group $k = 1, 2$. After choosing the conjugate prior $\mu_k \sim \mathcal{N}(b_0, B_0)$, these posteriors can be computed. The posteriors

$$\sigma_k^2|\mu_k, S, y \sim IG(c_k(S), C_k(S))$$

with

$$c_k(S) = c_0 + \frac{1}{2}N_k(S) \tag{A.14}$$

$$C_k(S) = C_0 + \frac{1}{2}\sum_{i:S_i=k}(y_i - \mu_k)^2 \tag{A.15}$$

for $k = 1, 2$ become

$$\sigma_1^2|\mu_1, S, y \sim IG\left(c_0 + \frac{1}{2}N_1(S), C_0 + \frac{1}{2}\sum_{i:S_i=1}(y_i - \mu_1)^2\right)$$

$$\sigma_2^2|\mu_2, S, y \sim IG\left(c_0 + \frac{1}{2}N_2(S), C_0 + \frac{1}{2}\sum_{i:S_i=2}(y_i - \mu_2)^2\right)$$

and again, after selecting a conjugate inverse-gamma prior $\sigma_k^2 \sim IG(c_0, C_0)$ for $k = 1, 2$, these posteriors are also completely determined. The necessary marginal posteriors for σ_k^2 for $k = 1, 2$ are then obtained, following the derivations in the above section, as

$$p(\sigma_1^2|S, y) \sim IG\left(c_0 + \frac{1}{2}N_1(S), C_0 + \frac{1}{2}\left(N_1(S)s_{y,1}^2(S) + \frac{N_1(S)N_0}{N_1(S) + N_0}(\bar{y}_1(S) - b_0)^2\right)\right)$$

and

$$p(\sigma_2^2|S, y) \sim IG\left(c_0 + \frac{1}{2}N_2(S), C_0 + \frac{1}{2}\left(N_2(S)s_{y,2}^2(S) + \frac{N_2(S)N_0}{N_2(S) + N_0}(\bar{y}_2(S) - b_0)^2\right)\right)$$

These marginal posteriors are completely determined, once c_0, C_0 is given by the selected prior $IG(c_0, C_0)$ and the group variances $s^2_{y,1}(S)$ and $s^2_{y,2}(S)$ are calculated. Again here, $N_1(S) = N_2(S) = N/2$ due to equal sizes of both groups, and the $\bar{y}_1(S)$ and $\bar{y}_2(S)$ are the means of the two groups. The joint posterior, which is the ultimate quantity of interest, then can be rewritten as

$$\begin{aligned}
p(\mu, \sigma^2|S, y) = p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y) = \prod_{k=1}^{2} p(\mu_k, \sigma_k^2|S, y) = \prod_{k=1}^{2} \underbrace{p(\mu_k|\sigma_k^2, S, y)}_{=:(A)} \cdot \underbrace{p(\sigma_k^2|S, y)}_{:=(B)} \\
= p(\mu_1|\sigma_1^2, S, y)p(\sigma_1^2|S, y) \cdot p(\mu_2|\sigma_2^2, S, y)p(\sigma_2^2|S, y)
\end{aligned}$$

A.2 Proof of Theorem 15.2 – Derivation of the full conditionals for the single-block Gibbs sampler

Proof. To make Gibbs sampling possible, the full conditionals of

$$p(\mu, \sigma^2|S, y) = p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y)$$

need to be derived. These are given as

$$\begin{aligned}
&p(\mu_1|\mu_2, \sigma_1^2, \sigma_2^2, S, y) \\
&p(\mu_2|\mu_1, \sigma_1^2, \sigma_2^2, S, y) \\
&p(\sigma_1^2|\mu_1, \mu_2, \sigma_2^2, S, y) \\
&p(\sigma_2^2|\mu_1, \mu_2, \sigma_1^2, S, y)
\end{aligned}$$

The first conditional distribution $p(\mu_1|\mu_2, \sigma_1^2, \sigma_2^2, S, y)$ is given as:

$$\begin{aligned}
p(\mu_1|\mu_2, \sigma_1^2, \sigma_2^2, S, y) &= \frac{p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, S, y)}{p(\mu_2, \sigma_1^2, \sigma_2^2, S, y)} \\
&= \frac{p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y)\cancel{p(S, y)}}{p(\mu_2, \sigma_1^2, \sigma_2^2|S, y)\cancel{p(S, y)}} \\
&\overset{(1)}{=} \frac{\prod_{k=1}^{K} p(\mu_k|\sigma_k^2, S, y)p(\sigma_k^2|S, y)}{p(\mu_2, \sigma_2^2|S, y)p(\sigma_1^2|S, y)} \\
&\overset{(2)}{=} \frac{p(\mu_1|\sigma_1^2, S, y)\cancel{p(\sigma_1^2|S, y)}\,\cancel{p(\mu_2|\sigma_2^2, S, y)}\,\cancel{p(\sigma_2^2|S, y)}}{\cancel{p(\mu_2|\sigma_2^2, S, y)}\,\cancel{p(\sigma_2^2|S, y)}\,\cancel{p(\sigma_1^2|S, y)}} \\
&= p(\mu_1|\sigma_1^2, S, y)
\end{aligned}$$

where in (1) the independence of σ_1^2 from μ_2, σ_2^2 was used, so $p(\mu_2, \sigma_1^2, \sigma_2^2|S, y) = p(\mu_2, \sigma_2^2|S, y) \cdot p(\sigma_1^2|S, y)$ holds, and $p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y) = p(\mu_1, \sigma_1^2|S, y)p(\mu_2, \sigma_2^2|S, y)$ uses the independence between groups. Also, $p(\mu_1, \sigma_1^2|S, y) = p(\mu_1|\sigma_1^2, S, y)p(\sigma_1^2|S, y)$ and $p(\mu_2, \sigma_2^2|S, y) = p(\mu_2|\sigma_2^2, S, y)p(\sigma_2^2|S, y)$. In (2), the factorization $p(\mu_2, \sigma_2^2|S, y) = p(\mu_2|\sigma_2^2, S, y) \cdot p(\sigma_2^2|S, y)$ was used.

The full conditional of μ_2 is given by $p(\mu_2|\mu_1,\sigma_1^2,\sigma_2^2,S,y)$, which is derived as

$$\begin{aligned}
p(\mu_2|\mu_1,\sigma_1^2,\sigma_2^2,S,y) &= \frac{p(\mu_2,\mu_1,\sigma_1^2,\sigma_2^2,S,y)}{p(\mu_1,\sigma_1^2,\sigma_2^2,S,y)}\\
&= \frac{p(\mu_2,\mu_1,\sigma_1^2,\sigma_2^2|S,y)\cancel{p(S,y)}}{p(\mu_1,\sigma_1^2,\sigma_2^2|S,y)\cancel{p(S,y)}}\\
&= \frac{\prod_{k=1}^{K} p(\mu_k|\sigma_k^2,S,y)p(\sigma_k^2|S,y)}{p(\mu_1,\sigma_1^2|S,y)p(\sigma_2^2|S,y)}\\
&= \frac{\cancel{p(\mu_1|\sigma_1^2,S,y)}\,\cancel{p(\sigma_1^2|S,y)}\,p(\mu_2|\sigma_2^2,S,y)\,\cancel{p(\sigma_2^2|S,y)}}{\cancel{p(\mu_1|\sigma_1^2,S,y)}\,\cancel{p(\sigma_2^2|S,y)}\,\cancel{p(\sigma_1^2|S,y)}}\\
&= p(\mu_2|\sigma_2^2,S,y)
\end{aligned}$$

where the reasoning is the same as in the derivation of the conditional distribution for μ_1. The conditional distribution of σ_1^2 is $p(\sigma_1^2|\mu_1,\mu_2,\sigma_2^2,S,y)$, which is derived as:

$$\begin{aligned}
p(\sigma_1^2|\mu_1,\mu_2,\sigma_2^2,S,y) &= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2,S,y)}{p(\mu_1,\mu_2,\sigma_2^2,S,y)}\\
&= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y)\cancel{p(S,y)}}{p(\mu_1,\mu_2,\sigma_2^2|S,y)\cancel{p(S,y)}}\\
&= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y)}{p(\mu_1,\mu_2,\sigma_2^2|S,y)}\\
&\overset{(1)}{=} \frac{\prod_{k=1}^{2} p(\sigma_k^2|\mu_k,S,y)\cdot p(\mu_k|S,y)}{p(\mu_1|S,y)\cdot p(\mu_2,\sigma_2^2|S,y)}\\
&\overset{(2)}{=} \frac{p(\sigma_1^2|\mu_1,S,y)\cdot\cancel{p(\mu_1|S,y)}\cdot\cancel{p(\sigma_2^2|\mu_2,S,y)}\cdot\cancel{p(\mu_2|S,y)}}{\cancel{p(\mu_1|S,y)}\cdot\cancel{p(\sigma_2^2|\mu_2,S,y)}\cdot\cancel{p(\mu_2|S,y)}}\\
&= p(\sigma_1^2|\mu_1,S,y)
\end{aligned}$$

where in (1) first the independence of parameters between both groups was used, that is, the independence of μ_1,σ_1^2 and μ_2,σ_2^2 and second (as a special case of this fact) the independence of μ_1 and μ_2,σ_2^2 was used. Therefore, $p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y) = p(\sigma_1^2,\mu_1|S,y)p(\sigma_2^2,\mu_2|S,y)$ holds, and as $p(\sigma_1^2,\mu_1|S,y) = p(\sigma_1^2|\mu_1,S,y)p(\mu_1|S,y)$ and $p(\sigma_2^2,\mu_2|S,y) = p(\sigma_2^2|\mu_2,S,y)p(\mu_2|S,y)$, it follows that

$$p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y) = \prod_{k=1}^{2} p(\sigma_k^2|\mu_k,S,y)\cdot p(\mu_k|S,y)$$

Also, $p(\mu_1,\mu_2,\sigma_2^2|S,y) = p(\mu_1|S,y)\cdot p(\mu_2,\sigma_2^2|S,y)$ because of the independence of μ_1 from μ_2,σ_2^2. In (2) the factorization $p(\mu_2,\sigma_2^2|S,y) = p(\sigma_2^2|\mu_2,S,y)\cdot p(\mu_2|S,y)$ was used.

The derivation of the conditional distribution $p(\sigma_2^2|\mu_1,\mu_2,\sigma_1^2,S,y)$ of σ_2^2 proceeds similarly as follows:

$$\begin{aligned}
p(\sigma_2^2|\mu_1,\mu_2,\sigma_1^2,S,y) &= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2,S,y)}{p(\mu_1,\mu_2,\sigma_1^2,S,y)} \\
&= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y)\cancel{p(S,y)}}{p(\mu_1,\mu_2,\sigma_1^2|S,y)\cancel{p(S,y)}} \\
&= \frac{p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y)}{p(\mu_1,\mu_2,\sigma_1^2|S,y)} \\
&\overset{(1)}{=} \frac{\prod_{k=1}^{2} p(\sigma_k^2|\mu_k,S,y)\cdot p(\mu_k|S,y)}{p(\mu_2|S,y)\cdot p(\mu_1,\sigma_1^2|S,y)} \\
&\overset{(2)}{=} \frac{\cancel{p(\sigma_1^2|\mu_1,S,y)}\cdot\cancel{p(\mu_1|S,y)}\cdot p(\sigma_2^2|\mu_2,S,y)\cdot p(\mu_2|S,y)}{p(\mu_2|S,y)\cdot\cancel{p(\sigma_1^2|\mu_1,S,y)}\cdot\cancel{p(\mu_1|S,y)}} \\
&= \frac{\cancel{p(\mu_2|S,y)}\cdot p(\sigma_2^2|\mu_2,S,y)}{\cancel{p(\mu_2|S,y)}} \\
&= p(\sigma_2^2|\mu_2,S,y)
\end{aligned}$$

where in (1) first the independence of parameters between both groups, that is, the independence of μ_1,σ_1^2 and μ_2,σ_2^2 and second (as a special case of this) the independence of μ_2 and μ_1,σ_1^2 was used. Therefore, $p(\sigma_1^2,\mu_1,\mu_2,\sigma_2^2|S,y) = \prod_{k=1}^{K} p(\sigma_k^2|\mu_k,S,y)\cdot p(\mu_k|S,y)$ holds and one also has $p(\mu_1,\mu_2,\sigma_2^2|S,y) = p(\mu_1|S,y)\cdot p(\mu_2,\sigma_2^2|S,y)$ because of the independence of μ from μ_2,σ_2^2. In (2) the factorization $p(\mu_1,\sigma_1^2|S,y) = p(\sigma_1^2|\mu_1,S,y)\cdot p(\mu_1|S,y)$ was used. In total, the full conditionals thus are given as follows:

$$\begin{aligned}
p(\mu_1|\mu_2,\sigma_1^2,\sigma_2^2,S,y) &= p(\mu_1|\sigma_1^2,S,y) \\
p(\mu_2|\mu_1,\sigma_1^2,\sigma_2^2,S,y) &= p(\mu_2|\sigma_2^2,S,y) \\
p(\sigma_1^2|\mu_1,\mu_2,\sigma_2^2,S,y) &= p(\sigma_1^2|\mu_1,S,y) \\
p(\sigma_2^2|\mu_1,\mu_2,\sigma_1^2,S,y) &= p(\sigma_2^2|\mu_2,S,y)
\end{aligned}$$

When using the independence prior, based on Appendix A.1, the full conditionals are therefore given as

$$\begin{aligned}
p(\mu_1|\mu_2,\sigma_1^2,\sigma_2^2,S,y) &= p(\mu_1|\sigma_1^2,S,y) \sim \mathcal{N}(b_1(S),B_1(S)) && \text{(A.16)} \\
p(\mu_2|\mu_1,\sigma_1^2,\sigma_2^2,S,y) &= p(\mu_2|\sigma_2^2,S,y) \sim \mathcal{N}(b_2(S),B_2(S)) && \text{(A.17)} \\
p(\sigma_1^2|\mu_1,\mu_2,\sigma_2^2,S,y) &= p(\sigma_1^2|\mu_1,S,y) \sim IG(c_1(S),C_1(S)) && \text{(A.18)} \\
p(\sigma_2^2|\mu_1,\mu_2,\sigma_1^2,S,y) &= p(\sigma_2^2|\mu_2,S,y) \sim IG(c_2(S),C_2(S)) && \text{(A.19)}
\end{aligned}$$

with $b_1(S), B_1(S), b_2(S), B_2(S), c_1(S), c_2(S), C_1(S)$ and $C_2(S)$ as specified in the Appendix A.1 in Equations (A.12), (A.13), (A.14) and (A.15), which completes the proof. □

A.3 Proof of Corollary 15.3

Proof. From standard MCMC theory, see e.g. Robert and Casella (2004), it follows that the full conditionals derived in Theorem 15.2 can be used to construct a Gibbs sampler, by iteratively updating each parameter via simulating step by step from the full conditionals (A.16) to (A.19). Using (A.16) to (A.19), this leads to the following Gibbs sampling algorithm:

1. Sample σ_k^2 in each group k, $k = 1, 2$ from an inverse Gamma distribution

$$IG(c_k(S), C_k(S))$$

   (which depends on μ_k).

2. Sample μ_k in each group k, $k = 1, 2$, from a normal distribution

$$\mathcal{N}(b_k(S), B_k(S))$$

   (which depends on σ_k^2).

where $B_k(S), b_k(S)$ and $c_k(S), C_k(S)$ are given by equations (A.12), (A.13), (A.14) and (A.15). The convergence to the joint posterior $p(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2|S, y)$ then follows then from standard MCMC theory, see Robert and Casella (2004, Theorem 10.8, Theorem 10.10 (ii)), where absolute continuity of the transition kernel of the Gibbs chain follows from choosing the Lebesgue-measure λ as the dominating measure. □

A.4 Proof of Theorem 15.8

Proof. In the case the ROPE R_j is correct, R_j includes the true effect size δ_0, so that $\delta_0 \subset R_j$. Estimation of δ via δ_{MPE} is then consistent: The posterior μ_n is said to be consistent for the parameter δ_0, if for every neighbourhood U of δ_0, $\mu_n(U) \xrightarrow[n\to\infty]{a.s.} 1$ π-almost surely for any prior π on δ. By Theorem 1 in Ghosal (1996) the consistency of the posterior follows for any prior π when choosing *any* ROPE U which contains the true parameter δ_0, except possibly on a set of π-measure zero.[1] As a direct consequence one therefore obtains: If the ROPE $R_j \neq \emptyset$ is correct and contains the true parameter δ_0, any prior π leads to a consistent posterior for which $\mu_n(R_j) \xrightarrow[n\to\infty]{a.s.} 1$ π-almost surely. This implies that

$$PMP(\delta_{MPE}) = \int_{R_j} p(\theta|x)d\theta \xrightarrow[n\to\infty]{\pi-a.s.} 1$$

If on the other hand R_j is incorrect, then $\delta_0 \notin R_j$. Then there exists a neighbourhood $N := (\delta_0 - \varepsilon, \delta_0 + \varepsilon)$ for $\varepsilon > 0$ around δ_0, so that $N \cap R_j = \emptyset$. Then, as $\mu_n(N) \xrightarrow[n\to\infty]{\pi-a.s.} 1$ almost surely it follows that on the complement N^c, $\mu_n(N^c) \xrightarrow[n\to\infty]{\pi-a.s.} 0$ and because of $R_j \subset N^c$ also that $\mu_n(R_j) \xrightarrow[n\to\infty]{\pi-a.s.} 0$ π-almost surely Thereby it follows that

$$PMP(\delta_{MPE}) = \int_{R_j} p(\theta|x)d\theta \xrightarrow[n\to\infty]{\pi-a.s.} 0$$

□

[1] This result is also known as Doob's consistency theorem, compare Doob (1949).

A.5 Proof of Theorem 15.11

Proof. An α type I error happens if the true parameter value $\delta_0 \in H$, with $H \subset R$ for a ROPE $R \subset \Theta$, but H is α-rejected for α. If any correct ROPE R is selected around the hypothesis $H \subset \Theta$ which makes a statement about the unknown parameter δ, then by Definition 15.5 the true value δ_0 of δ is inside R, that is $\delta_0 \subseteq R$. Then, under any prior π on δ, the posterior $\mu_n(R) \xrightarrow[n\to\infty]{\pi-a.s.} 1$ except on a set of π-measure zero, compare Ghosal (1996). Therefore, the corresponding α% HPD interval C_α which is based on the posterior density $p(\delta|x)$ lies inside R for $n \to \infty$, too, that is: $C_\alpha \subseteq R$, and by definition, H is then α-accepted. As $\alpha \in [0, 1]$ was arbitrary, the above holds in particular without loss of generality for $\alpha = 1$, and therefore, H is accepted always for $n \to \infty$ π-almost surely for any correct ROPE R around the hypothesis H. This in turn implies that H can only be α-rejected for $\alpha = 0$ under the above conditions. However, α rejection of H for $\alpha = 0$ is only stating that zero percent of the HPD interval are located outside the ROPE R around H, which implies that the α% HPD lies fully inside the ROPE. Thus, α-rejection of H for $\alpha = 0$ is equivalent to α-acceptance of H for $\alpha = 1$, and thus H is also accepted when $\alpha = 0$.

An α type II error happens if the true parameter value $\delta_0 \notin H$ and $\delta_0 \notin R$, with $H \subset R$ for a ROPE $R \subset \Theta$, but H is α-accepted for α. If any incorrect ROPE R is selected around the hypothesis $H \subset \Theta$ which makes a statement about the unknown parameter δ, then the true value δ_0 of δ is not inside R, that is $\delta_0 \notin R$. Then, under any prior π on δ, the posterior $\mu_n(R) \xrightarrow[n\to\infty]{\pi-a.s.} 0$, compare Ghosal (1996). Therefore, the corresponding α% HPD interval C_α which is based on $p(\delta|x)$ lies not inside R for $n \to \infty$, which means $C_\alpha \notin R$, and by definition, H is then α-rejected. As $\alpha \in [0, 1]$ was arbitrary, the above holds in particular for $\alpha = 1$, and therefore, H is rejected always for $n \to \infty$ π-almost surely under any prior π and for any incorrect ROPE R around the hypothesis H. This also implies that H can only be α-accepted for $\alpha = 0$ under the above conditions. Furthermore, α-acceptance of H for $\alpha = 0$ can be interpreted as zero percent of the HPD interval C_α being located inside the ROPE R around H, and this is equivalent to α-rejection of H for $\alpha = 1$. Thus, in this case, H is also rejected and no α type II error happens. □

Appendix B

The Relative Likelihood Principle for continuous Probability Spaces

B.1 The Relative Likelihood Principle

To extend the LP to the continuous case, Berger and Wolpert (1988) assumed an experiment with random variable X having probability distribution $\mathbb{P}_\theta$ without assuming the existence of a density. Also, the sample space Ω was assumed to be a locally-compact Hausdorff space whose topology admits a countable base. The treatment of Berger and Wolpert (1988) only assumed the measures $\mathbb{P}_\theta$ to be Borel measures.

The first difficulty are sets of measure zero. The likelihood function cannot be specified in a unique way anymore now, because if there exists no single σ-finite measure ν on Ω whose null sets are identical to the Borel null sets N with $\mathbb{P}_\theta(N) = 0$ for all $\theta \in \Theta$, the consequence is that no likelihood function exists. Even when there is a σ-finite measure ν fulfilling this property the Radon-Nikodym derivatives

$$f(x|\theta) = \frac{\mathbb{P}_\theta(dx)}{\nu(dx)}$$

are determined only up to null sets of ν.[1] The solution of Berger and Wolpert (1988, p. 30) was to specify a particular version of $\frac{\mathbb{P}_\theta(dx)}{\nu(dx)}$ by defining Ω_x as an open neighbourhood of $x \in \Omega$ and setting

$$L(\theta; x) := \inf_{V \in \Omega_x} \sup_{U \in \Omega_x, U \subset V} \frac{\mathbb{P}_\theta(U)}{\nu(U)}$$

for all x in the support of ν and $L(\theta; x) := 0$ elsewhere. The idea was to construct ν-almost everywhere continuous equivalents of the WCP and WSP and derive a continuous version of the LP, the relative LP. The generality achieved is huge, and the solution is applicable for experiments with discontinuous density functions or no likelihood function at all. Still, Berger and Wolpert (1988) noted, that

> "The price we pay for such generality is that our conclusions will all be weakened by the qualification "for all $x \in \Omega$ outside a fixed set N with $\mathbb{P}_\theta(N) = 0$

[1] Note the strong analogy to the spaces $\mathcal{L}^p$ and L^p in measure theory, where the solution is to use the quotient space and identify the Radon-Nikodým density as a representant of the equivalence class, where the representants differ only in their values on the null sets, see also Bauer (2001).

> for all θ", which we shall abbreviate "for $\{P_\theta\}$ a.e. x"."
> (Berger and Wolpert, 1988, p. 31), where notation has been modified for notational consistency in this appendix

Now, the only assurance that the actually observed values x are not in N is the 'faith that events of probability zero do not happen.' (Berger and Wolpert, 1988, p. 31). This is no severe limitation anyway, and the main steps were to first reformulate the WCP and WSP to the continuous equivalent, see Berger and Wolpert (1988, p. 31).

Continuous Weak Conditionality Principle (CWCP (Berger and Wolpert, 1988)). *Consider the mixture E^* of two experiments $E_i := (X_i, \theta, \{\mathbb{P}^i_\theta\})$ for $i = 1, 2$, defined as $E^* = (X^*, \theta, \{\mathbb{P}^*_\theta\})$, with $X^* = (J, X_J)$, $J = 1$ or 2 (as E_J is performed) with probability $\frac{1}{2}$ each independent of θ and*

$$\mathbb{P}^*_\theta(A) = \frac{1}{2}\mathbb{P}^1_\theta(\{x_1 : (1, x_1) \in A\}) + \frac{1}{2}\mathbb{P}^2_\theta(\{x_2 : (2, x_2) \in A\})$$

Then,

$$Ev(E^*, (j, x_j)) = Ev(E_j, x_j)$$

for $\{\mathbb{P}^*_\theta\}$-almost everywhere (j, x_j).

The concept of sufficiency is extended to the continuous case by starting with an experiment $E = (X, \theta, \{\mathbb{P}_\theta\})$ and a measurable map $T : \Omega \to \tilde{\Omega}$ from $\tilde{\Omega}$ to another locally-compact Hausdorff space $\tilde{\Omega}$ whose topology admits a countable base. The statistic T determines a family $\{\mathbb{P}^T_\theta\}$ of Borel measures on $\tilde{\Omega}$ given by the push-forward measures

$$\mathbb{P}^T_\theta(A) = \mathbb{P}_\theta(T^{-1}(A))$$

and thereby a new experiment $E^T = (T, \tilde{\Omega}, \{\mathbb{P}^T_\theta\})$. In general, unless T is one-to-one, the 'compressed' experiment E^T will yield less information about θ than the original experiment E. The concept of sufficiency was then defined as the exceptional case in which no information is lost, which also includes any one-to-one measurable mapping, see (Berger and Wolpert, 1988, p. 32):

Definition B.1 (Sufficiency (Continuous) (Berger and Wolpert, 1988))**.** For the experiment E^T, suppose there exists a family $(g_t : t \in \tilde{\Omega})$ of Borel probability measures on $\tilde{\Omega}$ satisfying

$$\mathbb{P}_\theta(A) = \int_{\tilde{\Omega}} g_t(A)\mathbb{P}^T_\theta(dt) = \int_{\Omega} g_{T(x)}(A)\mathbb{P}_\theta(dx)$$

for all Borel sets $A \subset \tilde{\Omega}$. Then T is called sufficient (for θ).[2]

Based on this definition, Berger and Wolpert (1988) defined the continuous version of the WSP simply as the principle that if the measurable map T is sufficient, then $T(\omega)$ in $\tilde{\Omega}$ yields the same evidence (about θ) as x in the original space E:

[2] For more information on the measure-theoretic motivation of this definition see Rüschendorf (2014, Chapter 2) and Rüschendorf (2014, Definition 4.1.1). Essentially, the continuous definition of sufficiency guarantees that the Radon-Nikodým equation for conditional probabilities and expectation holds, see also Definition C.50 and Equation (C.19).

