Fiducial and Objective Bayesian Inference
History, Theory, and Comparisons

Leiv Tore Salte Rønneberg
Master’s Thesis, Autumn 2017
This master’s thesis is submitted under the master’s programme *Modelling and Data Analysis*, with programme option *Statistics and Data Analysis*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group $E_8$, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.
Abstract

In 1930, Fisher presented his fiducial argument as a solution to the "fundamentally false and devoid of foundation" practice of using Bayes' theorem with uniform priors to represent ignorance about a parameter. His solution resulted in an "objective" posterior distribution on the parameter space, but was the subject of a long controversy in the statistical community. The theory was never fully accepted by his contemporaries, notably the Neyman-Wald school of thought, and after Fisher's death in 1962 the theory was largely forgotten, and widely considered his "biggest blunder".

In the past 20 years or so, his idea has received renewed attention, from numerous authors, yielding several more modern approaches. The common goal of these approaches is to obtain an objective distribution on the parameter space, summarizing what might be reasonably learned from the data – without invoking Bayes' theorem.

Similarly, from the Bayesian paradigm, approaches have been made to create prior distributions that are in a sense objective, based either on invariance arguments, or on entropy arguments – yielding an "objective" posterior distribution, given the data.

This thesis traces the origins of these two approaches to objective statistical inference, examining the underlying logic, and investigates when they give equal, similar or vastly different answers, given the same data.
Preface

First, and foremost, I owe thanks to my supervisor Nils Lid Hjort, for turning my attention towards such an exciting topic. What started as a simple idea, comparing confidence distributions to Bayesian posteriors, grew into a much more philosophical thesis than initially planned. It has been tremendously interesting to read these old foundational papers, and tracing the history of the fiducial and objective Bayesian arguments from their inception in the early 20th century, through to the beginnings of the 21st. Learning statistics first from an applied background, where the focus is on how and when to apply which test, it has been illuminating to learn more about the underlying logic that went in to the design of frequentist theory as we know it today.

I am also grateful for my employer, Statistics Norway, for providing me with an office to work in, after the study hall at the mathematics department closed for renovations, and for letting me use their servers for computations – though I never asked for permission. Particularly I would like to thank my boss, Randi Johannessen, for making it possible to pursue a masters degree while keeping a full time job, and my coworkers for providing new energy during long days. In addition, gratitude is owed to the people at study hall 802 for making Blindern a fun and interesting place to be. A special thanks goes to Jonas for proofreading.

I would like to thank my mother, for her endless support and for always being available on the phone after a long day of writing.

Lastly, thank you Stine, for always cheering me on, for listening to my musings about statistical inference, for grammatical suggestions, and for simply being who you are.

Leiv Tore Salte Rønneberg
Oslo 15.11.17
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Probability</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1 Paradigms of statistical inference</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Outline of the thesis</td>
<td>7</td>
</tr>
<tr>
<td>1.3 A note on notation</td>
<td>8</td>
</tr>
<tr>
<td><strong>2 Frequentist distribution estimators</strong></td>
<td>9</td>
</tr>
<tr>
<td>2.1 Fiducial probability and the Bayesian omelette</td>
<td>9</td>
</tr>
<tr>
<td>2.1.1 Interpretation of fiducial distributions</td>
<td>12</td>
</tr>
<tr>
<td>2.1.2 Simultaneous fiducial distributions</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Generalized Fiducial Inference and fiducial revival</td>
<td>16</td>
</tr>
<tr>
<td>2.3 The Confidence Distribution</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1 Constructing CDs</td>
<td>20</td>
</tr>
<tr>
<td>2.3.2 Inference with CDs</td>
<td>21</td>
</tr>
<tr>
<td>2.3.3 Optimality</td>
<td>23</td>
</tr>
<tr>
<td>2.3.4 Uniform Optimality in the exponential family</td>
<td>27</td>
</tr>
<tr>
<td><strong>3 Objective Bayesian Inference</strong></td>
<td>29</td>
</tr>
<tr>
<td>3.1 The case for objectivity</td>
<td>29</td>
</tr>
<tr>
<td>3.2 The principle of indifference</td>
<td>30</td>
</tr>
<tr>
<td>3.3 Jeffreys’ Prior – the invariance path</td>
<td>32</td>
</tr>
<tr>
<td>3.4 E. T. Jaynes – the entropy path</td>
<td>35</td>
</tr>
<tr>
<td>3.5 Reference Priors</td>
<td>37</td>
</tr>
<tr>
<td>3.5.1 Motivation and Definition</td>
<td>37</td>
</tr>
<tr>
<td>3.5.2 Explicit Forms of the Reference Prior</td>
<td>41</td>
</tr>
<tr>
<td>3.5.3 Shortcuts to a reference prior</td>
<td>45</td>
</tr>
<tr>
<td>3.5.4 On compact subsets and uniqueness</td>
<td>46</td>
</tr>
<tr>
<td>3.6 Frequentist properties of the Bayesian posterior</td>
<td>47</td>
</tr>
<tr>
<td>3.6.1 Probability matching priors</td>
<td>48</td>
</tr>
<tr>
<td><strong>4 Comparisons and Examples</strong></td>
<td>49</td>
</tr>
<tr>
<td>4.1 Difference of exponential parameters</td>
<td>49</td>
</tr>
<tr>
<td>4.1.1 An optimal confidence distribution</td>
<td>49</td>
</tr>
<tr>
<td>4.1.2 Objective Bayesian analysis</td>
<td>50</td>
</tr>
<tr>
<td>4.1.3 Comparisons</td>
<td>54</td>
</tr>
<tr>
<td>4.1.4 Boundary parameters</td>
<td>63</td>
</tr>
<tr>
<td>4.2 Unbalanced Poisson pairs</td>
<td>64</td>
</tr>
</tbody>
</table>