Continuous Weak Sufficiency Principle (CWSP (Berger and Wolpert, 1988)). *If $T : \Omega \to \tilde{\Omega}$ is sufficient, then*

$$Ev(E, x) = Ev(E^T, T(x))$$

for $\{\mathbb{P}_\theta\}$-almost everywhere $x \in \Omega$.

Based on these two principles, Berger and Wolpert (1988) showed that

1. For two experiments $E_i = (X_i, \theta, \{\mathbb{P}_\theta^i\}, i = 1, 2$ with countable sample space devoid of outcomes impossible under all θ, the LP and RLP are equivalent (Berger and Wolpert, 1988, p. 34, Theorem 2). This is important, because this implies that the RLP is a valid extension of the LP, which does not lead to contradictory results when applying it in the discrete case.[3]

2. The most important aspect: The CWCP and CWSP together imply the RLP (Berger and Wolpert, 1988, p. 35, Theorem 3), which is stated below.

Relative Likelihood Principle (RLP (Berger and Wolpert, 1988)). *Let $\phi : U_1 \to U_2$ be a Borel bimeasurable one-to-one mapping from $U_1 \subset \Omega$ onto $U_2 \subset \tilde{\Omega}$, and suppose there exists a strictly positive function c on U_1 such that $\forall \theta \in \Theta$*

$$\mathbb{P}_\theta^2(A) = \int_{\phi^{-1}(A)} [1/c(x_1)] \mathbb{P}_\theta^1(dx_1) \text{ for } A \subset U_2$$

Then, $Ev(E_1, x_1) = Ev(E_2, \phi(x_1))$ for $\{\mathbb{P}_\theta^1\}$-almost everywhere $x_1 \in U_1$.

[3] Note again the analogy to extending contents on algebras to measures on σ-algebras if the content is continuous in zero, see Bauer (2001).

Appendix C

Measure-theoretic Foundations of Statistical Inference

This appendix outlines the measure-theoretic foundations of statistical inference from a frequentist and Bayesian perspective.

First, the measure-theoretic foundations of frequentist inference will be outlined. Second, Bayesian inference will be discussed and contrasted with the frequentist approach. Subsequently, both approaches will be embedded in the framework of statistical decision theory which shows how frequentist and Bayesian approaches differ concerning parameter estimation, hypothesis testing or confidence set estimation.

C.1 Frequentist statistics

Frequentist statistical inference procedures consist of three ingredients: 1) The observation of data, 2) a statistical model and 3) an estimation procedure. Frequentist inference assumes that observing data $X(\cdot)$ during an experiment or study has a definite but unknown underlying probability distribution. The observed data is mathematically given by a realisation $X(\omega)$ of a random variable $X : (\Omega, \mathcal{B}, \mu) \to (\mathcal{X}, \mathcal{A})$ which maps from an unknown probability space $(\Omega, \mathcal{B}, \mu)$ into a measure space $(\mathcal{X}, \mathcal{A})$, which is called the *sample space*, compare Figure C.1. The sample space is assumed to be a measure space to enable the consideration of probability measures on $\mathcal{X}$, which formalise the uncertainty in observing $X(\omega) \in \mathcal{X}$ for $\omega \in \Omega$. The probability space $(\Omega, \mathcal{B}, \mu)$ is unknown in practice, and if it would be known – which would imply the precise distribution under which data $X(\omega) \in \mathcal{X}$ is observed would also be known – no randomness would be involved anymore. In practice, frequentist statistics hinges on the assumption that there exists a true probability measure P_0 on the sample space $(\mathcal{X}, \mathcal{A})$ which represents the "true distribution of the data". This means, $X \sim P_0$, or expressed differently, for all $A \in \mathcal{A}$ the probability $P_0(A)$ is given by the measure induced by μ:

$$P_0(A) := \mu(\{\omega \in \Omega : X(\omega) \in A\})$$

Assuming the existence of such a P_0 is quite strong, but allows for answering questions like "What does the data tell us about P_0?". So, from a frequentist perspective the data are realisations of a random variable X into a measurable space $(\mathcal{X}, \mathcal{A})$, where the true distribution P_0 of X is unknown but assumed to exist on $(\mathcal{X}, \mathcal{A})$. A statistical model formalises this notion:

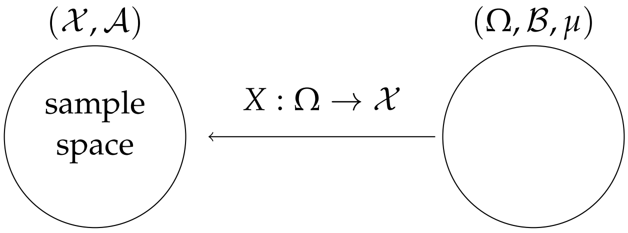
$(\mathcal{X}, \mathcal{A})$
sample space
$X : \Omega \to \mathcal{X}$
$(\Omega, \mathcal{B}, \mu)$

Figure C.1: Measure-theoretic background of frequentist inference

Definition C.1 (Statistical model)**.** The triple $\mathcal{E} := (\mathcal{X}, \mathcal{A}, \mathcal{P})$ is called statistical model, if $(\mathcal{X}, \mathcal{A})$ is a measurable space and $\mathcal{P} \subset M^1(\mathcal{X}, \mathcal{A})$ is a class of probability distributions on the space $(\mathcal{X}, \mathcal{A})$, where $M^1(\mathcal{X}, \mathcal{A})$ is the set of probability distributions on $(\mathcal{X}, \mathcal{A})$.

When it is clear which space the model refers to, $\mathcal{P}$ is simply called the statistical model. The model $\mathcal{P}$ can be interpreted as the distributions on $(\mathcal{X}, \mathcal{A})$ the statistician finds reasonable to explain the uncertainty in observing the data $X(\omega) \in \mathcal{X}$. To draw any inference about the true distribution P_0, an important assumption of frequentist methods is that the true distribution P_0, according to which data $X(\omega) \in \mathcal{X}$ is observed, is contained in $\mathcal{P}$, compare (Kleijn, 2022, Definition 1.6):

Definition C.2 (Well-specified)**.** A statistical model $\mathcal{P}$ is well-specified if it contains the true distribution P_0 of the data X. That is, $P_0 \in \mathcal{P}$.

Example C.3. Suppose we measure the deviation of reaction times of patients in a clinical trial from a known reference reaction time. After patients have been administered the drug, reaction times are recorded and the deviations from a reference reaction time are calculated. Suppose a sample of 50 patients is recruited and we are interested in the mean deviation from the reference reaction time. The data are shown in Figure C.2.

Let $(\mathcal{X}, \mathcal{A}) := (\mathbb{R}^{50}, \mathcal{B}(\mathbb{R}^{50}))$, that is, the observed data is vector $(x_1, ..., x_{50})$ with $x_i \in \mathbb{R}$, and $\mathcal{B}(\mathbb{R}^{50})$ is the Borel-σ-algebra on $\mathbb{R}^{50}$. Let $\mathcal{P} \subset M^1(\mathcal{X}, \mathcal{A})$ be any subset of $M^1(\mathcal{X}, \mathcal{A})$.

Now, the probability measures P_θ in a statistical model are commonly described by a parameterization, compare (Kleijn, 2022, Definition 1.4):

Definition C.4 (Parameterization)**.** A statistical model $\mathcal{P}$ is parameterized with parameter space Θ, if there exists a surjective map $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$, called the parameterization of $\mathcal{P}$.

Note that the set Θ, which is also called the parameter space is only a set and no measure space. Thus, there is no associated σ-algebra for Θ, nor a probability measure in the frequentist approach. However, if the parameterization is injective (and then, because of the last definition also bijective), one calls a parameterization identifiable, because each parameter $\theta \in \Theta$ identifies exactly one probability measure $P_\theta \in \mathcal{P}$, compare (Kleijn, 2022, Definition 1.5):

Definition C.5 (Identifiable)**.** A parameterization of a statistical model $\mathcal{P}$ is called identifiable, if the map $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ is injective.

Figure C.3 shows how the parameter space Θ (which is only a set, but is still called space in the frequentist approach) parameterizes the model $\mathcal{P}$ which is assumed on the sample space $(\mathcal{X}, \mathcal{A})$. Thus, the statistical model can be written as $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$.

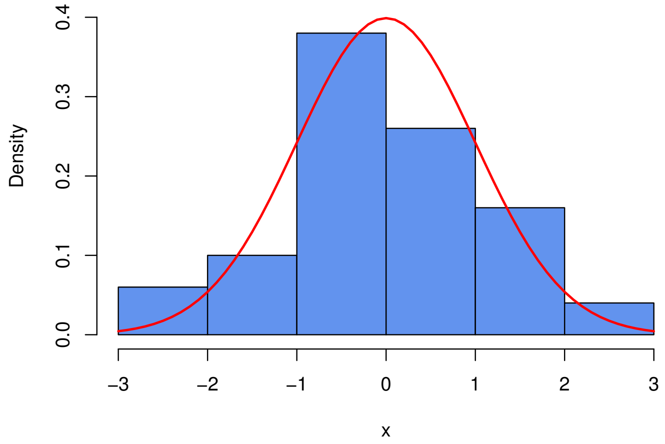

Figure C.2: Histogram of deviations from the reference reaction time for a sample of $n = 50$ patients. Data were simulated from a $\mathcal{N}(0,1)$ distribution.

Example C.6 (Continuation of Example B.1)**.** In the situation of Example B.1, select $\Theta := \mathbb{R}$, and an identifiable parameterization is given by $\theta \mapsto P_\theta$, where $P_\theta := \mathcal{N}(\theta,1)^{(50)}$.

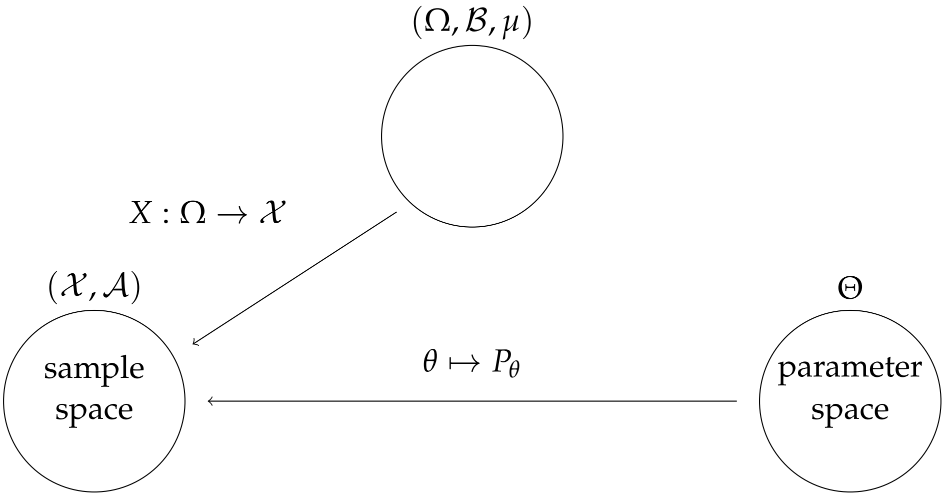
$(\Omega, \mathcal{B}, \mu)$
$X : \Omega \to \mathcal{X}$
$(\mathcal{X}, \mathcal{A})$
Θ
sample space
$\theta \mapsto P_\theta$
parameter space

Figure C.3: Measure-theoretic background of frequentist inference with an identifiable parameterization $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$

The most important distinction in (frequentist) statistics is between parametric and non-parametric models. A model $\mathcal{P}$ is called *parametric* of dimension d, if there exists an identifiable parameterization $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ where $\Theta \subset \mathbb{R}^d$ with non-empty interior $\mathring{\Theta} \neq \emptyset$. If there is no finite-dimensional Θ which parameterizes the model $\mathcal{P}$, then $\mathcal{P}$ is called *non-parametric model*.

Often in the statistical literature, models are described as families of probability densities rather than working with probability measures directly. From a mathematical perspective, to guarantee the existence of Radon-Nikodým densities, an absolutely continuous σ-finite measure is required, which leads to the following definition, compare

(Kleijn, 2022, Definition 1.3):

Definition C.7 (Dominated)**.** If there exists a σ-finite measure $\nu : \mathcal{A} \to [0, \infty]$ such that for all $P \in \mathcal{P}$, $P \ll \nu$, the model is *dominated* (notation: $\mathcal{P} \ll \nu$).

The Radon-Nikodým theorem guarantees that in dominated models one can work with probability densities $dP/d\nu : \mathcal{X} \to [0, \infty)$ instead of working with the measures $P \in \mathcal{P}$.

Example C.8 (Continuation of Example B.2)**.** In the situation of Example B.2, the probability measures P_θ, are dominated by the Lebesgue measure $\lambda^{(n)}$, that is $P_\theta \ll \lambda^{(n)}$ for $n = 50$ and all $\theta \in \Theta$.

The third ingredient of a frequentist procedure is a method for estimation. Although hypothesis testing and confidence sets seem at first glance different from estimation, all three tasks can be formalised under the framework of statistical decision theory, where the above method for estimation becomes a decision rule.

Definition C.9 ((Point) Estimator)**.** A point-estimator (or estimator) for P_0 is a map $\hat{P} : \mathcal{X} \to \mathcal{P}$, which represents the "best guess" $\hat{P} \in \mathcal{P}$ for P_0 based on the data $X(\omega) \in \mathcal{X}$.

Note that $\hat{P}(X)$ is random and depends on the observed $X(\omega) \in \mathcal{X}$. If the model is parameterized, one can equivalently define a (point) estimator as a map $\hat{\Theta} : \mathcal{X} \to \Theta$ for θ_0, from which one obtains $\hat{P} = P_{\hat{\Theta}}$ as an estimator for P_0. If the model is identifiable, assume the measure $P_0 := P_{\theta_0}$ (that is, the true model) corresponds to the parameter θ_0. Estimation of θ_0 in Θ is then equivalent to estimation of P_0 in $\mathcal{P}$, see Figure C.4.

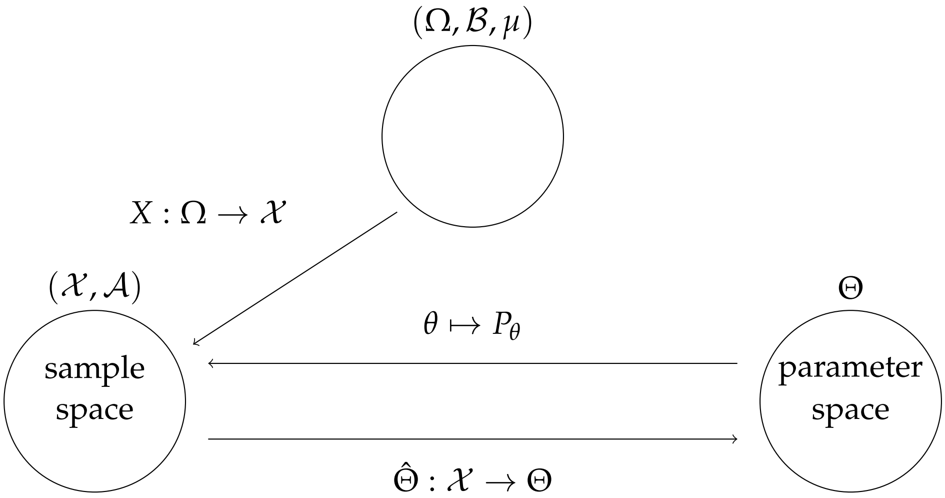
$(\Omega, \mathcal{B}, \mu)$
$X : \Omega \to \mathcal{X}$
$(\mathcal{X}, \mathcal{A})$
Θ
$\theta \mapsto P_\theta$
sample space
parameter space
$\hat{\Theta} : \mathcal{X} \to \Theta$

Figure C.4: Measure-theoretic background of frequentist inference with an identifiable parameterization $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ and a (point) estimator $\hat{\Theta} : \mathcal{X} \to \Theta$

Example C.10 (Continuation of Example B.3)**.** Suppose in the situation of Example B.3, $\hat{\Theta} : \mathbb{R}^{(50)} \to \mathbb{R}$ is defined as $\hat{\Theta}(x) := \frac{1}{50} \sum_{i=1}^{50} x_i$ for observed data $x := X(\omega)$ (for $\omega \in \Omega$). This is simply the sample mean, which estimates the mean parameter θ in P_θ in $\mathcal{P}$. For the data shown in Figure C.2, $\hat{\Theta}((x_1, ..., x_{50})) = -0.03$ based on two-digits precision. As data in Figure C.2 were indeed simulated from a $\mathcal{N}(0, 1)$ distribution, the estimator is quite close to the true parameter $\theta_0 = 0$. Note also that the model is well-specified because the precise distribution of X is known. In practice, this is, of course, not the case.

In summary, a frequentist needs to model a sample space $(\mathcal{X}, \mathcal{A})$ and associate a family of probability measures $\mathcal{P}$ with it, where the family expresses the uncertainty in observing the random quantity $X(\omega) \in \mathcal{X}$. The space $(\Omega, \mathcal{B}, \mu)$ is unknown in practice, and via a parameterization (which ideally is identifiable), and an estimation method (like a (point) estimator), P_0 is estimated by the data via $\hat{P}$ (or $\hat{\Theta}$). As the true model P_0 is assumed to be contained in the chosen family $\mathcal{P}$, enough data $X(\omega)$ should eventually "reveal" the true $\theta_0 \in \Theta$.

C.2 Bayesian statistics

The preceding section detailed the measure-theoretic foundations of frequentist inference. In this section, Bayesian statistics is contrasted with the frequentist measure-theoretic foundations. In the frequentist approach, the parameter θ was assumed to be a fixed but unknown value θ_0 in the set Θ, the parameter space. In the Bayesian framework, not only are the data X a random variable, but the parameter is random, too. Now, the parameter space Θ is assumed to be a measure space (Θ, τ) with σ-algebra τ and the parameter θ is a random variable ϑ which takes values in Θ. The main difference in the Bayesian approach now is that one assumes a probability measure $\mu : \sigma(\mathcal{X} \times \tau) \to [0, 1]$ on the product space $\Omega := \mathcal{A} \times \tau$ with product σ-algebra $\mathcal{B} := \sigma(\mathcal{A} \times \tau)$, compare Figure C.5. This probability measure provides a joint probability distribution for (X, ϑ), that is, for the data X and the parameter ϑ. Importantly, the choice of this measure on the product σ-algebra $\sigma(\mathcal{X} \times \tau)$ *defines* the statistical model $\mathcal{P}$ in the Bayesian approach, by the possibility to condition the distribution of X on fixed values $\vartheta = \theta \in \Theta$ (Kleijn, 2022, Section 2.1.1). The conditional distribution $X|\vartheta$ (X given ϑ) describes the distribution of the observation X given the parameter ϑ and as a consequence, the distributions $X|\vartheta = \theta$ can be identified as the elements P_θ of the (identifiably) parameterized model $\mathcal{P} = \{P_0 : \theta \in \Theta\}$ also used in frequentist statistics.

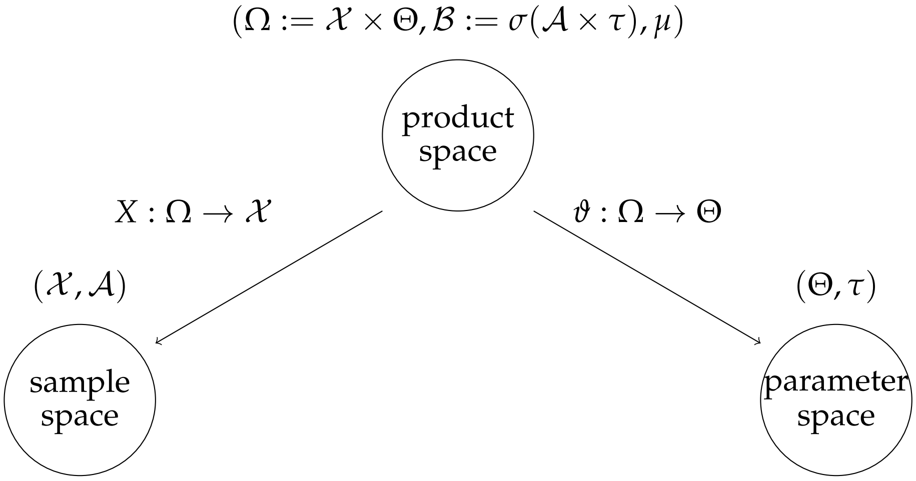
$(\Omega := \mathcal{X} \times \Theta, \mathcal{B} := \sigma(\mathcal{A} \times \tau), \mu)$
product space
$X : \Omega \to \mathcal{X}$
$\vartheta : \Omega \to \Theta$
$(\mathcal{X}, \mathcal{A})$
(Θ, τ)
sample space
parameter space

Figure C.5: Measure-theoretic background of Bayesian statistics

Definition C.11 (Model distribution). The distribution of the data X conditional on the parameter ϑ is a regular conditional distribution, $\mu_{X|\vartheta} : \mathcal{A} \times \Theta \to [0,1]$ which describes the model distributions P_θ.

Now, the Bayesian statistician also needs to incorporate some a priori information or beliefs about the uncertainty of the parameter θ. This is expressed via the prior distribution. The marginal distribution of the parameter is called the prior distribution.

Definition C.12 (Prior distribution). The marginal distribution $\mu_\Theta : \tau \to [0,1]$ is called the prior distribution for the parameter.

Example C.13 (Continuation of Example B.4). In the setting of Example B.4, we identify the model distribution $\mu_{X|\vartheta} : \mathcal{B}(\mathbb{R}^{(50)}) \times \mathbb{R} \to [0,1]$ as the distributions $\mathcal{N}(\theta,1)^{(50)}$. Let $\tau := \mathcal{B}(\mathbb{R})$ the Borel-σ-algebra on $\mathbb{R}$. Let $\mu_\Theta := \mathcal{N}(\mu_0, \sigma_0^2)$ the prior distribution on τ, where $\mu_0 \in \mathbb{R}$ and $\sigma_0^2 > 0$. This prior distribution reflects our a priori information or beliefs regarding the deviations of the patient reaction time from the reference reaction time in the clinical trial.

It is important to note that the product measure μ is already constructed by the Bayesian statistician when a prior distribution μ_Θ is selected for Θ and the statistical model $\mathcal{P}$ is chosen. Together, the prior distribution and the model $\{P_\theta : \theta \in \Theta\}$ determine a joint distribution on the product-space $\mathcal{X} \times \Theta$.

Example C.14 (Continuation of Example B.5). In the setting of Example B.5, suppose that the prior distribution μ_Θ has density f_θ (which is the Lebesgue density of the $\mathcal{N}(\mu_0, \sigma_0^2)$ distribution). Thus, it is implicitly assumed that X is continuous and let $B \subseteq \mathcal{X} \times \Omega$. Then the probability measure μ on $\sigma(\mathcal{A} \times \tau)$ is given as

$$\mu((X,\vartheta) \in B) = \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}_B(x,\theta) f_{X|\Theta}(x|\theta) f_\Theta(\theta) dx d\theta$$

In the Bayesian approach, X and ϑ are easily recognized as the projection of the joint space $\mathcal{X} \times \Theta$ to the respective components $\mathcal{X}$ and Θ. To be more specific, for $s := (x,\theta) \in \mathcal{X} \times \Theta$, $X(s) = x$ and $\vartheta(s) = \theta$. The selection of the statistical model is equivalent to the selection of the conditional distributions $\mu_{X|\vartheta} : \mathcal{A} \times \Theta \to [0,1]$ under the assumption that the model is parameterized and identifiable. If the prior measure μ_Θ is selected, too, the measure $\mu(\mathcal{X}, \tau) \to [0,1]$ is induced on the product σ-algebra $\sigma(\mathcal{A} \times \tau)$.

Now, central to the Bayesian framework is the conditional distribution for ϑ given X, called the posterior distribution. In Bayesian statistical inference, all inference about the parameter is made with respect to the posterior distribution after observing the data X.

Definition C.15 (Posterior distribution). The conditional distribution $\mu_{\vartheta|X} : \tau \times \mathcal{A} \to [0,1]$ for $\vartheta|X$ is called the posterior distribution.

The transition from prior to posterior is achieved via Bayes' theorem:

Theorem C.16 (Bayes' Theorem). Assume that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure ν on $(\mathcal{A}, \tau)$ with densities $f_\theta = dP_\theta / d\nu$. Then the posterior can be expressed as

$$\mu(\vartheta \in B|X) = \frac{\int_B f_\theta(X) d\mu_\Theta}{\int_\Theta f_\theta(X) d\mu_\Theta} \tag{C.1}$$

For a proof of Bayes' theorem see Kleijn (2022, Theorem 2.2) or Schervish (1995, 1.31).

Example C.17. Let $B := \theta \in \Theta$ and the observed data $X(\omega) := x \in \mathcal{X}$. Suppose ν is the Lebesgue-measure λ, and $\mathcal{P} \ll \lambda$ as well as $\mu_\Theta \ll \lambda$. Let $p(\cdot)$ denotes the density of the prior measure μ_Θ with respect to the dominating measure ν, that is $p = d\mu_\Theta / d\lambda$. Then, Equation (C.1) becomes the more familiar-looking

$$\mu(\vartheta = \theta|X = x) = \frac{\int_\theta f_\theta(X) \overbrace{d\mu_\Theta}^{=p(\theta)d\lambda}}{\int_\Theta f_\theta(X) \underbrace{d\mu_\Theta}_{=p(\theta)d\lambda}} = \frac{\int_\theta f_\theta(X)p(\theta)d\lambda}{\int_\Theta f_\theta(X)p(\theta)d\lambda}$$

In summary, as shown in Figure C.5, probabilities are calculated in the Bayesian approach with respect to the probability space $(\Omega, \mathcal{B}, \mu)$. The observed random sample is the realisation of a random variable $X : \Omega \to \mathcal{X}$, which is a mapping from Ω into the Borel space $(\mathcal{X}, \mathcal{A})$, called the *sample space*. The statistical model $\mathcal{P}$ is associated with $(\mathcal{X}, \mathcal{A})$ and is known up the unknown model parameter $\theta \in \Theta$. In the Bayesian perspective, $\vartheta : \Omega \to \Theta$ is called the *parameter* and is a random variable, that is, a measurable function from Ω into the Borel space (Θ, τ), where the latter is called the *parameter space*.

Example C.18 (Continuation of Example B.6)**.** In the setting of Example B.6, the prior distribution was $\mu_\Theta := \mathcal{N}(\mu_0, \sigma_0^2)$ and the model distribution $\mu_{X|\vartheta} := \mathcal{N}(\theta, 1)^{(50)}$. Standard calculus then yields the posterior distribution

$$\mu_{\vartheta|X} = \mathcal{N}\left(\frac{1}{\frac{1}{\sigma_0^2} + n}\left(\frac{\mu_0}{\sigma_0^2} + \sum_{i=1}^{50} x_i\right), \frac{1}{\frac{1}{\sigma_0^2} + n}\right)$$

For details, see Held and Sabanés Bové (2014, p. 181-182). Suppose we choose a prior $\mathcal{N}(0, 1)$ on θ, which corresponds to $\mu_0 := 0$ and $\sigma_0^2 = 1$. Then the posterior distribution is given as $\mathcal{N}(\frac{\sum_{i=1}^{50} x_i}{51}, \frac{1}{51})$. For the data shown in Figure C.2, $\sum_{i=1}^{(50)} x_i = -1.784$, so the posterior is $\mathcal{N}(-0.034, 0.019)$. Figure C.6 shows the prior and posterior together with the histogram of the observed data.

From a mathematical perspective, assuming the existence of a dominating measure ν on $(\mathcal{X}, \mathcal{A})$ for which each P_θ as a probability measure on $(\mathcal{X}, \mathcal{A})$ is absolutely continuous is important. The absolute continuity $P_\theta \ll \nu$ for all $\theta \in \Theta$ guarantees the existence of Radon-Nikodym densities $f_\theta = \frac{dP_\theta}{d\nu}$. One can assume that $f_\theta(x)$ is measurable with respect to the product σ-field $\mathcal{A} \otimes \tau$.[1] As a consequence, one can integrate $f_\theta(x)$ with respect to measures both on $\mathcal{X}$ and Ω, and for each $A \in \mathcal{A}$,

$$\mu_{X|\vartheta}(X \in A|\Theta = \theta) = \int_A f_\theta(x) d\nu(x)$$

[1] For a proof, see (Schervish, 1995, p. 13). Schervish (1995) uses the notation $f_{X|\theta}(x|\theta)$ for $f_\theta(x)$.

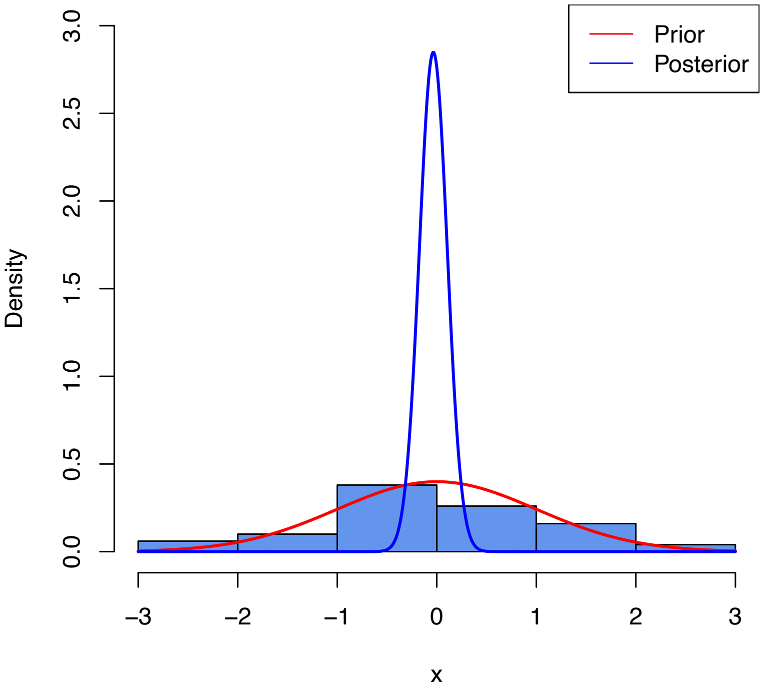

Figure C.6: Histogram, $\mathcal{N}(0,1)$ prior and $\mathcal{N}(-0.034, 0.019)$ posterior for the screw data. Data were simulated from a $\mathcal{N}(0,1)$ distribution.

The marginal distribution $\mu_X = \mu(X \in A, \vartheta \in \Theta)$ can be written as

$$\mu_X(A) = \int_\Omega \int_A f_\theta(x) d\nu(x) d\mu_\Theta(\theta) = \int_A \int_\Omega f_\theta(x) d\mu_\Theta d\nu(x)$$

where the last equality follows from Tonelli's theorem. From the above equation it follows that μ_X is absolutely continuous with respect to ν with density

$$f_X(x) = \int_\Omega f_\theta(x) d\mu_\Theta$$

which is called the prior predictive density of X or marginal density of X. To summarise, the Bayesian procedure consists of four steps:

1. Based on the available background information, the statistician chooses a model $\mathcal{P}$ of reasonable candidate distributions which express the uncertainty in observing X. Usually, the model $\mathcal{P}$ is parameterised with some (identifiable) parameterization $\Theta \to \mathcal{P}, \theta \mapsto P_\theta$.

2. A prior measure μ_Θ is chosen on (Θ, τ), which reflects the belief concerning the possible values the parameter(s) θ in the parameterised model $\mathcal{P}$ have. Usually, this is a probability measure on (Θ, τ).

3. Based on the conditional distributions, the prior, the available data and Bayes' theorem, the posterior is calculated as afunction of the data X.[2].

[2] For the case of a random i.i.d. sample, see Kleijn (2022, p. 28)

4. The statistician observes a realization $X(\omega) = x$ of the data and uses it to calculate a realisation of the posterior, upon which all further inference is based (e.g. point estimation, confidence set estimation or hypothesis testing).

Example C.19 (Continuation of Example B.8)**.** In the setting of Example B.8, one could estimate θ via the posterior distribution's mean, which is -0.034.

Notice that Bayesian inference yields a posterior *distribution*, upon which all further conclusions are based, while frequentist inference usually provides a *point in the model*, that is an estimate $\hat{\Theta}(x)$ (where $x \in \mathcal{X}$ are the observed data) for the parameter θ_0 parameterizing the true measure P_{θ_0} (Kleijn, 2022). However, although Bayesian inference provides a whole posterior distribution, Bayesians usually also use decision rules based on this posterior distribution which reduce to point estimates like the posterior mean or median.

C.3 Statistical Decision Theory

The above measure-theoretic theory of both frequentist and Bayesian statistics can be extended into the framework of statistical decision theory which goes back to Wald (1939, 1949). In this embedding it becomes apparent that frequentist procedures aim at minimising the risk with respect to a loss function, while Bayesian statistics aims at minimising the same risk with respect to the same loss function over the assumed prior distribution of the parameter. First, a decision space and decision rules (or functions) are introduced:

Definition C.20 (Decision space)**.** A decision space $(\Delta, \mathcal{A}_\Delta)$ is a measure space.

In practice, the decision space contains the actions taken when a statistical problem is considered. For example, if a parameter needs to be estimated, the decision space could be modelled as the parameter values which are possible, like $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{R}_+$. If a hypothesis H_0 is tested against an alternative H_1 (where $H_0 \subset \Theta$ and $H_1 := \Theta \setminus H_0$), the action space could be $\Delta := \{a_0, a_1\} = \{0, 1\}$, where a_0 means accept the null hypothesis H_0, and a_1 means accept H_1. Sometimes, the decision space is also called action space.

Definition C.21 (Non-randomized decision rule)**.** A non-randomized decision rule d is a mapping $d : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)$. The set

$$D := \{d : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)\}$$

is called the set of non-randomized decision rules.

The above definition states that a decision rule is simply a map from the sample space $(\mathcal{X}, \mathcal{A})$ into the decision space $(\Delta, \mathcal{A}_\Delta)$.

Example C.22. For the realisation $X(\omega) = (x_1, ..., x_n) =: x$, the decision $d(x) = a$ with $a \in \Delta$ is made (or the action a is taken).

Some statistical problems require a randomized decision rule:

Definition C.23 (Randomized decision rule)**.** A randomized decision rule δ is a Markov kernel of $\mathcal{X}$ to Δ, that is, $\delta : \mathcal{X} \times \mathcal{A}_\Delta \to [0, 1]$ is a map with

(i) $\forall A \in \mathcal{A}_\Delta : \delta(\cdot, A)$ is $\mathcal{A} - \mathcal{B}(\mathbb{R})_{[0,1]}$-measurable[3]

[3] $\mathcal{B}(\mathbb{R})_{[0,1]}$ denotes the Borel-σ-algebra restricted to $[0, 1]$.

(ii) $\forall x \in \mathcal{X} : \delta(x, \cdot) \in M^1(\Delta, \mathcal{A}_\Delta)$

The set of all randomized decision functions on $\mathcal{X}$ is denoted by $\mathcal{D} := \{\delta : \delta$ is a randomized decision

In contrast to $d \in D$, for $\delta \in \mathcal{D}$, $d(x, A)$ is the probability of deciding for $A \in \mathcal{A}_\Delta$ when observing $X(\omega) = x \in \mathcal{X}$. Note that it always suffices to treat randomized decision rules, because for $d \in D$ the following embedding formalized non-randomized decision rules as randomized ones:

$$\delta_d(x, A) := \begin{cases} 1, d(x) \in A \\ 0, d(x) \notin A \end{cases}$$

The map $D \hookrightarrow \mathcal{D}$, $d \to \delta_d$ is an injective embedding (that is, $D \subset \mathcal{D}$ and $\delta_{d_1} \neq \delta_{d_2} \Leftrightarrow d_1 \neq d_2$). Therefore, it suffices to treat only randomised decision rules. Figure C.7 shows how the measure-theoretic background of frequentist statistics in Figure C.3 is extended by statistical decision theory, and Figure C.8 shows the same situation for Bayesian statistics. Now, to quantify the loss incurred when using a specific decision rule, a loss function is introduced:

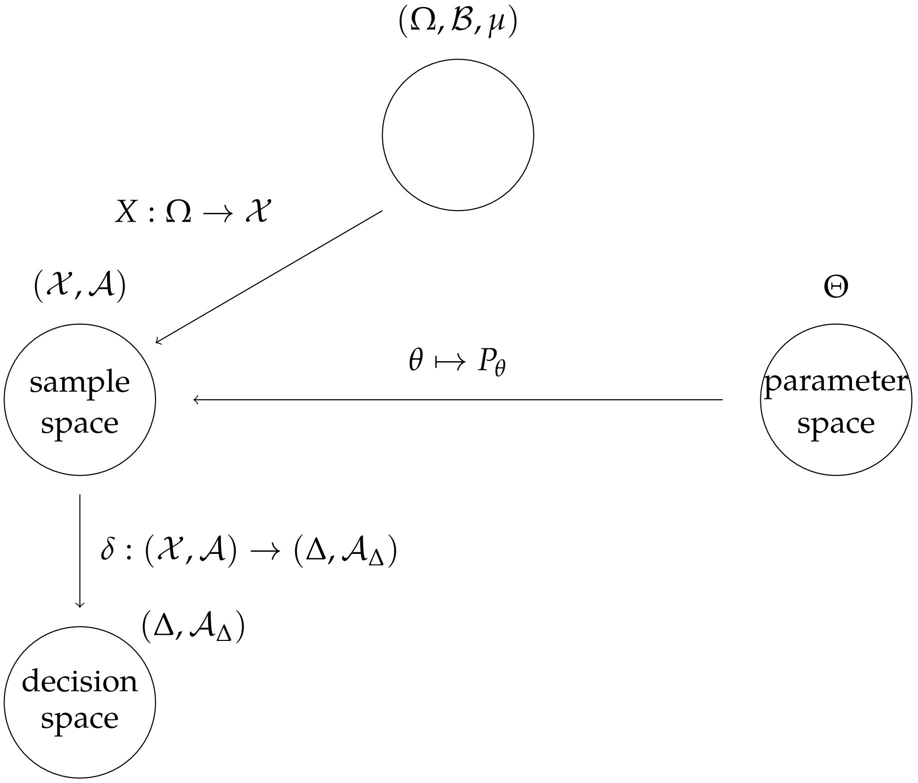
$(\Omega, \mathcal{B}, \mu)$
$X : \Omega \to \mathcal{X}$
$(\mathcal{X}, \mathcal{A})$
sample space
$\theta \mapsto P_\theta$
Θ
parameter space
$\delta : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)$
$(\Delta, \mathcal{A}_\Delta)$
decision space

Figure C.7: Measure-theoretic background of frequentist statistical inference and its connection to statistical decision theory

Definition C.24 (Loss function). $L : \Theta \times \Delta \to \overline{\mathbb{R}}_+$ is called loss function, if for all $\theta \in \Theta$:

$$L(\theta, \cdot) : (\Delta, \mathcal{A}_\Delta) \to (\mathbb{R}_+, \mathcal{B}(\overline{\mathbb{R}})_+)$$

where $\overline{\mathbb{R}}_+ := \mathbb{R} \cup \{\infty\}$. That is, a loss function $L(\theta, a)$ quantifies the positive loss incurred by a decision for $a \in \Delta$ when the parameter is $\theta \in \Theta$. In practice, there may be

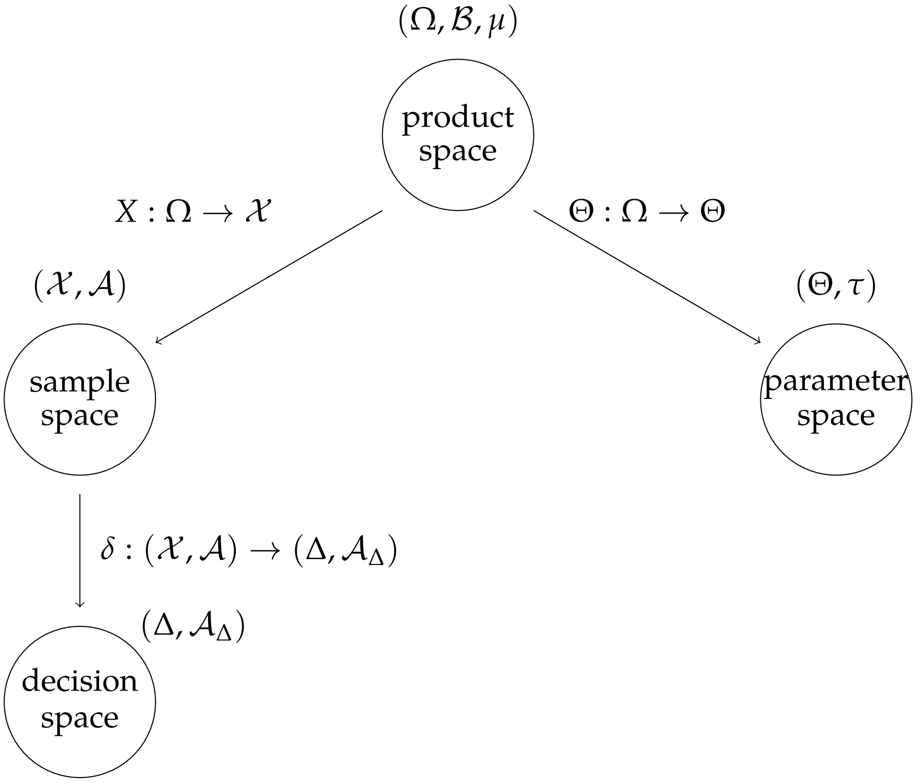
$(\Omega, \mathcal{B}, \mu)$
product space
$X : \Omega \to \mathcal{X}$
$\Theta : \Omega \to \Theta$
$(\mathcal{X}, \mathcal{A})$
(Θ, τ)
sample space
parameter space
$\delta : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)$
$(\Delta, \mathcal{A}_\Delta)$
decision space

Figure C.8: Measure-theoretic background of Bayesian statistical inference and its connection to statistical decision theory

multiple decision rules available and the statistician faces the problem to decide which one to use (no matter if a frequentist or Bayesian perspective is taken). Using the above definitions, a statistical decision problem is then formalised as follows:

Definition C.25 (Statistical decision problem)**.** The triple $(\mathcal{E}, \Delta, L)$ is called a statistical decision problem, if $\mathcal{E} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$ is a statistical model (also called a statistical experiment), $(\Delta, \mathcal{A}_\Delta)$ is a decision space and L a loss function.

As explained above, θ is the unknown state of the statistical model $\mathcal{P}$. Statistical decision theory allows to quantify the loss incurred by using a decision function after observing $x \in \mathcal{X}$ as follows:

1. Observe $x \in \mathcal{X}$ as the result of the statistical experiment, where x is assumed to follow the statistical model $\mathcal{P}$.

2. Decide for $a \in \Delta$ via a chosen non-randomized decision ($d(x) = a$) or randomized decision function (a is randomized via the Markov kernel $d(x, \cdot)$).

3. Quantify the incurred loss $L(\theta, a)$ for a chosen loss function L.

Based on the observation $X(\omega) = x$, a decision $\delta(x)$ is made. To show how broad the class of statistical decision problems in the statistical decision-theoretic framework is, the three archetypical examples of statistical inference are presented below.

C.3.1 Parameter estimation

First, consider parameter estimation. In this case, the decision space Δ is assumed to be a normed space: $(\Delta, ||\cdot||)$ where typically $\Delta := \mathbb{R}^k$ or $\Delta := L^p$, $\mathcal{A}_\Delta := \mathcal{B}(\Delta)$ is the Borel-σ-algebra on Δ. This (standard) setting for parameter estimation is visualised in

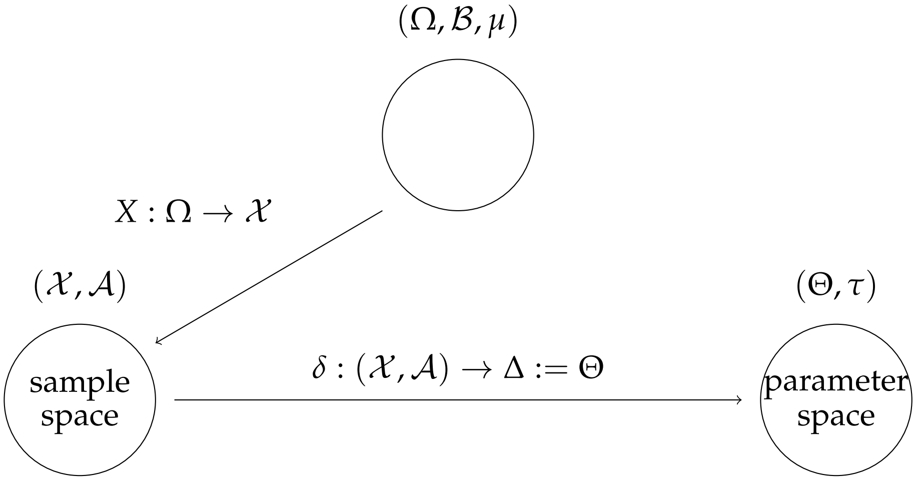
$(\Omega, \mathcal{B}, \mu)$
$X : \Omega \to \mathcal{X}$
$(\mathcal{X}, \mathcal{A})$
(Θ, τ)
sample space
$\delta : (\mathcal{X}, \mathcal{A}) \to \Delta := \Theta$
parameter space

Figure C.9: Measure-theoretic background of frequentist statistical inference and its connection to statistical decision theory for parameter estimation when $\Theta := \Delta$

Figure C.9: The action space $(\Delta, \mathcal{A}_\Delta)$ becomes the parameter space (Θ, τ). In general, the goal is to estimate θ. Frequently used loss functions in this setting are the Laplace loss L_1 or the Gauß-loss L_2, where $L_r(\theta, a) := ||a - g(\theta)||^r$ or the zero-one-loss

$$L_\varepsilon(\theta, a) := \begin{cases} 1, ||a - g(\theta)|| > \varepsilon \\ 0, ||a - g(\theta)|| \leq \varepsilon \end{cases}$$

The statistical decision problem $(\mathcal{E}, \Delta, L)$ is then called a (parameter) estimation problem for θ.

Example C.26 (Continuation of Example B.4)**.** Return to the setting of Example B.4, where the estimator $\hat{\Theta}(x) := \frac{1}{50}\sum_{i=1}^{50} x_i$ was used to estimate θ. Consider the non-randomised decision rule $d : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)$. Let $(\Delta, \mathcal{A}_\Delta) := (\Theta, \tau) = (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, then d becomes $d : (\mathbb{R}^{(50)}, \mathcal{B}(\mathbb{R}^{(50)})) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$. Let $d(x) := \bar{x}_{50} := \frac{1}{50}\sum_{i=1}^{50} x_i$ which estimates θ and let $L := L(\theta, a) = (\theta - a)^2$ the Gauß-loss L_2. The incurred loss after observing $x \in \mathcal{X}$ and deciding for $d(x)$ then is given as $(\bar{x}_{50} - \theta)^2$. In Example B.4, $\bar{x}_{50} = -0.03$, so for $\theta = 0$ the loss function yields $L(0, -0.03) = (-0.03 - 0)^2 = 0.0009$.

Example B.11 shows that a frequentist estimator is just a special case of a (randomised) decision rule when the parameter space Θ is additionally assigned a σ-algebra to fulfill the definition of a decision rule and $(\Delta, \mathcal{A}_\Delta) := (\Theta, \tau)$. The situation is shown in Figure C.9.

C.3.2 Confidence set estimation

Second, consider confidence set estimation. Let (Θ, τ) the parameter space. The map $C : \mathcal{X} \to \tau$ is called a confidence set for θ, if for all $\theta \in \Theta$:

$$A(\theta) := \{x \in \mathcal{X} : \theta \in C(x)\} \in \mathcal{A} \tag{C.2}$$

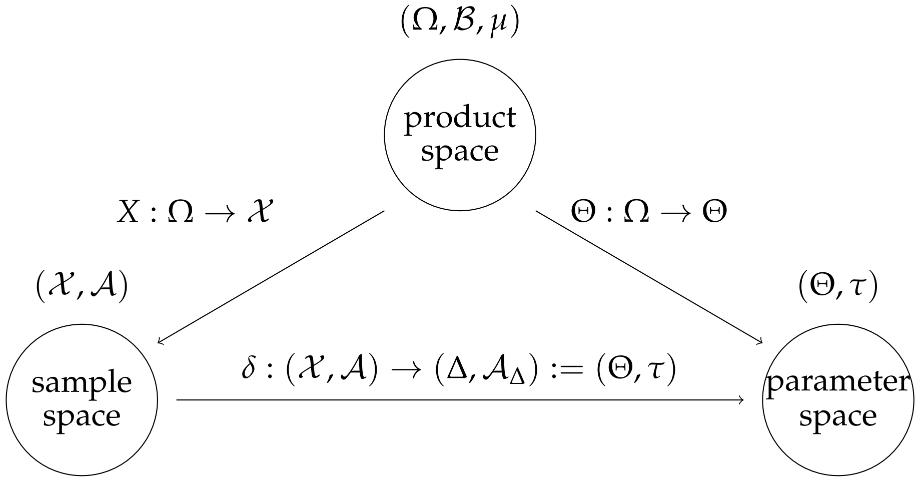
$(\Omega, \mathcal{B}, \mu)$
product space
$X : \Omega \to \mathcal{X}$
$\Theta : \Omega \to \Theta$
$(\mathcal{X}, \mathcal{A})$
(Θ, τ)
sample space
$\delta : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta) := (\Theta, \tau)$
parameter space

Figure C.10: Measure-theoretic background of Bayesian statistical inference and its connection to statistical decision theory for parameter estimation when $(\Theta, \tau) := (\Delta, \mathcal{A}_\Delta)$

$A(\theta)$ that is, the set of all values $x \in \mathcal{X}$ which are located inside $C(x)$ is $\mathcal{A}$-measurable. $A(\theta)$ is also called acceptance region of θ and is the set of all $x \in \mathcal{X}$ for which the parameter value θ is covered by the resulting confidence set $C(x)$. Let $\Delta := \tau$ and from a decision-theoretic perspective, in confidence set estimation one decides for a confidence set in the corresponding σ-algebra τ of the parameter space Θ which expresses the uncertainty about the parameter. Let $\mathcal{A}_\Delta := \sigma(\{T_\theta : \theta \in \Theta\})$, where $T_\theta := \{B \in \tau : \theta \in B\}$ is the set of all subsets $B \in \tau$ which cover the parameter value θ. Then

$$\begin{aligned} C : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta) &\Leftrightarrow \forall \theta \in \Theta : \{C \in T_\theta\} = \{x \in \mathcal{X} : \theta \in C(x)\} \in \mathcal{A} \\ &\Leftrightarrow \text{ C is a non-randomised decision rule} \end{aligned}$$

To see this, notice that

$$\underbrace{C : (\mathcal{X}, \mathcal{A}) \to (\Delta, \mathcal{A}_\Delta)}_{=(1)} = C : (\mathcal{X}, \mathcal{A}) \to (\tau, \sigma(\{T_\theta : \theta \in \Theta\}))$$

and if C is $\mathcal{A}_\Delta - \mathcal{A}$-measurable, it fulfills the definition of a non-randomised decision rule (1) in the above equation. However, this means that for all $\theta \in \Theta$ the set $\{C \in \sigma(T_\theta : \theta \in \Theta)\}$ (or equivalently, $\{C \in T_\theta\}$) needs to be $\in \mathcal{A}$ for all $\theta \in \Theta$. But the set $\{C \in T_\theta\}$ is equal to the set $\{x \in \mathcal{X} : \theta \in C(x)\}$ which – by the definition of a confidence set in Equation (C.2) above – is $\in \mathcal{A}$ for all $\theta \in \Theta$. As a consequence, C is $\mathcal{A}_\Delta - \mathcal{A}$-measurable, and it also is a non-randomised decision rule because (1) in the above equation holds.

In practice, the decision functions C are often restricted to subsets of $\Delta := \tau$ like convex or closed sets or intervals. A frequently used loss function for confidence set estimation is the zero-one loss

$$L(\theta, B) := \begin{cases} 1, \theta \notin B, & \forall \theta \in \Theta, \forall B \in \Delta \\ 0, \theta \in B, & \forall \theta \in \Theta, \forall B \in \Delta \end{cases}$$

That is, when deciding for a subset $C(x) = B \in \Delta$, the incurred loss is one if $\theta \notin B$, else zero.

Example C.27 (Continuation of Example B.4)**.** In the setting of Example B.4, Example B.11 showed that a point estimator is just a special case of a decision rule. However, a confidence set incorporates the uncertainty in the estimation procedure more directly than a point estimator. In the example, $\Theta := \mathbb{R}$, and $P_\theta := \mathcal{N}(\theta, 1)^{(50)}$. Define the two-sided confidence set

$$C(x) := \left[\bar{x}_{50} - \frac{1}{\sqrt{50}} z_{\frac{\alpha}{2}}, \bar{x}_{50} + \frac{1}{\sqrt{50}} z_{\frac{\alpha}{2}}\right]$$

for a prespecified confidence level $\alpha > 0$ (typical values are $\alpha = 0.01, 0.05$ or 0.1. Here, $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$-fractile of the standard normal distribution, that is $z_{\frac{\alpha}{2}} := \Phi^{-1}(1 - \frac{\alpha}{2})$ where Φ is the cumulative distribution function of the $\mathcal{N}(0, 1)$ distribution. For $\alpha = 0.05$, it follows that $z_{\frac{\alpha}{2}} = 1.96$, and in Example B.4 $\bar{x}_{50}$ was -0.03, so the resulting confidence set is given by $C((x_1, ..., x_{50}) = [-0.307, 0.247]$.

Note that one can also randomise decision rules for confidence set estimation: $\varphi : \mathcal{X} \times \Theta \to [0, 1]$ is called a randomised confidence set if for all $\theta \in \Theta$, $\varphi(\cdot, \theta)$ is measurable. One can interpret $\varphi(x, \theta)$ for $x \in \mathcal{X}, \theta \in \Theta$ as the probability that θ is covered by the resulting confidence set $C(x)$ when observing $x \in \mathcal{X}$. Notice that again an injective embedding can be constructed by $\varphi_C(x, \theta) = \mathbb{1}_{A(\theta)}(x) = \mathbb{1}_{C(x)}(\theta)$, so it suffices to treat randomised confidence sets.

C.3.3 Hypothesis testing

Third, consider hypothesis testing. A hypothesis test from a decision-theoretic perspective is a partition of the parameter space $\Theta = H_0 \cup H_1$ where $H_1 := \Theta \setminus H_0$. Let $\Delta := \{a_0, a_1\}$, $\mathcal{A}_\Delta := \mathcal{P}(\Delta)$, and interpret a_0 as deciding for the null hypothesis H_0 and a_1 as deciding for the alternative hypothesis H_1. A decision function $\delta : \mathcal{X} \times \mathcal{A}_\Delta \to [0, 1]$ is uniquely identified by

$$\varphi := \delta(\cdot, \{a_1\}) : (\mathcal{X}, \mathcal{A}) \to ([0, 1], \mathcal{B}(\mathbb{R})_{[0,1]})$$

as $\delta(x, \{a_0\}) = 1 - \delta(x, \{a_1\})$. That is, δ is uniquely determined by specifying the values $x \in \mathcal{X}$ for which a_1 is selected. One interprets $\varphi(x, \{a_1\})$ as the probability for a decision for the alternative hypothesis H_1 when observing $x \in \mathcal{X}$. Any such map is called a *hypothesis test* and the set of hypothesis tests is denoted as Φ:

$$\Phi := \{\varphi : (\mathcal{X}, \mathcal{A}) \to ([0, 1], \mathcal{B}(\mathbb{R})_{[0,1]})\}$$

For the hypothesis testing problem, the map of all randomised decision rules $\mathcal{D} \to \Phi$ onto all hypothesis tests is bijective, $\delta \to \varphi_\delta$. $\varphi \in \Phi$ is called a non-randomised test $\Leftrightarrow \exists A \in \mathcal{A} : \varphi = \mathbb{1}_A$. A common loss function for hypothesis tests is the Neyman-Pearson loss function. For $L_0, L_1 > 0$, let

$$L(\theta, a_1) := \begin{cases} 0, \theta \in H_1 & \text{correct decision} \\ L_0, \theta \in H_0 & \text{type I error} \end{cases}$$

and

$$L(\theta, a_0) := \begin{cases} L_1, \theta \in H_1 & \text{type II error} \\ 0, \theta \in H_0 & \text{correct decision} \end{cases}$$

Example C.28 (Continuation of Example B.4)**.** Return to the setting of Example B.4. Examples B.11 and B.12 already showed that a parameter estimation and confidence set estimation correspond to specific decision rules. Now, let $H_0 := (-\infty, \theta_0]$ and $H_1 := (\theta_0, \infty]$. Let α a prespecified error level for controlling the type I error rate and $z_\alpha = \Phi^{-1}(1-\alpha)$ the $(1-\alpha)$-fractile of the $\mathcal{N}(0,1)$ distribution. The resulting test under the Neyman-Pearson-loss function with $L_0 = L_1 = 1$ is called Gauß-test φ^* and is given as

$$\varphi^*(x) := \begin{cases} 1, \sqrt{n}\frac{\bar{x}_n - \theta_0}{1} \geq z_\alpha \\ 0, \sqrt{n}\frac{\bar{x}_n - \theta_0}{1} < z_\alpha \end{cases}$$

for $x \in \mathbb{R}^n$.[4] In Example *B*.4, for $n = 50$ patients the point estimate was given as $\bar{x}_{50} = -0.03$. Let $\alpha = 0.05$ so that $z_\alpha = 1.96$. Suppose interest lies in testing H_0 : $(-\infty, -1]$ against $H_1 := (-1, \infty]$, which implies $\theta_0 := -1$. This means, one tests if the mean deviation from the reference reaction time is smaller or equal to -1, against the alternative that the mean deviation from the reference reaction time is larger than -1. Thus, the null hypothesis states that the reaction time improves by a decrease from at least 1 unit from the reference reaction time. Given the data, $\frac{\bar{x}_n - \theta_0}{1} = \sqrt{50}(-0.03 + 1) = \sqrt{50} \cdot 0.97 = 6.86 \geq 1.96$. As a consequence, $H_0 : (-\infty, -1]$ is rejected and it is concluded that the mean θ of differences in reaction times from the reference reaction time is ≥ -1. Thus, the hypothesis that the reaction time does improve by at least one unit is rejected.

C.3.4 Differences in selecting decision rules between the Bayesian and frequentist approach

Now, as shown in the preceding section, parameter estimation, hypothesis testing and confidence set estimation can be unified in the statistical decision-theoretic framework. However, the above examples were mostly from a frequentist perspective. In this section, it is shown that the difference in selecting a decision rule between the Bayesian and frequentist approach primarily lies in how the risk function is minimised, which is introduced now.

To measure the incurred loss under a selected loss function L when using a decision rule δ, the concept of a risk function is introduced. One cannot compare decision functions directly via the loss function, as the loss $L(\theta, a) = L(\theta, \delta(x))$ depends on both the decision rule δ *and* and on the observed $x \in \mathcal{X}$. As a consequence, the risk function describes the expected loss under the decision rule δ averaged over all $x \in \mathcal{X}$.

Definition C.29 (Risk function)**.** Let $(\mathcal{E}, \Delta, L)$ a statistical decision problem. The map $R : \Theta \times \mathcal{D} \to [0, \infty)$,

$$R(\theta, \delta) := \int_{\mathcal{X}} \left(\int_{\Delta} L(\theta, y) \delta(x, dy) \right) dP_\theta(x)$$

is called risk function. $R_\delta := R(\cdot, \delta)$ denotes the risk function of δ as a function on Θ.

Now, the risk function still depends on the parameter value $\theta \in \Theta$. As a consequence, one introduces a partial ordering $\preceq$ on the set $\mathcal{D}$ of all randomised decision rules as the pointwise partial ordering $\leq$ on the set of all risk functions: For all $\delta_1, \delta_2 \in \mathcal{D}$

[4] For a derivation, see Rüschendorf (2014, Chapter 6).

let $\delta_1 \preceq \delta_2 :\Leftrightarrow: R_{\delta_1} \leq R_{\delta_2}$ for all $\theta \in \Theta$, see Rüschendorf (2014, Section 2.2). Using this ordering, a frequentist strives for admissible decision functions which minimise the risk function (the expected loss over $\mathcal{X}$ for a fixed decision rule δ) over Θ:

Definition C.30 (Admissibility)**.** Let $\mathcal{D}_0 \subset \mathcal{D}$ and $\delta \in \mathcal{D}_0$.

δ_0 is $\mathcal{D}_0$-admissible, if δ_0 is minimal concerning $\preceq$ in $\mathcal{D}_0$, that is, for all $\delta \in \mathcal{D}_0$:

$$\delta \preceq \delta_0 \Rightarrow \delta \sim \delta_0$$

where $\delta_1 \sim \delta_2 \Leftrightarrow \delta_1 \preceq \delta_2$ and $\delta_2 \preceq \delta_1$. If $\mathcal{D}_0 = \mathcal{D}$, δ_0 is called *admissible*.

The definition of admissibility states that a decision rule δ_0 is admissible, if no better decision rule can be found regarding the risk. A frequentist aims at finding an admissible decision rule which minimises the risk function R_{δ_0} among all decision rules $\delta \in \mathcal{D}_0$ (ideally for $\mathcal{D}_0 = \mathcal{D}$).[5] A Bayesian statistician takes a more balanced perspective by incorporating his prior distribution to find the decision rule which minimises the risk with respect to his prior beliefs about the parameter:

Definition C.31 (Bayes decision rule)**.** Let $\mathcal{D}_0 \subset \mathcal{D}$ and $\delta \in \mathcal{D}_0$. Let (Θ, τ) a measure space, so that for all $\theta \in \Theta$, $\{\theta\} \in \tau$ and let $L : (\Theta, \tau) \otimes (\Delta, \mathcal{A}_\Delta) \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$. Let $\mu \in M^1(\Theta, \tau)$ a prior distribution for the parameter θ. The functional

$$r(\mu, \delta) := \int_\Theta R(\theta, \delta) d\mu(\theta)$$

is called the *Bayes risk* of decision rule δ with respect to the prior μ. $\delta_0 \in \mathcal{D}_0$ is called $\mathcal{D}_0$*-Bayes decision rule* with respect to the prior distribution μ, if for all $\delta \in \mathcal{D}_0$

$$r(\mu, \delta_0) \leq r(\mu, \delta)$$

If $\mathcal{D}_0 = \mathcal{D}$, then δ_0 is called *Bayes decision rule* with respect to μ.

In the Bayesian approach, the statistician searches a decision rule which minimises the risk function with respect to his prior assumptions, expressed in form of the prior distribution on (Θ, τ).[6] Based on the risk function, statistical decision theory then allows to select decision rules based on specific criteria. The most prominent examples are admissibility and Bayes rules in the frequentist and Bayesian approach, other examples include minimaxity, for details see Rüschendorf (2014, Chapter 2).

C.3.5 Frequentist and Bayesian risk procedures from the decision-theoretic perspective

This section provides some intuitions how point estimation, confidence set estimation and hypothesis testing differ from a decision-theoretic perspective in the frequentist and Bayesian approach. Therefore, a few risk functions are derived and compared.

[5] Often, frequentists also search for minimax decision rules, that is, decision rules $\tilde{\delta}$ which minimise the maximum risk $R(\theta, \tilde{\delta}) := \sup_{\theta \in \Theta} R(\theta, \tilde{\delta})$. Although minimax decision rules often exist, they are rather pessimistic as Kleijn (2022) notes, because they optimise only with respect to the worst-case scenario. One can show that any Bayesian risk function (that is, the risk function integrated with respect to the prior distribution) is upper bounded by the minimax risk, see (Kleijn, 2022, Proposition 2.4). This strongly questions the use of minimax rules from a Bayesian perspective.

[6] Note that while it may seem a difficult task to find a Bayes decision rule, under quite general assumptions, a Bayes rule can be found by minimising the posterior expected loss. For details, see Kleijn (2022, Theorem 2.5) and (Robert, 2007, Chapter 2).

Example C.32 (Continuation of Example B.4)**.** If $\delta = \delta_d$ with $d \in D$, that is, δ is a non-randomised decision rule, then the risk is given as

$$R(\theta, d) := R(\theta, \delta_d) = \int_X L(\theta, d(x)) dP_\theta(x) = \mathbb{E}_\theta[L(\theta, d)]$$

For an estimation problem with Gauß loss L_2 and $\Delta = \mathbb{R}$, the risk of a non-randomised decision rule (a non-randomised estimator) $d \in D$ is

$$R(\theta, d) = \int_{\mathcal{X}} (d(x) - g(\theta))^2 dP_\theta(x) = \mathbb{E}_\theta[(d - g(\theta))^2]$$

In the situation of Example B.4, $\delta(x) := \bar{x}_{50} = -0.03$ and for the Gauß-loss L_2, $L(\delta(x), \theta) = (-0.03 - \theta)^2$. The risk is given as $\mathbb{E}_\theta[L(\theta, d)] = \mathbb{E}_\theta[(-0.03 - \theta)^2]$.

Example C.33 (Continuation of Example B.12)**.** In the setting of Example B.12, for the confidence set C the zero-one-loss yields the risk

$$\begin{aligned} R(\theta, C) &= 1 \cdot P_\theta(\{x \in \mathcal{X} : \theta \notin C(x)\}) + 0 \cdot P_\theta(\{x \in \mathcal{X} : \theta \in C(x)\}) \\ &= P_\theta(\{x \in \mathcal{X} : g(\theta) \notin C(x)\}) \end{aligned}$$

which is the probability of the parameter θ not being covered by $C(x)$.

Example C.34 (Continuation of Example B.13)**.** We return to Example B.13. The risk function of a hypothesis test under the Neyman-Pearson-loss is given as

$$\begin{aligned} R(\theta, \varphi) &= \int_{\mathcal{X}} \int_\Delta L(\theta, y) \delta_\varphi(x, dy) dP_\theta(x) \\ &= \int_{\mathcal{X}} \int_{\{a_0, a_1\}} L(\theta, y) \delta_\varphi(x, dy) dP_\theta(x) \\ &= \int [L(\theta, a_1)\varphi(x) + L(\theta, a_0)(1 - \varphi(x))] dP_\theta(x) \\ &= \begin{cases} L_0 \mathbb{E}_\theta[\varphi], & \theta \in H_0 \quad \text{type I error} \\ L_1 \mathbb{E}_\theta[(1 - \varphi)], & \theta \in H_1 \quad \text{type II error} \end{cases} \end{aligned}$$

Notice that the risk of φ is dependent only on the expectation $\mathbb{E}_\theta[\varphi]$. Importantly, this implies that even tests which minimise the risk (and are then called admissible), only guarantee statements about error probabilities in expectation, and not for individual tests carried out for an experiment or a study. This expresses mathematically what Neyman and Pearson stressed regarding the option to state anything about the truth of a hypothesis based on a Neyman-Pearson test, compare Chapter 4: The Neyman-Pearson theory is not able to quantify the truth of a hypothesis or the error probability of a hypothesis for a single experiment or study. For the Gauß-test in Example B.13, one can show that the risk function is constant and equal to α, for details see Rüschendorf (2014, p. 27).

From a frequentist perspective, finding an admissible decision rule δ_0 with $\delta \preceq \delta_0 \Rightarrow \delta \sim \delta_0$ in the above three examples would be the goal from a decision-theoretic point of view. From the Bayesian perspective, one would combine the risk functions above with a prior distribution $\mu_\Theta \in M^1(\Theta, \tau)$ on the Bayesian parameter space (Θ, τ). Subsequently, a Bayesian would look for a Bayes decision rule by minimising the Bayes risk

$$r(\mu, \delta) = \int_\Theta R(\theta, \delta) d\mu(\theta) \tag{C.3}$$

Thus, for a Bayesian examples B.14, B.15 and B.16 change by plugging the resulting risk functions $R(\theta, \delta)$ in examples B.14, B.15 and B.16 into equation Equation (C.3) and searching for a Bayes decision rule.

Note that the usually used quantities like the posterior mean or median can be derived as the corresponding Bayes decision rule under a specific loss function (L_2 loss for the mean, L_1-loss for the median). Decision theory justifies these "intuitive" Bayesian estimators for a parameter by guaranteeing that they minimise the risk under a given loss function with respect to the prior distribution. Details can be found in Held and Sabanés Bové (2014) and (Rüschendorf, 2014, Chapter 2). For a decision-theoretic justification of Bayesian confidence sets like highest-posterior-density intervals see Schervish (1995, Section 5.2.5), and for a decision-theoretic derivation of the Bayes factor see Robert (2007, p. 224-227).

C.4 Maximum-Likelihood estimation

To estimate the unknown parameter θ, the method of maximum likelihood presents a general approach. The *likelihood function* is defined as

Definition C.35 (Likelihood function). Let $\mathcal{P} \ll \mu$ and $f_\theta = dP_\theta/d\mu, \theta \in \Theta$. The density f_θ as a function of θ is called likelihood function, and defined for $x \in \mathcal{X}$ as

$$L(\theta; x) = f_\theta(x) = f(x; \theta) \tag{C.4}$$

for $\theta \in \Theta$.

In the above, $\ll$ denotes absolute continuity between the measures, that is for every measurable set A, $\mu(A) = 0$ implies $P(A) = 0$ for all $P \in \mathcal{P}$. The estimate which maximises the likelihood is the maximum likelihood estimator:

Definition C.36 (Maximum-Likelihood-Estimator). Let Θ be associated with a σ-algebra τ. A measurable map $\hat{\theta}_{ML} : \mathcal{X} \mapsto \Theta$ is called Maximum-Likelihood-Estimator, if

$$L(\hat{\theta}_{ML}(x)) := \sup_{\theta \in \Theta} L(\theta; x) \quad P_\theta\text{-almost surely} \tag{C.5}$$

To simplify notation, often instead of the likelihood function $L(\theta)$ only the *likelihood kernel* is reported and also denoted with $L(\theta)$, as multiplicative constants have no influence on $\hat{\theta}_{ML}$.

Definition C.37 (Likelihood kernel). The likelihood kernel is obtained from a likelihood function by removing all multiplicative constants.

The symbol $L(\theta)$ is used both for likelihood functions and kernels.

Definition C.38 (Log-likelihood function). The log-likelihood function is given as

$$l(\theta; x) := \log\, L(\theta; x) \tag{C.6}$$

The log-likelihood function can be used instead of $L(\theta; x)$ to derive $\hat{\theta}_{ML}$ as the logarithm is strictly monotone and therefore

$$\hat{\theta}_{ML} := \underset{\theta \in \Theta}{\arg\max}\; l(\theta; x) \tag{C.7}$$

Also in this case, often only the log-likelihood kernel will be reported and depending on the context, also be denoted by $l(\theta; x)$. An important property of the maximum likelihood estimate (MLE) is given by the following fact (compare Held and Sabanés Bové (2014, p. 24)):

Theorem C.39 (Invariance of the MLE)**.** Let $\hat{\theta}_{ML}$ be the MLE of θ, and let $\phi = h(\theta)$ be a one-to-one transformation of θ. The MLE of ϕ can be obtained by inserting $\hat{\theta}_{ML}$ in $h(\theta)$, and the MLE of ϕ is given as $\hat{\phi}_{ML} = h(\hat{\theta}_{ML})$.

That is, the MLE is invariant under one-to-one transformations.

C.5 Foundations of Point Estimation

The standard definition of a point estimator expresses the decision-theoretic perspective outlined above, compare Casella and Berger (2002, Definition 7.1.1), Rüschendorf (2014, p. 124):

Definition C.40 (Point Estimator)**.** A point estimator is any function $d : (\mathcal{X}, \mathcal{A}) \to (Y, \mathcal{C})$ of a random sample $(X_1, ..., X_n)$, where $(Y, \mathcal{C})$ is a measure space.

A common choice for parameter estimation is $(\mathcal{Y}, \mathcal{C}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

If, for example $\Theta := \mathbb{R}^n$ for $n \in \mathbb{N}$ and parameter estimation is the goal, Definition C.40 can be used by setting $\Delta = \mathbb{R}^n$ and $\mathcal{A}_\Delta := \mathcal{B}(\mathbb{R}^n)$, and choosing for example d as the sample mean or median. A loss function (like the L_2 loss) can then be used to quantify the risk of the estimator d. A point estimator d then becomes a non-randomized decision function of the decision-theoretic framework. Using the Gauß-loss L_2, also called the mean squared loss or mean squared error (MSE), one can compare different (point) estimators.

To compare multiple estimators, a widespread measure is given by the mean squared error (MSE), which assumes a quadratic loss function $L(\theta, d) := (\theta - d)^2$ for a (point) estimator d of a parameter θ, compare Casella and Berger (2002, Definition 7.3.1) and Rüschendorf (2014, p. 124):

Definition C.41 (Mean Squared Error)**.** The mean squared error (MSE) of an estimator d of a parameter θ is the function of θ defined by:

$$\mathbb{E}_\theta((d - \theta)^2)$$

Using the MSE and the familiar bias-variance-decomposition (Held and Sabanés Bové, 2014), one can introduce the concept of bias and unbiasedness of an estimator naturally, compare Casella and Berger (2002, Definition 7.3.2):

Definition C.42 (Bias of a point estimator)**.** The bias of a point estimator d of a parameter θ is given by

$$\text{Bias}_\theta(d) = \mathbb{E}_\theta(d) - \theta \tag{C.8}$$

An estimator whose bias is identically (in θ) equal to 0 is called unbiased and satisfies $\mathbb{E}_\theta(d) = \theta$ for all $\theta \in \Theta$.

When searching for the best estimator, that is, the estimator minimizing the MSE in Definition C.41, one often restricts attention to unbiased estimators. Then, only the variance is of importance as the bias part reduces to zero. Instead of unbiased estimators, it is purposeful to direct attention to the class of estimators

$$\mathcal{C}_g = \{d : \mathbb{E}_\theta(d) = g(\theta)\}$$

which includes unbiased estimators if the parameter function $g : \Theta \mapsto \Delta$ is selected as the identity $g(\theta) = \theta$ for $\theta \in \Theta$. This leads to the definition of best unbiased estimators, compare Casella and Berger (2002, Definition 7.3.7) and Rüschendorf (2014, Definition 5.1.5):

Definition C.43 (Best unbiased estimator (UMVUE))**.** An estimator d^* is a best unbiased estimator of $g(\theta)$ if it satisfies $\mathbb{E}_\theta(d^*) = g(\theta)$ for all θ and, for any other estimator d with $\mathbb{E}_\theta(d) = g(\theta)$, we have

$$\mathbb{V}_\theta(d^*) \leq \mathbb{V}_\theta(d) \tag{C.9}$$

for all $\theta \in \Theta$. The estimator d^* is also called a uniform minimum variance unbiased estimator (UMVUE) of $g(\theta)$.

For a comparison of unbiased estimators the Cramér-Rao-Inequality is essential. It gives a minimum variance bound for unbiased estimators $d(X)$ for a parameter θ, so if an unbiased estimator is found attaining the minimum variance bound, it is a UMVUE. The Cramér-Rao-Inequality is stated as follows, compare Casella and Berger (2002, Theorem 7.3.9) and Rüschendorf (2014, Theorem 5.4.6):

Theorem C.44 (Cramér-Rao-Inequality)**.** Let $X_1, ..., X_n$ be a random sample following probability density function $f_\theta := f(x|\theta)$, and let $d(X) = d(X_1, ..., X_n)$ be any estimator satisfying

$$\frac{\partial}{d\theta}\mathbb{E}_\theta(d) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[d(x)f(x|\theta)]dx \tag{C.10}$$

and

$$\mathbb{V}_\theta(d(X)) < \infty \tag{C.11}$$

Then

$$\mathbb{V}_\theta(d(X)) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(d(X)))^2}{\mathbb{E}_\theta[(\frac{\partial}{\partial\theta}\log f(X|\theta))^2]} \tag{C.12}$$

In the i.i.d. case the Cramer-Rao-Inequality reduces to (Casella and Berger, 2002, Corollary 7.3.10):

Corollary C.45 (Cramér-Rao-Inequality (i.i.d. case))**.** If the assumptions for Theorem C.44 are satisfied, and, additionally $X_1, ..., X_n$ are i.i.d. with probability density function $f(x|\theta)$, then

$$\mathbb{V}_\theta(d(X)) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(d(X)))^2}{n \cdot \mathbb{E}_\theta[(\frac{\partial}{\partial\theta}\log f(X|\theta))^2]} \tag{C.13}$$

Due to the fact that for an unbiased estimator $(\frac{\partial}{\partial\theta}\mathbb{E}_\theta(d(X)))^2 = (\frac{\partial}{\partial\theta}\theta)^2 = 1$, the lower variance bound is given by $\frac{1}{\mathbb{E}_\theta[(\frac{\partial}{\partial\theta}\log f(X|\theta))^2]}$. This is the inverse of the quantity in the denominator of Equation (C.12) in the Cramér-Rao-Inequality, and the quantity $\mathbb{E}_\theta[(\frac{\partial}{\partial\theta}\log f(X|\theta))^2]$ is called the Fisher-Information, compare Rüschendorf (2014, p. 157)

Definition C.46 (Fisher Information). The quantity

$$I(\theta) := \mathbb{E}_\theta \left[(\frac{d}{d\theta} \log f(X|\theta))^2 \right] \tag{C.14}$$

is called the Fisher-Information of the statistical model $\mathcal{P}$ in θ.

It can be shown that the Fisher-Information is the variance of the log-likelihood $l(\theta; x)$, so that a large Fisher-Information at the maximum likelihood estimate puts trust into the maximum likelihood estimate, while a small Fisher-Information indicates a flat likelihood curve, questioning the trust that can be put into the maximum likelihood estimate. By the Cramér-Rao-Inequality, the best possible efficiency of an unbiased estimator is given by the inverse Fisher-Information and therefore the theorem gives a lower bound on the variance of a UMVUE.

A desirable large sample property for an estimator is consistency, compare Casella and Berger (2002, p. 468):

Definition C.47 (Consistency). A sequence of (point) estimators $d_n := d_n(X_1, ..., X_n)$ is a consistent sequence of estimators for the parameter θ if, for every $\varepsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|d_n(X_1, ..., X_n) - \theta| < \varepsilon) = 1 \tag{C.15}$$

A consistent sequence of point estimators converges in probability to the true parameter θ, and therefore, as the sample becomes infinite and $n \to \infty$ the estimator will be arbitrarily close to θ with probability converging to one. The concept of an efficient estimator extends the requirement of unbiasedness of the estimator to additionally attaining the Cramér-Rao lower bound for the variance:

Definition C.48 (Efficiency). A sequence of estimators d_n is asymptotically efficient for a parameter function $g(\theta)$, if $\sqrt{n}[d_n - g(\theta)] \xrightarrow{d} \mathcal{N}(0, \nu(\theta))$ in distribution and

$$\nu(\theta) = \frac{[g'(\theta)]^2}{\mathbb{E}_\theta((\frac{\partial}{\partial\theta} \log f(X|\theta))^2)} \tag{C.16}$$

that is, d_n is unbiased for $g(\theta)$ and the asymptotic variance of d_n achieves the Cramér-Rao Lower Bound.

Now, if the Fisher-Regularity-Conditions hold (see Held and Sabanés Bové (2014, p. 80) or Rüschendorf (2014, Chapter 5.4)), the MLE is asymptotic normal and efficient (Casella and Berger, 2002, p. 516):

Theorem C.49. Let $X_1, X_2, ...,$ be i.i.d with probability density $f(x|\theta)$, let $\hat{\theta}_{ML}$ denote the MLE of θ, and let $g(\theta)$ be a continuous parameter function of $\theta \in \Theta$. Under the Fisher-regularity-conditions on $f(x|\theta)$ and, hence, $L(\theta; x)$,

$$\sqrt{n}[g(\hat{\theta}_{ML}) - g(\theta)] \xrightarrow{D} \mathcal{N}(0, \nu(\theta)) \tag{C.17}$$

where $\nu(\theta)$ is the Cramér-Rao Lower Bound. That is, $g(\hat{\theta}_{ML})$ is a consistent and asymptotically efficient estimator of $g(\theta)$.

Next to unbiasedness and consistency, another important element of likelihood inference is the concept of sufficiency (Casella and Berger, 2002, Definition 6.2.1):

Definition C.50 (Sufficiency)**.** A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Thus, a statistic $T = h(X_{1:n})$ is sufficient for θ if the conditional distribution of $X_{1:n}$ given $T = t$ is independent of θ, i.e. if

$$f(X_{1:n}|T = t) \tag{C.18}$$

does not depend on θ, compare Held and Sabanés Bové (2014). Here, $X_{1:n}$ denotes the sample $(X_1, ..., X_n)$ of size n. In an intuitive way, a statistic achieves sufficiency, if it captures the whole information of the data at hand, that is no loss of information is associated with the reduction of the data to the statistic $T(X_{1:n})$. From a measure-theoretic point of view, the compression of data happens when instead of $x \in \mathcal{X}$, the statistic $T(x)$ (e.g. a point estimator) is provided. This data compression can be described via sub-σ-algebras $\mathcal{B} \subset \mathcal{A}$, where sufficiently large sub-σ-algebras are sufficient to describe an experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$. As a consequence, a statistic T is called sufficient, if the generated σ-algebra $\sigma(T)$ is a sufficient σ-algebra, for details see Rüschendorf (2014, p. 82). In general, let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ a probability space and $\mathcal{B} \subset \mathcal{A}$ a σ-algebra. A $\mathcal{B}$-measurable function $f_A \in \mathcal{L}(\mathcal{X}, \mathcal{B})$ is then called the *conditional probability* of A given $\mathcal{B}$, written $f_A = P(A|\mathcal{B})$ if the Radon-Nikodým-equation

$$P(A \cap B) = \int_B f_A dP \qquad \forall B \in \mathcal{B}$$

holds. A $\mathcal{B}$-measurable function $Y \in \mathcal{L}(\mathcal{X}, \mathcal{B})$ is called *conditional expectation* of $X \in \mathcal{L}(\mathcal{X}, \mathcal{A})$, written $Y = \mathbb{E}[X|\mathcal{B}]$, if

$$\int_B X dP = \int_B Y dP \qquad \forall B \in \mathcal{B}$$

The Radon-Nikodým theorem guarantees the existence and $\mathcal{P}$-almost sure uniqueness of the conditional expectation (Bauer, 2001). Now, given a statistical model $(\mathcal{E}, \mathcal{A}, \mathcal{P})$, a σ-algebra $\mathcal{B} \subset \mathcal{A}$ is called *sufficient* for $\mathcal{P}$, if for all $A \in \mathcal{A}$ there exists an $f_A \in \mathcal{L}(\mathcal{X}, \mathcal{B})$ so that

$$f_A = P(A|\mathcal{B}) \qquad P\text{-almost surely} \qquad \forall P \in \mathcal{P}$$

which means the Radon-Nikodým equation holds (for all $A \in \mathcal{A}$). A statistic $T : (\mathcal{X}, \mathcal{A}) \to (Y, \mathcal{C})$ is called sufficient for $\mathcal{P}$, if for all $A \in \mathcal{A}$ there exists an $f_A \in \mathcal{L}(\mathcal{X}, \sigma(T))$, so that

$$f_A = P(A|T) \qquad P\text{-almost surely} \qquad \forall P \in \mathcal{P}$$

To understand the definition, suppose $\mathcal{B}$ is sufficient for $\mathcal{P}$. Then

$$P(A) = \int_{\mathcal{B}} P(A|\mathcal{B}) dP = \int P(A|\mathcal{B}) dP|_{\mathcal{B}} = \int f_A dP|_{\mathcal{B}} \tag{C.19}$$

where in the last equality the assumption $f_A = P(A|\mathcal{B})$, P-almost surely for all $P \in \mathcal{P}$ was used. As a consequence, the probability measures $P \in \mathcal{P}$ differ only on the sub-σ-algebra $\mathcal{B}$, and this shows that sufficiency of a sub-σ-algebra and a statistic implies no loss of information, although the transition from the original σ-algebra $\mathcal{A}$ to the sub-σ-algebra $\mathcal{B}$ can be interpreted as a compression of data. Equivalently, the transition from the original sample $X_1, ..., X_n$ in $\mathcal{A}$ to the statistic $T(X_1, ..., X_N)$ in $\mathcal{B}$ can be interpreted as the compression of data.

The Fisher-Neyman factorization theorem provides a convenient characterization of a sufficient statistic, compare Casella and Berger (2002, Theorem 6.2.6).

Theorem C.51 (Neyman-Fisher factorization Theorem)**.** Let $f(x|\theta)$ denote the joint probability mass or density function of a random sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,

$$f(x|\theta) = g(T(x)|\theta)h(x) \tag{C.20}$$

A more precise measure-theoretic version of the Neyman-Fisher factorization is given by Rüschendorf (2014, Theorem 4.1.15):

Theorem C.52 (Neyman-Fisher factorization Theorem)**.** $T : (\mathcal{X}, \mathcal{A} \to (\mathcal{Y}, \mathcal{C})$ is sufficient for the statistical model $\mathcal{P}$ if and only if there exist functions $h : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+)$ and for all $P \in \mathcal{P}$, $g_P : (\mathcal{Y}, \mathcal{C}) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+)$ so that

$$\frac{dP}{d\mu}(x) = g_P(T(x))h(x) \tag{C.21}$$

μ-almost surely, where μ is the dominating measure for the statistical model $\mathcal{P}$, that is, $\mathcal{P} << \mu$.

The concept of minimal sufficiency answers the question for the largest possible data compression that is possible, see Casella and Berger (2002, Definition 6.2.11) and Rüschendorf (2014, Section 4.2):

Definition C.53 (Minimal sufficient statistic)**.** A sufficient statistic $T(X)$ is called minimal sufficient if, for any other sufficient statistic $T'(X)$, $T(X)$ is a function of $T'(X)$.

This means, that $T'(x) = T'(y)$ implies $T(x) = T(y)$. To identify minimal sufficient statistics, the following theorem is helpful (Casella and Berger, 2002, 6.2.13):

Theorem C.54. Let $f(x|\theta)$ be the joint probability mass function or probability density of a random sample X. Suppose there exists a function $T(x)$, such that, for every two sample points x and y, the ratio $\frac{f(x|\theta)}{f(y|\theta)}$ is constant as a function of θ if and only if $T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

The Rao-Blackwell-Theorem formalizes how one can improve upon an existing unbiased estimator by using a sufficient statistic, for details see Casella and Berger (2002, Theorem 7.3.17) and Rüschendorf (2014, p. 130):

Theorem C.55 (Rao-Blackwell)**.** Let d be any unbiased estimator of $g(\theta)$ (where $g : \Theta \mapsto \Delta$), and let T be a sufficient statistic for θ. Define $\phi(T) := \mathbb{E}(d|T)$. Then $\mathbb{E}_\theta(\phi(T)) = g(\theta)$ and $\mathbb{V}_\theta(\phi(T)) \leq \mathbb{V}_\theta(d)$ for all θ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $g(\theta)$ than d.

Again, the most common case in the above is when $g(\theta) := \theta$ is the identity. If a best unbiased estimator is found (where best is interpreted as the estimator having minimum variance), one can show that it is unique (Casella and Berger, 2002, Theorem 7.3.19).

Theorem C.56. If d is a best unbiased estimator of $g(\theta)$, then d is unique.

To find a best unbiased estimator, the correlation of the estimator at hand with unbiased estimators of zero is essential (Casella and Berger, 2002, Theorem 7.3.20).

Theorem C.57. If $\mathbb{E}_\theta(W) = g(\theta)$, d is the best unbiased estimator of $g(\theta)$ if and only if d is uncorrelated with all unbiased estimators of 0.

Therefore, if there were no such unbiased estimators of zero, any statistic d would satisfy $\text{Cov}_\theta(d, 0) = 0$ and is a UMVUE. The following property of completeness is essential, as it guarantees such a situation (Casella and Berger, 2002, Definition 6.2.21):

Definition C.58 (Completeness). Let $f(t|\theta)$ be a family of probability densities or probability mass functions for a statistic $T(X)$. The family of probability distributions is called complete if $\mathbb{E}_\theta[g(T)] = 0$ for all θ implies $\mathbb{P}_\theta(g(T) = 0) = 1$ for all θ. Equivalently $T(X)$ is called a complete statistic.

Completeness in a measure-theoretic interpretation can be introduced by requiring that a sub-σ-algebra $\mathcal{T} \subset \mathcal{A}$ for the statistical model $\mathcal{P}$ is sufficiently small, or equivalently, only large enough to separate between all elements in $\mathcal{L}^1(\mathcal{X}, \mathcal{T}, \mathcal{P})$. Interestingly, complete and sufficient sub-σ-algebras are (if they exist) minimal sufficient. From a measure-theoretic point of view, first, the set of $\mathcal{P}$-zero-estimators is introduced as

$$D_0(\mathcal{P}) := \{f \in \mathcal{L}^1(\mathcal{X}, \mathcal{T}, \mathcal{P}) : \mathbb{E}_P[f] = 0 \text{ for all } P \in \mathcal{P}\}$$

A class of distributions $\mathcal{P} \subset M^1((\mathcal{X}, \mathcal{A})$ is complete, if for all $f \in D_0$, the equality $f = 0$ $\mathcal{P}$-almost surely holds (Rüschendorf, 2014, p. 105-106). A statistic $T : (\mathcal{X}, \mathcal{A}) \to (Y, \mathcal{B})$ is called complete, if $\mathcal{P}|_{\sigma(T)}$ is complete. Note the analogy to Definition C.58. Phrased differently, completeness of a class of distributions $\mathcal{P}$ states that there exist only trivial estimators of zero. The relationship between best unbiased estimators and completeness is described as follows (Casella and Berger, 2002, Theorem 7.3.23).

Theorem C.59. Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based on T only. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Compared with sufficient statistics, a statistic with a complementary purpose is given by ancillary statistics (Casella and Berger, 2002, 6.2.16):

Definition C.60 (Ancillary statistic). A statistic $T(X)$ whose distribution does not depend on the parameter θ is called an ancillary statistic.

Ancillary statistics, when used as a complement for a statistic can help to recover information that is lost by the statistic. From a measure-theoretic perspective, ancillarity is first defined for σ-algebras similar to the concept of sufficiency:

Definition C.61 (Ancillarity). A σ-algebra $\mathcal{B} \subset \mathcal{A}$ is called ancillary for $\mathcal{P}$, if for all $B \in \mathcal{B}$ and $P, Q \in \mathcal{P}$ the following equality holds: $P(B) = Q(B)$. A statistic $S : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ is called ancillary, if $\sigma(S)$ is an ancillary σ-algebra. This is true if and only if for all $P, Q \in \mathcal{P}$ we have: $P^S = Q^S$.

While a sufficient statistic contains all relevant information about the parameter(s) without any loss of information due to data compression, an ancillary statistic contains no information about the parameter.[7]

[7] It is possible to show that an ancillary statistic for the model $\mathcal{P}$ has Fisher-information $I_T(\theta) = 0$,

Example C.62. In a normal distribution model $P_\theta = \mathcal{N}(\theta, 1)^{(n)}$, $\theta \in \Theta := \mathbb{R}^1$, the estimator $T(x) := \bar{x}_n$ is sufficient for $g(\theta) = \theta$. The function $S(x) := \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2$ is an estimator for the constant variance $\sigma^2 = 1$ of the statistical model. From S no information about the parameter θ can be obtained. As a consequence, S is an ancillary statistic. Clearly, for two distributions P and Q from $\mathcal{P}$, the only difference is the mean θ, so that one can express this as $Q = P_0$ where it is assumed that $\theta = 0$ without loss of generality, and $P = P_\theta$, shifted by some $\theta \in \Theta$. As a consequence the probability P^S and Q^S of observing a specific value for $S(x)$ are identical, that is, for any $B \in \mathcal{B}$ with $S(x) \in B$ the equality $P^S(B) = Q^S(B)$ holds.

One of the most important connections between minimal sufficiency, completeness and ancillarity is given by Basu's Theorem (Casella and Berger, 2002, Theorem 6.2.24):

Theorem C.63 (Basu's Theorem). If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

A more measure-theoretic formulation of Basu's theorem is given by Rüschendorf (2014, p. 112-113), and a brief statement in the decision-theoretic framework from above is: Let $(X, \mathcal{A}, \mathcal{P})$ a statistical model, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of distribution indexed by the parameter(s) θ, and T and A maps from $(\mathcal{X}, \mathcal{A})$ to some measurable decision space $(\Delta, \mathcal{A}_\Delta)$. If T is a boundedly complete sufficient statistic for θ, and A is ancillary to θ, then T is independent of A. Here, $\mathcal{P}$ is called boundedly complete if for all $f \in D_0 \cap \mathcal{B}(\mathcal{X}, \mathcal{A})$ the equality $f = 0$ holds $\mathcal{P}$-almost surely. $\mathcal{B}(\mathcal{X}, \mathcal{A})$ is the set of bounded $\mathcal{A}$-measurable functions. A statistic $T : (\mathcal{X}, \mathcal{A}) \to (Y, \mathcal{B})$ is called boundedly complete, if $\mathcal{P}|_{\sigma(T)}$ is boundedly complete.

Given the appeal of minimal sufficient statistics, one of the main reasons for the popularity of maximum likelihood estimation as an inference method is due to the following fact (Held and Sabanés Bové, 2014, Chapter 2).

Theorem C.64 (Minimal sufficiency of the likelihood). The likelihood function L(θ;x) is minimal sufficient.

A measure-theoretic proof can be found in Rüschendorf (2014, Theorem 4.2.9) and Rüschendorf (2014, Corollary 4.2.9). As a consequence, the likelihood function $L(\theta; x)$ achieves the maximum possible data compression without causing any loss of information.

C.6 Foundations of Hypothesis Testing

In a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ we are interested in the testing problem $\Theta = \Theta_0 \cup \Theta_1$, where $H_0 : \theta \in \Theta_0$ denotes the null hypothesis and $H_1 : \theta \in \Theta_1$ denotes the alternative hypothesis.

Definition C.65 (Null and alternative hypothesis). The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_1 = \Theta \setminus \Theta_0$.

that is, it contains no information about the model parameter(s). See Rüschendorf (2014, Proposition 5.4.4) for the definition of $I_T(\theta)$ (the Fisher-Information of a statistic T) and a proof.

In a formal measure-theoretic sense, the hypothesis testing problem is interpreted as a binary decision problem, where the experimenter either has to accept or reject H_0. It is important to note, that on a philosophical level, this is a strong simplification, as the decisions 'accepting H_0' and 'not rejecting H_0' are two non-identical events. Also, 'rejecting H_0' is a different event as 'accepting H_1' whenever hypotheses are tested for example in scientific research, compare Part I and Part II. However, the measure-theoretic and decision-theoretic definition of a hypothesis test is as follows (Casella and Berger, 2002, Definition 8.1.3):

Definition C.66 (Hypothesis test)**.** A statistical hypothesis test is a (decision) rule that specifies:

i) For which sample values the decision is made to accept H_0 as true.

ii) For which sample values the null hypothesis H_0 is rejected and H_1 is accepted as true.

This vague definition is made more precise by introducing critical functions, also called randomized tests (Rüschendorf, 2014):

Definition C.67 (Critical function / randomized test)**.** A function

$$\varphi : (\mathcal{X}, \mathcal{A}) \mapsto ([0,1], \mathcal{B}(\mathbb{R})|_{[0,1]})$$

is called a critical function (or randomized test), and

$$\Phi := \{\varphi : (\mathcal{X}, \mathcal{A}) \mapsto ([0,1], \mathcal{B}(\mathbb{R})|_{[0,1]})\}$$

is the set of all critical functions (or randomized tests).

Traditionally, hypothesis tests were described by rejection and acceptance regions (Casella and Berger, 2002).

Definition C.68 (Rejection and acceptance region)**.** The subset $R \subset \Theta$, for which H_0 will be rejected is called the rejection region. The complement $R^c \subset \Theta$ is called the acceptance region.

In a measure-theoretic interpretation, the value of the critical function (or randomized test) is interpreted as the probability to reject H_0. Consequentially, from a measure-theoretic interpretation, one can specify the rejection and acceptance region:

Definition C.69 (Rejection and acceptance region)**.** The subset $\mathcal{C} := \{x \in \mathcal{X} : \varphi(x) = 1\}$ is called the rejection region. The complement is called the acceptance region.

In practice, the definition of a critical region is central:

Definition C.70 (Critical region for level α)**.** Let $\alpha \in [0,1]$. Then C is called a critical region for level α for H_0 against H_1 if

$$\forall \theta \in H_0 : P_\theta(C) \leq \alpha$$

A hypothesis test for level α is then defined as follows:

Definition C.71 (Hypothesis test for level α for H_0 against H_1)**.** A non-randomized test φ is called hypothesis test for level α for H_0 against H_1, if

$$\mathbb{E}_\theta[\varphi] := \int \varphi dP_\theta \leq \alpha$$

The power function is important to judge the property of a test to correctly reject the null hypothesis when it is false:

Definition C.72 (Power function). Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ a statistical model and $\Theta = \Theta_0 \cup \Theta_1$ be a partition of the parameter space Θ into the null hypothesis $H_0 : \theta \in \Theta_0$ and alternative hypothesis $H_1 : \theta \in \Theta_1$. Let $\varphi : (\mathcal{X}, \mathcal{A}) \mapsto ([0,1], \mathcal{B}(\mathbb{R})|_{[0,1]})$ a randomized test. Then, the function

$$G_\varphi : \Theta \mapsto [0,1], G_\varphi(\theta) := \mathbb{E}_\theta[\varphi]$$

is called the power function for the randomized test φ at $\theta \in \Theta$.

To evaluate tests, the probabilities of making mistakes is compared in the frequentist Neyman-Pearson theory. Therefore, the classic error probabilities and the power function are necessary (Casella and Berger, 2002).

Definition C.73 (Type I Error). If $\theta \in \Theta_0$ but the hypothesis test φ incorrectly decides to reject H_0, then the test has made a *type I error*. The type I error probability is given as $G_\varphi(\theta) = \mathbb{E}_\theta[\varphi]$ for $\theta \in \Theta_0$.

Definition C.74 (Type II Error). If $\theta \in \Theta_1$ but the hypothesis test φ incorrectly accepts H_0, then the test has made a *type II error*. The type II error probability is given as $1 - G_\varphi(\theta) = 1 - \mathbb{E}_\theta[\varphi]$ for $\theta \in \Theta_1$.

For an overview, see Table C.1: Notice that the power function is the probability of making a type I error, if the null hypothesis is true. It is the probability of correctly rejecting the null hypothesis, if the alternative hypothesis is true. Of course, the ideal power function would be zero for all $\theta \in \Theta_0$ and one for all $\theta \in \Theta_0^c$, but this is not possible as type I and type II error probabilities balance each other out.

		Decision	
		Accept H_0, $x \notin \mathcal{C}$, $\varphi(x) = 0$	Reject H_0, $x \in \mathcal{C}$, $\varphi(x) = 1$
Truth	H_0, $\theta \in \Theta_0$	Correct decision	Type I Error, $\mathbb{E}_\theta[\varphi]$, $P_\theta(C)$
	H_1, $\theta \in \Theta_1$	Type II Error, $1 - \mathbb{E}_\theta[\varphi]$, $1 - P_\theta(C)$	Correct decision

Table C.1: Type I and type II errors in hypothesis Tests

The most universal method to find a hypothesis test is based on the likelihood function. This is one of the simplest reasons why likelihood theory and frequentist statistical hypothesis testing according to the Neyman-Pearson theory have been and are still widely applied in research. The corresponding statistic is called the likelihood ratio test statistic, see Casella and Berger (2002, Definition 8.2.1) and Rüschendorf (2014, p. 189-190).

Definition C.75 (Likelihood ratio test). Let $\Theta_i := \{\theta_i\}$ for $i = 1, 2$, $P_i := P_{\theta_i}$ with density f_i and $L := f_1/f_0$, where $a/0 := \infty$ for all $a > 0$ and $0/0 := 0$. ϕ is called likelihood ratio test, if it has the following form:

$$\varphi(x) = \begin{cases} 1, \text{ if } L(x) > k \\ \gamma(x), \text{ if } L(x) = k \ P_1 + P_2 \text{ almost surely} \\ 0, \text{ if } L(x) \leq k \end{cases} .$$

k is called critical value of φ and $\{\varphi = k\}$ is called the randomization set.

The Neyman-Pearson lemma gives an explicit construction for optimal tests for simple test problems:

Lemma C.76 (Neyman-Pearson). Let $\Theta_i := \{\theta_i\}$ for $i = 1, 2$ and $0 < \alpha < 1$. Then

(i) There exists a likelihood ratio test φ^* with $\gamma(x) = \gamma \in [0, 1]$ and precise level α, that is $\mathbb{E}_{\theta_0}[\varphi] = \alpha$.

(ii) If φ^* is a likelihood ratio test with level α, that is, $\mathbb{E}_{\theta_0}[\varphi] = \alpha$, φ^* then it is a best test for level α.

(iii) If φ is a best test for level α, φ is a likelihood ratio test and $\mathbb{E}_{\theta_0}[\varphi] < \alpha$ implies $\mathbb{E}_{\theta_1}[\varphi] = 1$.

A generalization for composite hypotheses which consist of more than a single parameter value is given by the generalized Neyman-Pearson lemma, see Rüschendorf (2014, Definition 6.3.7, Theorem 6.3.8). Sometimes, the likelihood ratio test is also written as

$$\lambda(x) = \frac{\sup\limits_{\Theta_0} L(\theta; x)}{\sup\limits_{\Theta} L(\theta; x)} \tag{C.22}$$

where $\lambda : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and a likelihood ratio test (LRT) is then defined as any test that has a rejection region of the form

$$\{x \in \mathcal{X} : \lambda(x) \leq c\} \tag{C.23}$$

where $0 \leq c \leq 1$. The rationale behind the LRT is intuitive: If there are parameter points in the alternative hypothesis H_1, for which the observed sample x is a lot more likely than for any parameter in the null hypothesis H_0, reject H_0. It is now possible to choose c so that the restriction of level α is guaranteed. As the definition of a hypothesis test for level α does not control the type II error probability, it is natural to search for level α tests for which the probability to reject H_0 is higher if $\theta \in \Theta_1$ than if $\theta \in \Theta_0$. This leads to the definition of an unbiased test, compare Casella and Berger (2002, Definition 8.3.9) and Rüschendorf (2014, Definition 6.4.1):

Definition C.77 (Unbiased test). Let $\Phi_\alpha := \{\varphi \in \Phi : \mathbb{E}_\theta[\varphi] \leq \alpha \text{ for all } \theta \in \Theta_0\} = \Phi_\alpha(\Theta_0)$ the set of all tests for level α with Φ as given in Definition C.67. A hypothesis test $\varphi \in \Phi$ is called unbiased (for level α) if $\varphi \in \Phi_\alpha$ and

$$\mathbb{E}_\theta[\varphi] \geq \alpha \qquad \forall \theta \in \Theta_1 \tag{C.24}$$

It is natural to demand that a good test in a class of tests $\mathcal{C}$ (e.g. LRTs for a specified level α) has a small type II error probability. If another test in $\mathcal{C}$ has a smaller type II error probability, it would be a better contender. This leads to the definition of the uniformly most powerful (UMP) class $\mathcal{C}$ test (Casella and Berger, 2002, Definition 8.3.11), (Rüschendorf, 2014, Definition 6.1.1).

Definition C.78 (Uniformly most powerful test for level α). The hypothesis test $\varphi^* \in \Phi_\alpha$ is called uniformly most powerful (UMP) test for level α, if for all $\theta \in \Theta_1$

$$\mathbb{E}_\theta[\varphi^*] = \sup_{\varphi \in \Phi_\alpha} \mathbb{E}_\theta[\varphi]$$

One drawback of the Neyman-Pearson lemma is the fact, that only tests involving simple hypotheses can be shown to be UMP level α tests. In most realistic applications and especially exploratory research, however, composite hypotheses are used which contain more than just a single parameter value:

Definition C.79 (One-sided hypothesis). Hypotheses $H : \theta \geq \theta_0$ or $H : \theta > \theta_0$, or $H : \theta \leq \theta_0$ or $H : \theta < \theta_0$, for a prespecified $\theta \in \Theta$ are called one-sided hypotheses.

Definition C.80 (Two-sided hypothesis). Hypotheses $H : \theta \neq \theta_0$ for a prespecified $\theta \in \Theta$ are called two-sided hypotheses.

In realistic applications, a large class of tests to consider consists of one-sided hypotheses and probability distributions or density functions with the monotone likelihood ratio property, see Casella and Berger (2002, p. 391) or Rüschendorf (2014, Definition 6.2.2).

Definition C.81 (Monotone likelihood ratio property). Let $(\Theta, \leq)$ be totally ordered. $\mathcal{P}$ has a (strictly) monotone density quotient in $T : (\mathcal{X}, \mathcal{A}) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$, if for all $\theta, \theta^{'} \in \Theta, \theta \leq \theta^{'}$ there exists a (strictly) isotone function $f_{\theta,\theta'} : (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}})) \to (\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$ so that

$$L_{\theta,\theta'} := \frac{f_{\theta'}}{f_\theta} = f_{\theta,\theta'} \circ T \qquad P_\theta + P_{\theta'} \text{ almost surely}$$

Phrased differently, the likelihood ratio (the density quotient) is a (strictly) monotone function of $x \in \mathcal{X}$ for every $\theta^{'} \geq \theta$ (where $\geq$ becomes $>$ in the case of a strictly monotone density quotient). As many families fulfill the monotone likelihood ratio property (e.g. $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, μ unknown, any univariate exponential family with certain conditions, see for example Rüschendorf (2014, Remark 6.2.3)), the Karlin-Rubin-Theorem answers the question to a UMP level α test for one-sided hypotheses, compare Casella and Berger (2002, Theorem 8.3.17) and Rüschendorf (2014, Theorem 6.2.6).

Theorem C.82 (Karlin-Rubin). Let Θ a totally ordered and identifiable parameterization (that is, $\varphi \to P_\varphi$ is injective). Suppose $\mathcal{P}$ has a monotone density quotient in T and let $\alpha \in [0,1]$ and for $\theta_0 \in \Theta, \Theta_0 := \{\theta \in \Theta : \theta \leq \theta_0\}, \Theta_1 := \{\theta \in \Theta : \theta > \theta_0\} \neq \emptyset$. Then there exists a uniformly most powerful test $\tilde{\varphi}$ for level α for (Θ_0, Θ_1) of the form

$$\tilde{\varphi}(x) := \tilde{\varphi}_{c,\gamma}(x) := \begin{cases} 1, \text{ if } T(x) > c \\ \gamma, \text{ if } T(x) = c \text{ with } c \in \overline{\mathbb{R}}, \gamma \in [0,1] \\ 0, \text{ if } T(x) < c \end{cases}$$

The Karlin-Rubin-Theorem applies also to situations where $H_0 : \theta \geq \theta_0$, and the test rejecting $H_0 : \theta \geq \theta_0$ in favour of $H_1 : \theta < \theta_0$ if and only if $T < c$ is a UMP level $\alpha = P_{\theta_0}(T < c)$ test.

C.7 p-values

After a hypothesis test is calculated, one way to report the conclusions drawn is to report the chosen size α and the decision made to reject or accept H_0. The usual heuristic based on the theory developed in part I is that the rejection of H_0 if α is small is fairly convincing, while if α is large, it certainly is not. The p-value quantifies the strength of this argument in Fisher's sense (Casella and Berger, 2002, Definition 8.3.26):

Definition C.83 (P-VALUE)**.** A p-value $p : (\mathcal{X}, \mathcal{A}) \to ([0,1], \mathcal{B}|_{[0,1]})$ is a (test) statistic satisfying $0 \leq p(x) \leq 1$ for every sample point $x \in \mathcal{X}$. A p-value is valid, if for all $\theta \in \Theta_0$ and for all $0 \leq \alpha \leq 1$,

$$P_\theta(p(x) \leq \alpha) \leq \alpha \tag{C.25}$$

Small values are often interpreted evidence that H_0 is not true, or as evidence for H_1 (notice that the latter is a much stronger statement than the former). The left hand side of Equation (C.25) can be interpreted as the probability to observe a p-value $p(x) \leq \alpha$ under assumption of the null hypothesis H_0. The right hand side of Equation (C.25) then restricts a valid p-value to have such a probability of less than or equal to α. If a valid p-value is found, the construction of a level α test is then possible based on $p(x)$: A test φ rejecting H_0 if and only if $p(x) \leq \alpha$ is a level α test due to Equation (C.25), because

$$\mathbb{E}_\theta[\varphi] = \int \varphi dP_\theta = \int \mathbb{1}_{\{p(x)\leq\alpha\}} dP_\theta = \int_{\{p(x)\leq\alpha\}} 1 dP_\theta = P_\theta(p(x) \leq \alpha) \leq \alpha \qquad \forall\theta \in \Theta_0$$

where in the last equality it was used that p is a valid p-value, compare Definition C.71. Then, instead of reporting the size α of the test and the decision to accept and reject H_0, one simply reports a test result via a p-value $p(x)$ which quantifies the evidence against the null hypothesis H_0 continuously. Importantly, the p-value resolves the decision-theoretic dichotomy of 'Reject H_0' and 'Accept H_0':

> "(...) a p-value reports the results of a test on a more continuous scale, rather than just the dichotomous decision "Accept H_0" or "Reject H_0"."
> Casella and Berger (2002, p. 397)

Clearly this is Fisher's interpretation and not Neyman's and Pearson's, compare Part I. A widespread method to construct a valid p-value is given by Theorem C.84 (Casella and Berger, 2002, Theorem 8.3.27):

Theorem C.84. Let $d(X)$ be a (test) statistic such that large values of d give evidence that H_1 is true. For each sample point, define

$$p(x) = \sup_{\theta\in\Theta_0} P_\theta(d(X) \geq d(x)) \tag{C.26}$$

Then, $p(x)$ is a valid p-value.

The right-hand side of C.26 can be interpreted as the probability of obtaining a test statistic equal to or more extreme than $d(x)$, maximized over all $\theta \in \Theta_0$. In a compact way, Held and Sabanés Bové (2014) summarise this as follows:

> "The p-value is the probability, under the assumption of the null hypothesis H_0, to obtain a result equal to or more extreme than what was actually observed."
> Held and Sabanés Bové (2014, p. 70)

C.8 Foundations of Interval Estimation

In realistic applications, also interval estimators are necessary to give not only a single value to estimate the parameter θ of interest, but instead provide a possible set of plausible values. The goal of constructing confidence sets is to find a set $C(x)$ based on the observation $x \in \mathcal{X}$, which – if θ is supposed to be the true parameter – includes a parameter function value $g(\theta)$ (which in most cases defaults to the identity $g(\theta) := \theta$) of θ with high probability $\geq 1-\alpha$:

$$P_\theta(\{x \in \mathcal{X} : g(\theta) \in C(x)\}) \geq 1-\alpha$$

The confidence set $C(x)$ thus should cover $g(\theta)$ with high probability. The goal of interval estimation via confidence sets is to find an interval which includes the parameter $g(\theta)$ with high probability.

Definition C.85 (Confidence set). Let $g : \Theta \to \Gamma$ (typically $\Gamma \subset \mathbb{R}^k$) a parameter function to be estimated (typically the identity $g(\theta) := \theta$). A function $C : \mathcal{X} \to \mathscr{P}(\Gamma)$ is called *confidence set* for g, if $A(\gamma') := \{x \in \mathcal{X} : \gamma' \in C(x)\} \in \mathcal{A}$ for all $\gamma' \in \Gamma$.

In the above, $\mathscr{P}(\Gamma)$ is the power set of Γ. Definition C.86 essentially states that the function C needs to be $\mathcal{A}$-measurable. Let $\mathcal{E}$ the set of all confidence sets for g.

Definition C.86 (Confidence set for g with confidence level $1-\alpha$). Let $\alpha \in [0,1]$ and C a confidence set. C is called confidence set for g with confidence level $1-\alpha$, if for all $\theta \in \Theta$:

$$P_\theta(\{x \in \mathcal{X} : g(\theta) \in C(x)\}) \geq 1-\alpha$$

The set of all confidence sets for g with confidence level $1-\alpha$ is henceforth denoted as $\mathcal{E}_{1-\alpha}$.

Example C.87. To illustrate the definition, let $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ with $\Theta := \mathbb{R}$, $P_\theta := \mathcal{N}(\theta, \sigma_0^2)^{(n)}$ and $g(\theta) := \theta$. Let $C_2(x) := [\bar{x}_n - \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2}, \bar{x}_n + \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2}]$ for $x \in \mathbb{R}$ a two-sided confidence set, also called confidence interval, where $u_{\alpha/2} := \Phi^{-1}(1-\frac{\alpha}{2})$ is the α-fractile of the standard normal distribution. For all $\theta \in \mathbb{R}$, it follows that

$$\begin{aligned}
P_\theta(\{x \in \mathcal{X} : g(\theta) \in C_2(x)\}) &= P_\theta(\{x \in \mathbb{R}^n : \theta \in C_2(x)\}) \\
&= P_\theta\left(\left\{\bar{x}_n - \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2} \leq \theta \leq \bar{x}_n + \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2}\right\}\right) \\
&= P_\theta\left(\left\{-u_{\alpha/2} \leq \sqrt{n}\frac{\bar{x}_n - \theta}{\sigma_0} \leq u_{\alpha/2}\right\}\right) \\
&= \Phi(u_{\alpha/2}) - \Phi(-u_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha
\end{aligned}$$

As a consequence, $C_2(x) \in \mathcal{E}_{1-\alpha}$, and wrong parameter values $\theta' \neq \theta$ are covered with a probability $< \alpha$.

Importantly, the random quantity in frequentist confidence set estimation is the confidence set (or interval) $C(x)$ and *not* the parameter $\theta \in \Theta$. As can easily be seen from the definition, the parameter θ is fixed and assumed to take a prespecified value. The probability statement

$$P_\theta(\{x \in \mathcal{X} : g(\theta) \in C(x)\}) \tag{C.27}$$

therefore refers to the random variable X with realisation $x \in \mathcal{X}$, and not to $\theta \in \Theta$. As a consequence, the coverage probability condition $P_\theta(\{x \in \mathcal{X} : g(\theta) \in C(x)\}) \geq 1-\alpha$ requires that the resulting confidence interval $C(x)$ for realisation $x \in \mathcal{X}$ covers the true and fixed parameter θ with probability $\geq 1-\alpha$, for any prespecified and fixed parameter value $\theta \in \Theta$. In context, this is often interpreted as follows: As randomness is concerned with $C(x)$ and the random variable X, under indefinite repetition of the experiment or study at least $1-\alpha$ percent of the confidence sets $C(x_1), C(x_2), C(x_3), \ldots$ need to include the true parameter value θ, for any $\theta \in \Theta$.

The concept of a *pivot* is helpful when searching for confidence intervals (Casella and Berger, 2002, Section 9.2.2), and is a generalisation of the confidence set (Rüschendorf, 2014, Definition 7.1.3)

Definition C.88 (Pivot). Let $g : \Theta \to \Gamma$. A measurable function $T : \mathcal{X} \times \Gamma \to \Gamma$ is called pivot (for g), if:

1. $Q = P_\theta^{T(\cdot, g(\theta))}$ does not depend on $\theta \in \Theta$.

2. For $B \in \mathcal{A}_\Gamma$ and $\theta \in \Theta$ the set
$$\{x \in \mathcal{X} : T(x, g(\theta)) \in B\}$$
is $\in \mathcal{A}$.

$C_B(x) := \{\gamma \in \Gamma : T(x, \gamma) \in B\}$ is called the confidence set which is induced through B and T. For $\theta \in \Theta$, the following equality holds:
$$\{g(\theta) \in C_B(\cdot)\} = \{T(\cdot, g(\theta)) \in B\} \in \mathcal{A}$$

A pivot generalises the concept of ancillary statistics. This generalisation allows for the construction of confidence sets C_B, where the choice of B determines the geometric form and the confidence level of C_B. The following example demonstrates how useful pivots are in practice.

Example C.89. Let $\Theta = \mathbb{R} \times \mathbb{R}_+$ and $P_\theta = \mathcal{N}(\mu, \sigma^2)^{(n)}$ for $\theta = (\mu, \sigma^2)$. Interest lies in constructing a pivot for μ, so let the parameter function $g(\theta) := \mu$, that is, the projection of θ onto the first coordinate μ (as a consequence, $\Gamma := \mathbb{R}$, so $g : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$). Let
$$T(x, \mu) := \sqrt{n}\frac{\bar{x}_n - \mu}{s_n}$$
with
$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

Then, $T(x, \mu)$ is a pivot with distribution $P_\theta^{T(\cdot, \mu)}$ that is given as t_{n-1} (Casella and Berger, 2002), which is the t-distribution with $n-1$ degrees of freedom. Clearly, $T : \mathcal{X} \times \Gamma \to \Gamma$, and the t_{n-1} distribution does not depend on $\theta = \mu \in \Theta$. For $B \in \mathcal{A}_\Gamma$ (where $A_\Gamma := \mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$) and $\theta \in \Theta$ the set $\{x \in \mathcal{X} : T(x, g(\theta)) \in B\}$ is $\in \mathcal{A}$ (where $\mathcal{A} := \mathcal{B}(\mathbb{R}^n)$ is the Borel σ-algebra on $\mathbb{R}^n$). Consequently, $T(x, \mu)$ fufills conditions 1. and 2. above and is a pivot for $g(\theta) = \mu$. To construct a confidence set, this pivot can be employed. Let $C_2(x) = [\bar{x}_n - \frac{s_n}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}, \bar{x}_n + \frac{s_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}]$, then

$$P_\theta(\{\mu \in C_2\}) = P_\theta(\{T(\cdot,\mu) \in [t_{n-1,\frac{\alpha}{2}}, t_{n-1,1-\frac{\alpha}{2}}]\}) = 1 - \alpha$$

because $P_\theta^{T(\cdot,\mu)} = t_{n-1}$, that is, $T(\cdot,\mu)$ is distributed as t_{n-1}. As a consequence, C_2 is a two-sided confidence interval for μ with confidence level $1-\alpha$.

It is also possible to construct confidence sets with minimal volume based on a differentiation condition via pivots, see Rüschendorf (2014, Theorem 7.1.12) for measure-theoretic details and Casella and Berger (2002, Theorem 9.3.2) for a more applied treatment.

An important systematic relationship is given between frequentist confidence sets and hypothesis tests, compare Rüschendorf (2014, Section 7.2). Again, a parameter function $g : \Theta \to \Gamma$ is specified and a confidence set $C : \mathcal{X} \to \mathcal{P}(\Gamma)$ is of interest. To state that $C(x)$ should not contain "wrong values" γ' when the true parameter is θ, for example $\gamma' < g(\theta)$ for one-sided confidence intervals or $\gamma' \in [g(\theta) - \varepsilon, g(\theta) + \varepsilon]^c$ for two-sided confidence intervals, define for $\theta \in \Theta$ subsets $\tilde{H}_{0,\theta} \subset \Gamma$ which are interpreted as "correct values" γ' and $\tilde{H}_{1,\theta} \subset \Gamma$ which are interpreted as "wrong values" γ', where $\tilde{H}_{0,\theta} \cap \tilde{H}_{1,\theta} = \emptyset$. The system $(\tilde{H}_{0\theta}, \tilde{H}_{1\theta}), \theta \in \Theta$ is called *form hypotheses* for g. Using this notation, the definition of a confidence set for g with confidence level $1-\alpha$ can be extended as follows (Rüschendorf, 2014, Definition 7.2.1):

Definition C.90 (UNBIASED AND UNIFORMLY MOST POWERFUL CONFIDENCE SETS WITH CONFIDENCE LEVEL $1-\alpha$). Let $(\tilde{H}_{0,\theta}, \tilde{H}_{1,\theta})$, given form hypotheses for g and $\theta \in \Theta$.

(a) A confidence set C is called confidence set for g with confidence level $1-\alpha \Leftrightarrow P_\theta(\{\gamma' \in C\}) \geq 1-\alpha, \forall \gamma' \in \tilde{H}_{0,\theta}, \forall \theta \in \Theta$. As previously, the set $\mathcal{E}_{1-\alpha} := \mathcal{E}_{1-\alpha}((\tilde{H}_{0,\theta}, \tilde{H}_{1,\theta})_{\theta \in \Theta})$ denotes the set of all confidence sets for level $1-\alpha$.

(b) $C^* \in \mathcal{E}_{1-\alpha}$ is called uniformly most powerful confidence set with confidence level $1-\alpha \Leftrightarrow P_\theta(\{\gamma' \in C^*\}) = \inf_{C \in \mathcal{E}_{1-\alpha}} \{P_\theta(\{\gamma' \in C^*\}) : C \in \mathcal{E}_{1-\alpha}\}, \forall \gamma' \in \tilde{H}_{1,\theta}, \forall \theta \in \Theta$.

(c) $C \in \mathcal{E}_{1-\alpha}$ is called unbiased confidence set with confidence level $1-\alpha$ for $g \Leftrightarrow P_\theta(\{\gamma' \in C\}) \leq 1-\alpha, \forall \gamma' \in \tilde{H}_{1,\theta}, \forall \theta \in \Theta$. $\mathcal{E}_{1-\alpha,u}$ denotes the set of unbiased confidence sets with confidence level $1-\alpha$ for g.

(d) $C^* \in \mathcal{E}_{1-\alpha,u}$ is called uniformly most powerful unbiased confidence set with confidence level $1-\alpha$ for $g \Leftrightarrow P_\theta(\{\gamma' \in C^*\}) = \inf_{C \in \mathcal{E}_{1-\alpha,u}} P_\theta(\{\gamma' \in C\}), \forall \gamma' \in \tilde{H}_{1,\theta}, \forall \theta \in \Theta$.

If for $g(\theta) = \theta$ the form hypotheses are selected as $\tilde{H}_{0,\theta} := \{\theta\}$ and $\tilde{H}_{1,\theta} := \Theta \setminus \{\theta\}, \theta \in \Theta$, then all parameters $\gamma \in \Theta$ with $\gamma \neq \theta$ are interpreted as wrong, and a uniformly most powerful confidence set C^* should cover as few wrong parameters $\gamma \neq \theta$ as possible. This means that C^* should be a set with smallest possible volume. If ϱ is a volume measure on $(\Theta, \mathcal{A}_\Theta)$ with $\varrho(\{\theta\}) = 0$ for all $\theta \in \Theta$ (e.g. the Lebesgue-measure λ), and it is assumed that $\{\theta\} \in \mathcal{A}_\Theta$ for all $\theta \in \Theta$, then

$$\mathrm{vol}_\varrho(C) := \int \varrho(C(x)) dP_\theta(x)$$

is the mean volume of $C(x)$ and

$$\beta_{\varrho}(C) := \int_{\Theta \setminus \{\theta\}} P_{\theta}(\{\gamma' \in C\}) d\varrho(\gamma')$$

is the mean weighted coverage probability of wrong parameter values $\gamma' \neq \theta$, where the weights are from ϱ. The theorem of Pratt states that these two quantities are equal:

Theorem C.91 (Pratt (1961)). Let C a confidence set for $g(\theta) = \theta$ with $C(x) \in \mathcal{A}_{\Theta}, \forall x \in \mathcal{X}$. For the form hypotheses $\tilde{H}_{0\theta} = \{\theta\}$, $\tilde{H}_{1\theta} = \Theta \setminus \{\theta\}$ and the volume measure ϱ as defined above, for $\theta \in \Theta$ the following equality holds:

$$\text{vol}_{\varrho}(C) = \beta_{\varrho}(C)$$

For a proof, see Rüschendorf (2014, p. 242). As a consequence of Pratt's theorem, a uniformly most powerful confidence set for $g(\theta) = \theta$ with confidence level $1 - \alpha$ also minimises the mean volume of $C(x)$. Thus, UMP confidence sets in the Neyman-Pearson theory are therefore sometimes also called "Neyman-shortest" (Casella and George, 1992).

Now, an important dualism between hypothesis tests and confidence intervals is given as follows: C is a uniformly most powerful (unbiased) confidence set with confidence level $1 - \alpha$, if and only if for $A_C(\gamma') := \{\gamma' \in C\}$, the indicator function $\mathbb{1}_{(A(\gamma'))^c}$ is a uniformly most powerful (unbiased) non-randomized test for level α for $(H_{0,\gamma'}, H_{1,\gamma'})$, for all $\gamma' \in \Gamma$. This fact is established via the correspondence theorem, compare Rüschendorf (2014, Theorem 7.2.3):

Theorem C.92 (Correpondence Theorem). $C^* \in \mathcal{E}_{1-\alpha}$ is a uniformly most powerful (unbiased) confidence set with confidence level $1 - \alpha$ for $g \Leftrightarrow \forall \gamma' \in \Gamma$ $\mathbb{1}_{(A_{C^*}(\gamma'))^c}$ is a uniformly most powerful (unbiased) non-randomized test for level α for the test problem $(H_{0,\gamma'}, H_{1,\gamma'})$.

Phrased differently, the indicator function on the complement of the uniformly most powerful (unbiased) confidence set C^* (the test rejects $H_{0,\gamma'}$ when $x \notin A_C(\gamma')$) with confidence level $1 - \alpha$ for g is a uniformly most powerful (unbiased) non-randomized test for level α for $H_{0,\gamma'}$ against $H_{1,\gamma'}$ ($H_{0,\gamma'} := H_0 : \theta = \gamma' \in \Theta$ and $H_{1,\gamma'} := H_1 : \theta \neq \gamma' \in \Theta$). For an example for the dualism between a uniformly most powerful (unbiased) confidence set and a uniformly most powerful (unbiased) non-randomized test in the normal distribution model, see Rüschendorf (2014, Example 7.2.4).

Bibliography

Achinstein, P. (2001). *The Book of Evidence*. Oxford University Press, Oxford.

Achinstein, P. (2010). Mill's Sins or Mayo's Errors? In Mayo, D. and Spanos, A., editors, *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, pages 170–188. Cambridge University Press, Cambridge.

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176.

Aldrich, J. (2006). The Statistical Education of Harold Jeffreys. *International Statistical Review*, 73(3):289–307.

Aldrich, J. (2008). R. A. Fisher on Bayes and Bayes' Theorem. *Bayesian Analysis*, 3(1):161–170.

Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine*, 1(1):59–71.

Altman, D. G. (1991a). *Practical statistics for medical research*. Chapman and Hall, Boca Raton.

Altman, D. G. (1991b). Statistics in medical journals: Developments in the 1980s. *Statistics in Medicine*, 10(12):1897–1913.

Altman, D. G. (2000). Statistics in medical journals: Some recent trends. *Statistics in Medicine*, 19(23):3275–3289.

Altman, D. G., Gore, S. M., Gardner, M. J., and Pocock, S. J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal (Clinical research ed.)*, 286(6376):1489–93.

Anderson, H. L. (1986). Metropolis, Monte Carlo, and the MANIAC. *Los Alamos Science*, 14:96–108.

Andrieu, C., Moulines, É., and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference, CDC-ECC '05*, 2005(3):6656–6661.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Anscombe, F. (1963). Sequential Medical Trials. *Journal of the American Statistical Association*, 58:365–383.

Azevedo-Filho, A. and Shachter, R. D. (1994). Laplace's Method Approximations for Probabilistic Inference in Belief Networks with Continuous Variables. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 28–36. Elsevier.

Bai, Y., Craiu, R. V., and Di Narzo, A. F. (2011). Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC. *Journal of Computational and Graphical Statistics*, 20(1):63–79.

Baker, M. and Dolgin, E. (2017). Reproducibility project yields muddy results: Gates Foundation demands open access. *Nature*, 541:269–270.

Baker, M. and Penny, D. (2016). Is there a reproducibility crisis? *Nature*, 533(7604):452–454.

Barker, A. A. (1969). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119–133.

Barnard, G., Jenkins, G., and Winsten, C. (1962). Likelihood Inference and Time Series. *Journal of the Royal Statistical Society. Series A (General)*, 125(3):321–372.

Barnard G.A. (1947). A review of 'Sequential Analysis' by Abraham Wald. *Journal of the American*

Statistical Association, 42:658–669.

Barnard G.A. (1949). Statistical inference (with Discussion). *Journal of the Royal Stastical Society Series B (Methodological)*, 11:115–139.

Bartholomew, D. (1967). Hypothesis Testing When the Sample Size is Treated as a Random. *Journal of the Royal Statistical Society. Series B (Methodological)*, 29(1):53–82.

Bartlett, M. S. (1933). Probability and Chance in the Theory of Statistics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 141(845):518–534.

Basu, D. (1975). Statistical Information and Likelihood (with discussion). *Sankhya: The Indian Journal of Statistics, Series A*, 37(1):1–71.

Bauer, H. (2001). *Measure and integration theory*. De Gruyter, Berlin.

Bayes, M. and Price, M. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53(0):370–418.

Begley, C. G. (2013). Six red flags for suspect work. *Nature*, 497(7450):433–434.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, 116(1):116–126.

Benjamin, D. J. and Berger, J. (2019). Three Recommendations for Improving the Use of p-Values. *The American Statistician*, 73(sup1):186–191.

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A. P., Forster, M., George, E. I., Gonzalez, R., Goodman, S., Green, E., Green, D. P., Greenwald, A. G., Hadfield, J. D., Hedges, L. V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D. J., Imai, K., Imbens, G., Ioannidis, J. P. A., Jeon, M., Jones, J. H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S. E., McCarthy, M., Moore, D. A., Morgan, S. L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T. H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F. D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D. J., Winship, C., Wolpert, R. L., Xie, Y., Young, C., Zinman, J., and Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10.

Bennett, J. (1990). *Statistical Inference and Analysis: Selected Correspondance of R.A.Fisher*. Clarendon Press, Oxford.

Berger, J. (1980). *Statistical Decision Theory : Foundations, Concepts, and Methods*. Springer New York.

Berger, J. (1984). In defense of the likelihood principle: axiomatics and coherency. In Bernardo, J., DeGroot, M., Lindley, D., and Smith, A., editors, *Bayesian Statistics II - Proceedings of the Second Valencia International Meeting*, pages 33–67, Valencia. Elsevier.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

Berger, J., Brown, L., and Wolpert, R. (1994). A Unified Conditional Frequentist and Bayesian Test for fixed and sequential Hypothesis Testing. *The Annals of Statistics*, 22(4):1787–1807.

Berger, J. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3):317–335.

Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, 82(397):112–122.

Berger, J. and Wolpert, R. L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, California.

Bergh, D. V. D., Doorn, J. V., Marsman, M., Gupta, K. N., Sarafoglou, A., Jan, G., Stefan, A., Ly, A., and Hinne, M. (2019). A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP. *psyarxiv preprint, https://psyarxiv.com/spreb*.

Beribisky, N., Davidson, H., and Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, 2019(5):e6853.

Bernado, J. (1999). Nested hypothesis testing: the Bayesian reference criterion. In Bernado, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics (Vol. 6)*, pages 101–130 (with discussion). Oxford University Press, Valencia.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *1986*, 48(3):259–302.

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv preprint, https://arxiv.org/abs/1701.02434*.

Birnbaum, A. (1962). On the Foundations of Statistical Inference (with discussion). *Journal of the American Statistical Association*, 57(298):269–306.

Birnbaum, A. (1972). More on concepts of statistical evidence. *Journal of the American Statistical Association*, 67(340):858–861.

Borges, W. and Stern, J. M. (2007). The Rules of Logic Composition for the Bayesian Epistemic e-Values. *Logic Journal of the IGPL*, 15(5-6):401–420.

Box, G. E. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356):791.

Box, J. F. (1978). *R.A. Fisher. The Life of a Scientist*. Wiley, New York.

Brard, C., Le Teuff, G., Le Deley, M.-C., and Hampson, L. V. (2017). Bayesian survival analysis in clinical trials: What methods are used in practice? *Clinical Trials: Journal of the Society for Clinical Trials*, 14(1):78–87.

Brémaud, P. (2020). *Probability Theory and Stochastic Processes*. Springer Nature Switzerland, Cham, Switzerland.

Broad, C. (1918). On the Relation between Induction and Probability (Part I). *Mind*, 27(108):389–404.

Brooks, S. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press,Taylor & Francis, Boca Raton.

Brooks, S. P., Dellaportas, P., and Roberts, G. O. (1997). An Approach to Diagnosing Total Variation Convergence of MCMC Algorithms. *Journal of Computational and Graphical Statistics*, 6(3):251–265.

Brooks, S. P. and Roberts, G. O. (1998). Assessing Convergence of Markov Chain Monte Carlo Algorithms. *Statistics and Computing*, (8):319–335.

Brownstein, N. C., Louis, T. A., O'Hagan, A., and Pendergast, J. (2019). The Role of Expert Judgment in Statistical Inference and Evidence-Based Decision-Making. *The American Statistician*, 73(sup1):56–68.

Buchanan-Wollaston, H. J. (1935). Statistical tests. *Nature*, 136(3444):722.

Buehler, R. J. (1959). Some Validity Criteria for Statistical Inferences. *The Annals of Mathematical Statistics*, 30(4):845–863.

Buehler, R. J. and Feddersen, A. P. (1963). Note on a conditional property of student's t. *The Annals of Mathematical Statistics*, 34(3):1098–1100.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Carlin, B. and Louis, T. (2009). *Bayesian Methods for Data Analysis*. Chapman & Hall, CRC Press, Boca Raton.

Carnap, R. (1950). *Logical foundations of probability*. University of Chicago Press, Chicago.

Carnap, R. (1962). *Logical Foundations of Probability*. University of Chicago Press, Chicago, 2nd edition.

Carpenter, B., Guo, J., Hoffman, M. D., Brubaker, M., Gelman, A., Lee, D., Goodrich, B., Li, P., Riddell, A., and Betancourt, M. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Thomson Learning, Stamford, Connecticut.

Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., and Stanley, D. J. (2019). Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological Science*, 2(3):233–239.

Center for Open Science (2020). OSF Open Science Foundation. *https://osf.io/*.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Clifford, P. (1990). Markov random fields in statistics. In Grimmett, G. R. and Welsh, D. J. A., editors, *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, pages 19–32. Oxford University Press, Oxford.

Clyde, M. (2020). BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling. *R software package*, (version 1.5.5).

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3):145–153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, Hillsdale, N.J, 2nd edition.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12):1304–1312.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3):140216–140216.

Colquhoun, D. (2017). The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12).

Cook, J., Hislop, J., Adewuyi, T., Harrild, K., Altman, D., Ramsay, D., Fraser, C., Buckley, B., Fayers, P., Harvey, I., Briggs, A., Norrie, J., Fergusson, D., Ford, I., and Vale, L. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technology Assessment*, 18(28):1–172.

Cook, J., Julious, S., Sones, W., Hampson, L., Hewitt, C., Berlin, J., Ashby, D., Emsley, R., Fergusson, D., Walters, S., Wilson, E., MacLennan, G., Stallard, N., Rothwell, J., Bland, M., Brown, L., Ramsay, C., Cook, A., Armstrong, D., Altman, D., and Vale, L. (2018). DELTA 2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Trials*, 19(1):1–6.

Coro, G. (2017). A Lightweight Guide on Gibbs Sampling and JAGS. Technical report, Institute of Science and Technology, University of Pisa, Pisa.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434):883–904.

Cox, D. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Craiu, R. and Rosenthal, J. S. (2014). Bayesian Computation Via Markov Chain Monte Carlo. *Annual Review of Statistics and Its Applications*, 1(1):179–201.

Davis, B. J. and Erlanson, D. A. (2013). Learning from our mistakes: The 'unknown knowns' in fragment screening. *Bioorganic and Medicinal Chemistry Letters*, 23(10):2844–2852.

Dawid, A. (1977). Recent Developments in Statistics. In *Proceedings of the European Meeting of Statisticians*, Grenoble. North-Holland Pub. Co.

de Finetti, B. (2017). *Theory of Probability*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, Chichester, UK.

Diaconis, P. (2009). The Markov Chain Monte Carlo Revolution. *Bulletin of the American Mathematical Society*, 46(208):179–205.

Dickey, J. M. and Lientz, B. P. (1970). The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *Annals of Mathematical Statistics*, 41(1):214–226.

Diebolt, J. and Robert, C. P. . (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society Series B (Methodological)*, 56(2):363–375.

Dobruschin, P. L. (1968). The Description of a Random Field by Means of Conditional Probabilities and Conditions of Its Regularity. *Theory of Probability & Its Applications*, 13(2):197–224.

Dongarra, J. and Sullivan, F. (2000). Guest Editors Introduction to the top 10 algorithms. *Computing in Science & Engineering*, 2(1):22–23.

Doob, J. (1949). Le Calcul des Probabilités et ses Applications. *Colloques Internationaux Du Centre National de La Recherche Scientifique, No. 13. Centre National de la Recherche Scientifique, Paris*, 13:23–27.

Dronamraju, K. (2012). Recollections of J.B.S. Haldane, with special reference to Human Genetics in India. *Indian Journal of Human Genetics*, 18(1):3–8.

Dronamraju, K. (2015). J.B.S. Haldane as I knew him, with a brief account of his contribution to mutation research. *Mutation Research*, 765:1–6.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Eckhardt, R. (1987). Stan Ulam, John Von Neumann, and the Monte Carlo Method. *Los Alamos Science*, 15:131–136.

Edwards, A. L. (1950). *Experimental design in psychological research*. Rinehart, New York.

Edwards, A. L. (1954). *Statistical methods for the behavioral sciences*. Rinehart, New York.

Edwards, J. H. (1993). Haldane and the Analysis of Linkage. In Majumder, P. P., editor, *Human Population Genetics: A Centennial Tribute to J.B.S. Haldane*, pages 153–164. Springer, Boston, MA.

Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3):193–242.

Efron, B. (1998). R. A. Fisher in the 21st century (Invited paper presented at the 1996 R. A. Fisher Lecture). *Statistical Science*, 13:95–122.

Efron, B. and Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press, Cambridge.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, 11(2):e0149794.

Etz, A. and Wagenmakers, E.-J. (2017). J. B. S. Haldane's Contribution to the Bayes Factor Hypothesis Test. *Statistical Science*, 32(2):313–329.

Faraway, J. J. (2016). *Extending the linear model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC, New York, 2nd edition.

Faulkenberry, T. J., Ly, A., and Wagenmakers, E. J. (2020). Bayesian Inference in Numerical Cognition: A Tutorial Using JASP. *Journal of Numerical Cognition*, 6(2):231–259.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5):532–538.

Fiedler, K. and Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1):45–52.

Fisher, R. (1921a). Studies in Crop Variation. I. An examination of the yield of dressed grain from broadbalk. *The Journal of Agricultural Science*, XI(6):107–135.

Fisher, R. (1925a). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fisher, R. (1935). The Logic of Inductive Inference. *Journal of the Royal Statistical Society*, 98(1):39–82.

Fisher, R. (1936). Uncertain Inference. *Proceedings of the American Academy of Arts and Sciences*, 71(4):245–258.

Fisher, R. (1948). Conclusions fiduciaires. *Annales de l'I.H.P.*, 3:191–213.

Fisher, R. (1950). *Contributions to Mathematical Statistics*. John Wiley & Sons, New York.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Stastical Society Series B (Methodological)*, 17(1):69–78.

Fisher, R. (1956a). On a test of significance in Pearson's Biometrika Tables (No. 11). *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 56–60.

Fisher, R. (1958a). *Statistical Methods for Research Workers*. Hafner, New York, 12th edition.

Fisher, R. A. (1912). On an Absolute Criterion for Fitting Frequency Curves. *Messenger of Mathematics*, 41:155–160.

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, (52):399–433.

Fisher, R. A. (1920). A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error. *Monthly Notices of the Royal Astronomical Society*, (80):758–770.

Fisher, R. A. (1921b). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, (1):3–32.

Fisher, R. A. (1922a). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society*, A(222):309–368.

Fisher, R. A. (1922b). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.

Fisher, R. A. (1922c). The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients. *Journal of the Royal Statistical Society*, (85):597–612.

Fisher, R. A. (1925b). *Statistical Methods for Research Workers*. Oliver and Boyd, Hafner Publishing Company, Edinburgh.

Fisher, R. A. (1925c). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(05):700.

Fisher, R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain*, (33):501–513.

Fisher, R. A. (1930). Inverse Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(4):528–535.

Fisher, R. A. (1932). Inverse probability and the use of Likelihood. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3):257–261.

Fisher, R. A. (1934a). Probability Likelihood and Quantity of Information in the Logic of Uncertain Inference. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 146(856):1–8.

Fisher, R. A. (1934b). *Statistical Methods for Research Workers*. Oliver and Boyd LTD., Edinburgh, 5th edition.

Fisher, R. A. (1934c). Two New Properties of Mathematical Likelihood. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 144(852):285–307.

Fisher, R. A. (1939). A Note on Fiducial Inference. *The Annals of Mathematical Statistics*, 10(4):383–388.

Fisher, R. A. (1956b). *Statistical Methods and Scientific Inference*. Oliver and Boyd, London.

Fisher, R. A. (1958b). The nature of probability. *Centennial Review*, (2):261–274.

Flinn, P. A. (1974). Monte Carlo calculation of phase separation in a two-dimensional Ising system. *Journal of Statistical Physics*, 10(1):89–97.

Fraser, D. (1963). On the Sufficiency and Likelihood Principles. *Journal of the American Statistical Association*, 58(303):641–647.

Fraser, D. (1969). The Structure of Inference. *Biometrika*, 56(2):453–456.

Fraser, D. (1972). Bayes, Likelihood of Structural. *The Annals of Mathematical Statistics*, 43(4):777–790.

Freedman, L. S., Lowe, D., and Macaskill, P. (1983). Stopping rules for clinical trials. *Statistics in Medicine*, 2(2):167–174.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Gandenberger, G. (2015). A New Proof of the Likelihood Principle. *The British Journal for the Philosophy of Science*, 66(3):475–503.

Gaver, D. and O'Muircheartaigh, I. (1987). Robust Empirical Bayes Analyses of Event Rates. *Technometrics*, 29(1):1–15.

Geisser, S. (1980). Basic Theory of the 1922 Mathematical Statistics Paper. In Fienberg, S. E. and Hinkley, D. V., editors, *R.A. Fisher - An Appreciation*, pages 59–66. Springer Verlag New York, New York.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian data analysis*. CRC Press/Taylor & Francis, Boca Raton, 3rd edition.

Gelman, A., Haig, B., Hennig, C., Owen, A., Cousins, R., Young, S., Robert, C., Yanofsky, C., Wagenmakers, E.-J., Kenett, R., and Lakeland, D. (2019). Many perspectives on Deborah Mayo's "Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars". *arXiv preprint, http://arxiv.org/abs/1905.08876*.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483.

Ghosal, S. (1996). A Review of Consistency and Convergence of Posterior Distribution. In *Proceedings of Varanashi Symposium in Bayesian Inference*. Banaras Hindu University.

Giere, R. N. (1977). Allan Birnbaum's conception of statistical evidence. *Synthese*, 36(1):5–13.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606.

Gigerenzer, G. and Marewski, J. N. (2015). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, 41(2):421–440.

Glauber, R. J. (1963). Time-Dependent Statistics of the Ising Model. *Journal of Mathematical Physics*, 4(2):294–307.

Glick, J. L. (1992). Scientific Data Audit—A Key Management Tool. *Accountability in Research*, 2(3):153–168.

Godambe, V. (1979). On Birnbaum's Axiom of the Mathematically Equivalent Experiments. *Journal of the Royal Statistical Society. Series B. (Methodological)*, 41(1):107–110.

Gönen, M., Johnson, W. O., Lu, Y., and Westfall, P. H. (2005). The Bayesian Two-Sample t Test. *The American Statistician*, 59(3):252–257.

Good, I. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

Good, I. (1960). Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(2):319–331.

Good, I. (1968). Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. Technical Report 2.

Good, I. (1981). Some Logic and History of Hypothesis Testing. In Pitt, J. C., editor, *Philosophy in Economics*, pages 149–174. Springer Netherlands, Virginia.

Good, I. (1985). Weight of Evidence: A brief survey. In Bernado, J., DeGroot, M. H., Lindley, D., and Smith, A., editors, *Bayesian Statistics (Vol. 2)*, pages 249–277, Valencia, Spain. Elsevier Science Publishers B.V. (North-Holland).

Good, I. (1988). The Interface between Statistics and Philosophy of Science. *Statistical Science*, 3(4):386–412.

Good, I. (1993). C397. Refutation and Rejection Versus Inexactification, and Other Comments Concerning Terminology. *Journal of Statistical Computation and Simulation*, 47(1-2):91–92.

Good, I. (1994). C420. The existence of sharp null hypotheses. *Journal of Statistical Computation and Simulation*, 49(3-4):241–242.

Goodman, S. N. (1999). Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130(12):1005.

Goodrich, B. K., Wawro, G., and Katznelson, I. (2012). Designing Quantitative Historical Social Inquiry: An Introduction to Stan.

Gordon, G. S. D., Joseph, J., Alcolea, M. P., Sawyer, T., Macfaden, A. J., Williams, C., Fitzpatrick, C. R. M., Jones, P. H., di Pietro, M., Fitzgerald, R. C., Wilkinson, T. D., and Bohndiek, S. E. (2018). Quantitative phase and polarisation endoscopy applied to detection of early oesophageal tumourigenesis. *arxiv preprint, https://arxiv.org/abs/1811.03977*.

Greenland, S. (2019). Valid p-Values Behave Exactly as They Should: Some Misleading Criticisms of p-Values and Their Resolution With s-Values. *The American Statistician*, 73(sup1):106–114.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350.

Gronau, Q. F., Ly, A., and Wagenmakers, E.-J. (2020). Informed Bayesian t -Tests. *The American Statistician*, 74(2):137–143.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97.

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., and Matzke, D. (2019). A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models Using Warp-III Bridge Sampling. *Psychometrika*, 84(1):261–284.

Grossman, J. (2011). The likelihood principle. In Bandyopadhyay, P. S. and Forster, M. R., editors, *Philosophy of Statistics*, chapter 7, pages 553–580. Elsevier North-Holland, Amsterdam.

Gubernatis, J. E. (2005). Marshall Rosenbluth and the Metropolis algorithm. *Physics of Plasmas*, 12(5):057303.

Haaf, J. M., Ly, A., and Wagenmakers, E. J. (2019). Retire significance, but still test hypotheses. *Nature*, 567(7749):461.

Haario, H., Saksman, E., and Tamminen, J. (2001). An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223.

Hacking, I. (1980). The Theory of Probable Inference: Neyman, Peirce and Braithwaite. In Mellor, D., editor, *Science, Belief and Behaviour: Essays in Honour of R. B. Brathwaite*, pages 141–160. Cambridge University Press, Cambridge.

Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(1):55–61.

Halpin, P. F. and Stam, H. J. (2006). Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940-1960). *The American Journal of Psychology*, 119(4):625–653.

Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15(5):20190174.

Hammersley, J. M. and Clifford, P. (1971). Markov Fields on Finite Graphs and Lattices. Technical report, University of California.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Springer Netherlands.

Harlow, F. H. and Metropolis, N. (1983). Computing & Computers: Weapons Simulation Leads to the Computer Era. *Los Alamos Science*, pages 132–141.

Hastie, T., Tibshirani, R., and Friedman, J. H. J. H. (2017). *The Elements of Statistical Learning : Data mining, Inference, and Prediction*. Springer-Verlag New York, New York, 2nd edition.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with Sparsity : the lasso and generalizations*. Chapman and Hall/CRC, New York, 1st edition.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109.

Healy, M. J. R. (2003). R. A. Fisher the statistician. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):303–310.

Held, L. and Ott, M. (2016). How the Maximal Evidence of p-Values Against Point Null Hypotheses Depends on Sample Size. *The American Statistician*, 70(4):335–341.

Held, L. and Ott, M. (2018). On p-Values and Bayes Factors. *Annual Review of Statistics and Its Application*, 5(1):393–419.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer, Berlin, Heidelberg.

Hendriksen, A., de Heide, R., and Grünwald, P. (2020). Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis*, in press.

Hitchcock, D. B. (2003). A History of the Metropolis-Hastings algorithm. *American Statistician*, 57(4):254–257.

Hobbs, B. P. and Carlin, B. P. (2007). Practical Bayesian design and analysis for drug and device clinical trials. *Journal of Biopharmaceutical Statistics*, 18(1):54–80.

Hodges, J. L. and Lehmann, E. L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):261–268.

Hoffman, M. D., Carpenter, B., and Gelman, A. (2012). Stan, scalable software for Bayesian modeling. In *Proceedings of the NIPS Workshop on Probabilistic Programming*.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381.

Howie, D. (2002). *Interpreting Probability : Controversies and Developments in the Early Twentieth Century*. Cambridge University Press, Cambridge.

Hubbard, R. (2004). Alphabet Soup. *Theory & Psychology*, 14(3):295–327.

Hubbard, R. (2019). Will the ASA's Efforts to Improve Statistical Practice be Successful? Some Evidence to the Contrary. *The American Statistician*, 73(sup1):31–35.

Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997). The Spread of Statistical Significance Testing in Psychology. *Theory & Psychology*, 7(4):545–554.

Hubbard, R. and Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology - And its future prospects. *Educational and Psychological Measurement*, 60(5):661–681.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, 61(4):317–333.

Hunter, J. E. (1997). Needed: A Ban on the Significance Test. *Psychological Science*, 8(1):3–7.

Hurlbert, S. H., Levine, R. A., and Utts, J. (2019). Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. *The American Statistician*, 73(sup1):352–357.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer New York, New York.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294(2):218–228.

Ioannidis, J. P. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psycho-*

logical Science, 7(6):645–654.

Ioannidis, J. P. (2016). Why Most Clinical Research Is Not Useful. *PLoS Medicine*, 13(6).

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701.

Ishwaran, H. (1999). Applications of Hybrid Monte Carlo to Bayesian Generalized Linear Models: Quasicomplete Separation and Neural Networks. *Journal of Computational and Graphical Statistics*, 8(4):779–799.

Jaeschke, R., Singer, J., and Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4):407–415.

JASP Team (2019). Jeffreys Awesome Statistics Package (JASP). *https://jasp-stats.org/*.

Jeffreys, H. (1931). *Scientific Inference*. Cambridge University Press, Cambridge.

Jeffreys, H. (1933). Probability, Statistics and the Theory of Errors. *Proceedings of the Royal Society A*, 140:523–535.

Jeffreys, H. (1934). Probability and Scientific Method. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 146(856):9–16.

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222.

Jeffreys, H. (1936). Further significance tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(03):416–445.

Jeffreys, H. (1939). *Theory of Probability*. The Clarendon Press, Oxford, 1st edition.

Jeffreys, H. (1948). *Theory of Probability*. The Clarendon Press, Oxford, 2nd edition.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, 3rd edition.

Jeffreys, H. (1980). Some General Points in Probability Theory. In Zellner, A. and Kadane, J. B., editors, *Bayesian Analysis in Econometrics and Statistics : Essays in Honor of Harold Jeffreys*, pages 451–453. North-Holland Publishing Company, Amsterdam, The Netherlands.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5):524–532.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1989). Optimization by Simulated Annealing: An Experimental Evaluation; Part I, Graph Partitioning. *Operations Research*, 37(6):865–892.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1991). Optimization by Simulated Annealing: An Experimental Evaluation; Part II, Graph Coloring and Number Partitioning. *Operations Research*, 39(3):378–406.

Johnson, T. R. and Kuhn, K. M. (2013). Bayesian Thurstonian models for ranking data using JAGS. *Behavior Research Methods*, 45(3):857–872.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317.

Joshi, V. M. (1976). A Note on Birnbaum's Theory of the Likelihood Principle. *Journal of the American Statistical Association*, 71(354):345–346.

Kadane, J. B. (1987). [Testing precise hypotheses]: Comment. *Statistical Science*, 2(3):347–348.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, 62(2):251–259.

Kalbfleisch, J. D., Berger, J., Dawid, A., and Sprott, D. (1986). On Principles and Arguments to Likelihood: Discussion. *The Canadian Journal of Statistics*, 14(3):194–199.

Kamary, K., Mengersen, K., Robert, C. P., and Rousseau, J. (2014). Testing hypotheses via a mixture estimation model. *arXiv preprint, https://arxiv.org/abs/1412.2044*, pages 1–37.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kelter, R. (2020a). Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Medical Research Methodology*, 20(88).

Kelter, R. (2020b). Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology*, 20(142).

Kelter, R. (2020c). Bayesian survival analysis in STAN for improved measuring of uncertainty in parameter estimates. *Measurement: Interdisciplinary Research and Perspectives*, 18(2):101–119.

Kelter, R. (2020d). New Perspectives on Statistical Data Analysis: Challenges and Possibilities of Digitalization for Hypothesis Testing in Quantitative Research. In Radtke, J., Klesel, M., and Niehaves, B., editors, *Proceedings on digitalization at the Institute for Advanced Study of the University of Siegen*, pages 100–108, Siegen. Institute for Advanced Study of the University of Siegen.

Kelter, R. (2020e). Simulation data for the analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research. *BMC Research Notes*, 13(452).

Kelter, R. (2021a). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, 36:1263–1288.

Kelter, R. (2021b). Bayesian and frequentist testing for differences between two groups with parametric and nonparametric two-sample tests. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1523.

Kelter, R. (2021c). Bayesian model selection in the M-open setting — Approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. *Journal of Mathematical Psychology*, 100.

Kelter, R. and Stern, J. (2020). The Full Bayesian Significance Test and the e-value - Foundations, theory and application in the cognitive sciences. *arXiv preprint, https://arxiv.org/abs/2006.03334*.

Kemeny, J. G. and Oppenheim, P. (1952). Degree of factual support. *Philosophy of Science*, 19(4):307–234.

Kenakin, T., Bylund, D. B., Toews, M. L., Mullane, K., Winquist, R. J., and Williams, M. (2014). Replicated, replicable and relevant-target engagement and pharmacological experimentation in the 21st century. *Biochemical Pharmacology*, 87(1):64–77.

Kendall, D., Bartlett, M., and Thornton, L. (1982). Jerzy Neyman. *Biographical Memoirs of Fellows of the Royal Society*, 28:379–412.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics*, 27(4):887–906.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by Simulated Annealing. *Science*, 220(4598):671–680.

Kirkwood, T. B. L. (1981). Bioequivalence Testing - A Need to Rethink. *Biometrics*, 37(3):589–594.

Kleijn, B. (2022). *The frequentist theory of Bayesian statistics*. Springer, Amsterdam.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2014). *Handbook of Survival Analysis*. Boca Raton.

Kolmogorov, A. (1950). *Foundations of the Theory of Probability*. Chelsea Pub. Co., New York.

Kordsmeyer, T. and Penke, L. (2017). The association of three indicators of developmental instability with mating success in humans. *Evolution and Human Behavior*, 38:704–713.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, 142(2):573–603.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, Oxford, 2nd edition.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

Kruschke, J. K. and Liddell, T. (2018a). Bayesian data analysis for newcomers. *Psychonomic*

Bulletin and Review, 25(1):155–177.

Kruschke, J. K. and Liddell, T. (2018b). The Bayesian New Statistics : Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, 25:178–206.

Laarhoven, P. J. M. and Aarts, E. H. L. (1987). *Simulated Annealing: Theory and Applications*. Springer Netherlands.

Lakatos, I. and Musgrave, A. (1970). *Criticism and the growth of knowledge*. Cambridge University Press, Cambridge.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Lane, D. A. (1980). Fisher, Jeffreys, and the Nature of Probability. In Fienberg, S. E. and Hinkley, D. V., editors, *R.A. Fisher - An Appreciation*, pages 148–160. Springer Verlag New York, New York.

Lee, M. D. and Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling : a practical course*. Cambridge University Press, Amsterdam.

Lehmann, E. (1993). The Fisher, Neyman-Pearson Theories of Testign Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88(424):1242–1249.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer, New York.

Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press, Cambridge.

Lenhard, J. (2006). Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science*, 57(1):69–91.

Lindley, D. (1957). A Statistical Paradox. *Biometrika*, 44(1):187–192.

Lindley, D. (1972). *Bayesian statistics, A Review*. Society for Industrial and Applied Mathematics, Philadelphia.

Lindquist, E. (1940). *Statistical analysis in educational research*. Houghton Mifflin, Boston.

Lindquist, E. (1953). *Design and analysis of experiments in psychology and education*. Houghton Mifflin, Boston.

Liu, J., Wong, W., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Liu, J. S. (2004). Molecular Dynamics and Hybrid Monte Carlo. In Liu, J. S., editor, *Monte Carlo Strategies in Scientific Computing*, pages 183–203. Springer Verlag New York, New York.

Loftus, G. R. (1991). On the Tyranny of Hypothesis Testing in the Social Sciences. *Contemporary Psychology: A Journal of Reviews*, 36(2):102–105.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Ly, A. (2017). *Bayes factors for research workers*. PhD thesis, University of Amsterdam.

Ly, A., Stefan, A., van Doorn, J., Dablander, F., van den Bergh, D., Sarafoglou, A., Kucharský, S., Derks, K., Gronau, Q. F., Raj, A., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., and Wagenmakers, E.-J. (2020). The Bayesian Methodology of Sir Harold Jeffreys as a Practical Alternative to the P Value Hypothesis Test. *Computational Brain & Behavior*, 3(2):153–161.

Ly, A., van den Bergh, D., Bartoš, F., and Wagenmakers, E. (2021). Bayesian inference with JASP. *The ISBA Bulletin*, 28(7-15).

Ly, A., Verhagen, J., and Wagenmakers, E.-J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, 72:43–55.

Ly, A., Verhagen, J., and Wagenmakers, E.-J. (2016b). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32.

Madruga, M. R., Esteves, L. G., and Wechsler, S. (2001). On the Bayesianity of Pereira-Stern tests. *Test*, 10(2):291–299.

Madruga, M. R., Pereira, C. A. d. B., and Stern, J. M. (2003). Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, 117(2):185–198.

Makowski, D., Ben-Shachar, M., and Lüdecke, D. (2019a). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, 4(40):1541.

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., and Lüdecke, D. (2019b). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10:2767.

Marin, J.-M. and Robert, C. (2014). *Bayesian Essentials With R*. Springer, New York.

Marino, M. J. (2014). The use and misuse of statistical methodologies in pharmacology research. *Biochemical Pharmacology*, 87(1):78–92.

Matthews, R. (1998). Bayesian Critique of Statistics in Health: The Great Health Hoax. Technical report, Department of Computer Science, Aston University, Birmingham.

Matthews, R., Wasserstein, R., and Spiegelhalter, D. (2017). The ASA's p-value statement, one year on. *Significance*, 14(2):38–41.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis?: What does 'failure to replicate' really mean? *American Psychologist*, 70(6):487–498.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, Cambridge.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. Chapman & Hall, CRC Press, Boca Raton.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press, Leipzig, 2nd edition.

McElreath, R. and Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE*, 10(8):1–16.

McGonigle, P. and Ruggeri, B. (2014). Animal models of human disease: Challenges in enabling translation. *Biochemical Pharmacology*, 87(1):162–171.

McGrayne, S. B. (2011). *The Theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press, Devon, Pennsylvania.

McKeigue, P. M., Campbell, H., Wild, S., Vitart, V., Hayward, C., Rudan, I., Wright, A. F., and Wilson, J. F. (2010). Bayesian methods for instrumental variable analysis with genetic instruments ('Mendelian randomization'): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome. *International Journal of Epidemiology*, 39(3):907–918.

McNemar, Q. (1949). *Psychological statistics*. John Wiley, New York.

McNemar, Q. (1955). *Psychological statistics*. John Wiley, New York, 2nd edition.

Mcouat, G. (2017). J. B. S. Haldane's passage to India: reconfiguring science. *Journal of Genetics*, 96(5):845–852.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2):103–115.

Mengersen, K., Robert, C. P., and Guihenneuc-Jouyaux, C. (1998). MCMC Convergence Diagnostics: A "Reviewww". In Berger, J., Bernado, J., Dawid, A., Lindley, D., and Smith, A., editors, *Bayesian Statistics 6*, volume 91, pages 415–440. Oxford University Press, Oxford.

Mengersen K. L. and Tweedie, R. L. (2012). Rates of Convergence of the Hastings and Metropolis Algorithms. *the Annals of Statistics*, 24(1):101–121.

Mestek, M. L., Plaisance, E., and Grandjean, P. (2008). The relationship between pedometerdetermined and self-reported physical activity and body composition variables in college-aged men and women. *Journal of American College Health*, 57(1):39–44.

Metropolis, N. (1987). The Beginning of the Monte Carlo Method. *Los Alamos Science*, (Special Issue):125–130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics*, 21(6):1087–1092.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247):335.

Meyn, S. and Tweedie, R. L. (2009). *Markov chains and stochastic stability, second edition*. Cambridge University Press, Cambridge.

Mills, J. A. (2018). Objective Bayesian Precise Hypothesis Testing. Technical report, University of Cincinnati.

Monnahan, C. C., Thorson, J. T., and Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3):339–348.

Moore, D. S., McCabe, G. P., and Craig, B. A. (2012). *Introduction to the practice of statistics*. W. H. Freeman, New York, 9th edition.

Morant, G. M. (1939). *A Bibliography of the Statistical and Other Writings of Karl Pearson (issued by the Biometrika Office, University College London)*. Cambridge University Press, Cambridge.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1):103–123.

Morey, R. D. and Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, 16(4):406–419.

Morey, R. D. and Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. *R package version 0.9.12-4.2*.

Moussouris, J. (1974). Gibbs and Markov random systems with constraints. *Journal of Statistical Physics*, 10(1):11–33.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100.

Natanegara, F., Neuenschwander, B., Seaman, J. W., Kinnersley, N., Heilmann, C. R., Ohlssen, D., and Rochester, G. (2014). The current state of Bayesian methods in medical product development: survey results and recommendations from the DIA Bayesian Scientific Working Group. *Pharmaceutical Statistics*, 13(1):3–12.

Neal, R. M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical report, Department of Computer Science, University of Toronto, Toronto.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer Verlag New York, New York.

Neal, R. M. (1997). Markov chain Monte Carlo Methods Based on 'Slicing' the Density Function. Technical report, Department of Statistics, University of Toronto, Toronto.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 31(3):743–748.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5, pages 113–162. Chapman and Hall/CRC, Boca Raton, 1st edition.

Neyman, A. J. and Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference : Part I. *Biometrika*, 20(1):175–240.

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 236(767):333–380.

Neyman, J. (1957). "Inductive Behavior" as a Basic Concept of Philosophy of Science. *Review of the International Statistical Institute*, 25(1):7–22.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A*, 231(694-706):289–337.

Neyman, J. and Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. *Statistical Research Memoirs*, 1:1–37.

Neyman, J. and Pearson, E. S. (1938). Contributions to the theory of testing statistical hypotheses, Parts II, III. *Statistical Research Memoirs*, (2):25–57.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., and Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, 48(4):1205–1226.

Nuzzo, R. (2014). Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487):150–152.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., and Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, 9(4):734–745.

Pashler, H. and Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, 7(6):531–536.

Pearson, E. (1933). A Survey of the Uses of Statistical Method in the Control and Standardization of the Quality of Manufactured Products. *Journal of the Royal Statistical Society*, 96(1):21–75.

Pearson, E. (1939). William Sealy Gosset, 1876-1937. *Biometrika*, 30(3-4):210–250.

Pearson, E. S. (1966). The Neyman-Pearson story. In David, F., editor, *Research Papers in Statistics* (*Festschrift for J. Neyman*). Wiley, London.

Pearson, E. S., Plackett, R. L., and Barnard, G. A. (1990). *"Student": a statistical biography of William Sealy Gosset: based on writings by E. S. Pearson*. Clarendon Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175.

Pereira, C. A. d. B. and Stern, J. M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110.

Pereira, C. A. d. B. and Stern, J. M. (2020). The e-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program. *São Paulo Journal of Mathematical Sciences*, pages 1–19.

Pereira, C. A. d. B., Stern, J. M., and Wechsler, S. (2008). Can a Significance Test be genuinely Bayesian? *Bayesian Analysis*, 3(1):79–100.

Perugini, M., Gallucci, M., and Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, 9(3):319–32.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612.

Philippe, A. and Robert, C. P. (2001). Riemann sums for MCMC estimation and convergence monitoring. *Statistics and Computing*, 11(2):103–115.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (*DSC 2003*).

Plummer, M. and Northcott, B. (2017). JAGS Version 4.3.0 User Manual.

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge, London, New York.

Popper, K. (2005). *The Logic Of Scientific Discovery*. Routledge, London.

Porter, T. M. (2006). *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton University Press, Princeton.

Pratt, J. W. (1961). Reviewed Work(s): Testing Statistical Hypotheses by E. L. Lehmann. *Journal of the American Statistical Association*, 56(293):163–167.

Pratt, J. W. (1962). On the Foundations of Statistical Inference: Discussion. *Journal of the American Statistical Association*, 57(298):307–326.

Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements. *Journal of the Royal*

Statistical Society: Series B (Methodological), 27(2):169–192.

Pratt, J. W. (1977). 'Decisions' as Statistical Evidence and Birnbaum's 'Confidence Concept'. *Synthese*, 36(1):59–69.

Quintana, D. S. and Williams, D. R. (2018). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP. *BMC Psychiatry*, 18(1):178.

Raftery, A. E. and Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. *Statistical Science*, 7(4):493–497.

Rao, C. R. (1992). R . A . Fisher : The Founder of Modern Statistics. *Statistical Science*, 7(1):34–48.

Rao, C. R. and Lovric, M. M. (2016). Testing point null hypothesis of a normal mean and the truth: 21st Century perspective. *Journal of Modern Applied Statistical Methods*, 15(2):2–21.

Reid, C. (1982). *Neyman*, volume 91. Springer, New York.

Richardson, S. and Green, P. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society Series B (Methodological)*, 59(4):731–792.

Richey, M. (2010). The evolution of Markov chain Monte Carlo methods. *American Mathematical Monthly*, 117(5):383–413.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association*, 87(419):861–868.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York.

Robert, C. and Casella, G. (2008). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science*, 26(1):102–115.

Robert, C. P. (2007). *The Bayesian Choice*. Springer New York, Paris, 2nd edition.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232.

Robert, C. P. (2015). The Metropolis-Hastings algorithm.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, 72(2009):33–37.

Robert, C. P. and Casella, G. (2010). *Introducing Monte Carlo methods with R*. Springer.

Robert, C. P., Chopin, N., and Rousseau, J. (2008). Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2):141–172.

Roberts, G. O. and Rosenthal, J. S. (1997). Markov chain Monte Carlo: Some practical implications of theoretical results. *The Canadian Journal of Statistics*, 26(1):5–20.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms. *Applied Probability Trust*, 44(2):458–475.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Roberts, G. O. and Tweedie, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110.

Rochon, J., Gondan, M., and Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12.

Rogers, J. L., Howard, K. I., and Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3):553–565.

Rosenthal, J. S. (2005). W.K. Hastings, Statistician and Developer of the Metropolis-Hastings Algorithm. *Web archive: http://probability.ca/hastings/ (accessed at 01/03/2021)*.

Rosenthal, J. S. (2014). Optimizing and Adapting the Metropolis Algorithm. In Lawless, J. F., editor, *Statistics in Action: A Canadian Outlook*, pages 93–108. Chapman and Hall/CRC.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, 16(2):225–237.

Rousseau, J. (2007). Approximating Interval hypothesis : p-values and Bayes factors. In Bernado, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics (Vol. 8)*, pages

417–452. Oxford University Press, Valencia.

Royall, R. (1997). *Statistical Evidence: A likelihood paradigm for statistical evidence*. Chapman and Hall, London.

Rubin, D. (1987). A noniterative sampling-importance resampling alternative to the data augmen-tation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82:543–546.

Rüschendorf, L. (2014). *Mathematische Statistik*. Springer.

Rusticus, S. and Eva, K. (2016). Defining equivalence in medical education evaluation and research: does a distribution-based approach work? *Practical Assessment, Research & Evaluation*, 16(7):1–6.

Salmon, W. C. (1966). *The Foundations of Scientific Inference*. University of Pittsburgh Press, Pittsburgh.

Salmon, W. C. (1988). Dynamic Rationality: Propensity, Probability, and Credence. In *Probability and Causality*, pages 3–40. Springer Netherlands, Dordrecht.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century: How Statisticians Revolutionized Science in the 20th Century*. Henry Holt and Company, New York.

Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):e1003285.

Savage, L., Bartlett, M., Barnard, G., Cox, D., Pearson, E., and Smith, C. (1962a). *The Foundations of Statistical Inference - A Discussion*. Methuen & Co Ltd., London.

Savage, L. J. (1954). *The Foundations of Statistics*. John Wiley & Sons, New York.

Savage, L. J. (1976). On Rereading R. A. Fisher. *Annals of Statistics*, 4(3):441–500.

Savage, L. J., Barnard, G., Cornfield, J., Bross, I., Box, G. E. P., Good, I. J., Lindley, D. V., Clunies-Ross, C. W., Pratt, J. W., Levene, H., Goldman, T., Dempster, A. P., Kempthorne, O., and Birnbaum, A. (1962b). On the Foundations of Statistical Inference: Discussion. *Journal of the American Statistical Association*, 57(298):307–326.

Sawilowsky, S. (2016). Rao-Lovric and the Triwizard Point Null Hypothesis Tournament. *Journal of Modern Applied Statistical Methods*, 15(2):11–12.

Schervish, M. J. (1995). *Theory of Statistics*. Springer Verlag, New York.

Schmidt, M. N. (2009). Function factorization using warped Gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York. ACM Press.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–80.

Sellke, T., Bayarri, J. J., and Berger, J. (2001a). Calibration of p-values for testing precise hypotheses. *American Statistician*, 55(1):62–71.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001b). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55(1):62–71.

Semmens, B. X., Ward, E. J., Moore, J. W., and Darimont, C. T. (2009). Quantifying inter-and intra-population niche variability using hierarchical bayesian stable isotope mixing models. *PLoS ONE*, 4(7):1–9.

Senn, S. (2001). Statistical issues in bioequivalance. *Statistics in Medicine*, 20(17-18):2785–2799.

Siebert, S., Machesky, L. M., and Insall, R. H. (2015). Overflow in science and its implications for trust. *eLife*, 4(e10825).

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, 26(5):559–569.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9).

Snedecor, G. W. (1937). *Statistical Methods*. Collegiate Press, Ames, Iowa, 1st edition.

Soper, H., Young, A., Cave, B., Lee, A., and Pearson, K. (1917). On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of "Student" and R. A. Fisher. *Biometrika*, 11(4):328–413.

Sprenger, J. (2016). Confirmation and Induction. In Humphreys, P., editor, *Oxford Handbook of the Philosophy of Science*. Oxford University Press, Oxford.

Sprenger, J. and Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford University Press.

Stan Development Team (2018). Stan Modeling Language Users Guide and Reference Manual.

Stern, J. M. (2003). Significance tests, Belief Calculi, and Burden of Proof in legal and Scientific Discourse. *Frontiers in Artificial Intelligence and its Applications*, 101:139–147.

Stigler, S. (2005). Fisher in 1921. *Statistical Science*, 20(1):32–49.

Stigler, S. M. (2006). How Ronald Fisher became a mathematical statistician. *Mathematics and Social Sciences*, 44(4):23–30.

Stigler, S. M. (2008). Karl Pearson's Theoretical Errors and the Advances They Inspired. *Statistical Science*, 23(2):261–271.

Student (1908a). The Probable Error of a Correlation Coefficient. *Biometrika*, 6(1):302–319.

Student (1908b). The Probable Error of a Mean. *Biometrika*, 6(1):1–25.

Sugden, L. A., Tackett, M. R., Savva, Y. A., Thompson, W. A., and Lawrence, C. E. (2013). Assessing the validity and reproducibility of genome-scale predictions. *Bioinformatics*, 29(22):2844–2851.

Tanner, M. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Thompson, E. A. (1990). R.A. Fisher's contributions to genetical statistics. *Biometrics*, 46(4):905–14.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1762.

Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18(17-18):2507–2515.

Turing, A. M. (1942). The Applications of Probability to Cryptography. *National Archives of the UK*, (HW 25/37).

Ulam, S. M. (1991). *Adventures of a mathematician*. University of California Press, California.

U.S. Food and Drug Administration Center for Drug Evaluation and Research (2001). Guidance for industry: Statistical approaches to establishing bioequivalence. *Web archive: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-approaches-establishing-bioequivalence (accessed 01/03/2021)*.

U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research (2016). Non-inferiority clinical trials to establish effectiveness: Guidance for industry. *Web archive: https://www.fda.gov/media/78504/download (accessed 01/03/2021)*.

U.S. Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research (2019). Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry. *Web archive: https://www.fda.gov/media/78495/download (accessed 01/03/2021)*.

U.S. Food and Drug Administration Center for Veterinary Medicine (2016). Guidance for industry: Bioequivalence: Blood level bioequivalence study VICH GL52. *Web archive: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cvm-gfi-224-vich-gl52-bioequivalence-blood-level-bioequivalence-study (accessed 01/03/2021)*.

van den Bergh, D., Clyde, M. A., Gupta, A. R. N., de Jong, T., Gronau, Q. F., Marsman, M., Ly, A., and Wagenmakers, E. J. (2021). A tutorial on Bayesian multi-model linear regression with

BAS and JASP. *Behavior Research Methods*, 53(6):2351–2371.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van Doorn, J., Ly, A., Marsman, M., and Wagenmakers, E.-J. (2020). Bayesian Rank-Based Hypothesis Testing for the Rank Sum Test, the Signed Rank Test, and Spearman's rho. *Journal of Applied Statistics*, 47(16):2984–3006.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., and Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin and Review*, 28(3):813–826.

van Laarhoven, P. (1988). *Theoretical and computational aspects of simulated annealing*. Phd thesis, Erasmus Universiteit Rotterdam, Rotterdam.

Vats, D. and Knudson, C. (2018). Revisiting the Gelman-Rubin Diagnostic. pages 1–22.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Vovk, V. G. (1993). A Logic of Probability, with Application to the Foundations of Statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):317–341.

Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., Ly, A., Verhagen, J., Selker, R., Sasiadek, A., Gronau, Q. F., Love, J., and Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, 6.

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., and Etz, A. (2020). The Support Interval. *Erkenntnis*.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., and Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60(3):158–189.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., and Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, 25(1):58–76.

Wagenmakers, E.-J. and Pashler, H. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Wald, A. (1939). Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics*, 20(2):165–205.

Wang, M. and Liu, G. (2016). A Simple Two-Sample Bayesian t-Test for Hypothesis Testing. *American Statistician*, 70(2):195–201.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a World Beyond "p<0.05". *The American Statistician*, 73(sup1):1–19.

Weber, R. and Popova, L. (2012). Testing equivalence in communication research: theory and application. *Communication Methods and Measures*, 6(3):190–213.

Westlake, W. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, 32(4):741–744.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., and Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests.

Perspectives on Psychological Science, 6(3):291–298.

Wetzels, R., Raaijmakers, J. G., Jakab, E., and Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin and Review*, 16(4):752–760.

Wilkinson, G. (1977). On Resolving the Controversy in Statistical Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(2):119–171.

Winquist, R. J., Mullane, K., and Williams, M. (2014). The fall and rise of pharmacology - (Re-)defining the discipline? *Biochemical Pharmacology*, 87(1):4–24.

Wrinch, D. and Jeffreys, H. (1919). LXXV. On some aspects of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 38(228):715–731.

Wrinch, D. and Jeffreys, H. (1921). XLII. On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42(249):369–390.

Wrinch, D. and Jeffreys, H. (1923a). I. The theory of mensuration. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 46(271):1–22.

Wrinch, D. and Jeffreys, H. (1923b). XXXVII. On certain fundamental principles of scientific inquiry (Second Paper). *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 45(266):368–374.

Yang, J., Levi, E., Craiu, R. V., and Rosenthal, J. S. (2019). Adaptive Component-wise Multiple-Try Metropolis Sampling. *Journal of Computational and Graphical Statistics*, 28(2):276–289.

Yang, J. and Rosenthal, J. S. (2017). Automatically Tuned General-Purpose MCMC via New Adaptive Diagnostics. *Computational Statistics*, 32:315–348.

Zabell, S. (1989a). R. A. Fisher on the History of Inverse Probability. *Statistical Science*, 4(3):247–256.

Zabell, S. L. (1989b). The rule of succession. *Erkenntnis*, 31(2-3):283–321.

Zehna, P. . (1966). Invariance of Maximum Likelihood Estimators. *The Annals of Mathematical Statistics*, 3(37):744.

Zellner, A. (1980). Introduction. In Zellner, A. and Kadane, J. B., editors, *Bayesian Analysis in Econometrics and Statistics : Essays in Honor of Harold Jeffreys*, chapter 1. Elsevier North-Holland, Amsterdam.

Ziliak, S. T. (2019). How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little " p" Is Not Enough. *The American Statistician*, 73(sup1):281–290.

Zumbo, B. D. and Kroc, E. (2016). Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective'. *Journal of Modern Applied Statistical Methods*, 15(2):11–2016.