V
## VI CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1</td>
<td>Exact Matching</td>
<td>67</td>
</tr>
<tr>
<td>4.3</td>
<td>Linear combination of Normal means</td>
<td>69</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Unknown variance</td>
<td>71</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Behrens-Fisher</td>
<td>73</td>
</tr>
<tr>
<td>4.4</td>
<td>The Fieller-Creasy problem</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>Concluding remarks</td>
<td>83</td>
</tr>
<tr>
<td>5.1</td>
<td>Exact matching and uniformly optimal CDs</td>
<td>83</td>
</tr>
<tr>
<td>5.2</td>
<td>Approximate matching and PMPs</td>
<td>84</td>
</tr>
<tr>
<td>5.3</td>
<td>Paired Exponentials</td>
<td>85</td>
</tr>
<tr>
<td>5.4</td>
<td>Behrens-Fisher</td>
<td>87</td>
</tr>
<tr>
<td>5.5</td>
<td>Epistemic probability</td>
<td>87</td>
</tr>
<tr>
<td>B</td>
<td>Bibliography</td>
<td>89</td>
</tr>
<tr>
<td>A</td>
<td>Proofs</td>
<td>93</td>
</tr>
<tr>
<td>A.1</td>
<td>Proof of Lemma 4.1</td>
<td>93</td>
</tr>
<tr>
<td>A.2</td>
<td>Proof of Lemma 4.2</td>
<td>96</td>
</tr>
<tr>
<td>A.3</td>
<td>The error in the normalizing constant</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>Selected R code</td>
<td>101</td>
</tr>
<tr>
<td>B.1</td>
<td>Algorithm from remark 4.2</td>
<td>101</td>
</tr>
<tr>
<td>B.2</td>
<td>Marginal posterior from section 4.3.1</td>
<td>102</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Since this thesis concerns itself with some unfamiliar concepts, it is natural to start it off by giving some historical context.

Around the start of the 20th century, the statistician’s toolbox consisted of a series of ad-hoc mechanisms for statistical inference. These included “Bayes theorem, least squares, the normal distribution and the central limit theorem, binomial and Poisson methods for count data, Galton’s correlation and regression, multivariate distributions, Pearson’s $\chi^2$ and Student’s t” (Efron 1998, p. 96). What was missing, says Efron, was a central core for these ideas. “There were two obvious candidates to provide a statistical core: ’objective’ Bayesian statistics in the Laplace tradition of using uniform priors for unknown parameters, and a rough frequentism exemplified by Pearson’s $\chi^2$ test. (Efron 1998, p. 97)

The core was to be supplied by Fisher in several landmark papers during the 1920s, which gave us many of the tools and concepts in modern estimation theory; sufficiency, maximum likelihood, Fisher information and more. There is no doubt that Fisher is the father of modern mathematical statistics, and the paradigm he laid out is in the same spirit as that of Pearson – crucially, it involves a complete rejection of the ’objective’ Bayesianism of Laplace. Fisher’s initial core was built upon by the works of Neyman and Wald over the next decades to provide the framework for frequentism as we know it today. With frequentist theory, the logic of statistical inference were put on a solid, perhaps narrow, mathematical framework – but one that did not depend on Bayesian reasoning.

While modern statistics is inherently a mathematical subject, it is also in essence an epistemological subject. The nature of statistical inference is to reason under uncertainty, about quantities that are often intrinsically unobservable, on the basis of smaller pieces of evidence, confirming or contradicting some hypothesis or prior beliefs. It was on a philosophical basis that Laplacian Bayesianism, with its ’uniform’ priors, was rejected in the first place, which led to the development of the frequentist school of thought. The goal was to be able to make inferences about unknown quantities, without appealing to Bayes theorem – especially in cases where a good prior distribution could not be given. The theory of Neyman and Fisher delivered what Zabell (1989, p. 247) deems; “a nearly lethal blow to Bayesian statistics”.

Of course, the Bayesian paradigm is alive and kicking, for several reasons. First and foremost, it works. Without worrying too much about philosophical foundations, the Bayesian estimation method provides good results, even in complex situations. Secondly, simulation methods have been created, that make calculations feasible even when the number of parameters are large, and the models highly
complex. Third, when framed correctly\footnote{see the next section}, the posterior distributions have a clear interpretation, more akin to the everyday interpretation of probability, and are completely coherent, i.e. marginal distributions can be obtained simply by integration, and regular probability calculus holds. Lastly, the Bayesian method provides a distribution of uncertainty over the entire parameter space, summarizing what may be reasonably inferred about the underlying parameters.

This last point is an appealing property of the Bayesian paradigm. A posterior distribution provides a quick and visual summary of the uncertainties present in the model, given the data. A sharp, localized posterior indicates that we can be quite certain about the location of our parameter, while a wide, diffuse posterior should lead us to be more careful in our judgements. Similarly, the frequentist confidence intervals provide a measure of our uncertainty, where by fixing a level of confidence $\alpha$, we can derive intervals that will cover the true parameter in an $\alpha$ proportion of experiments. A narrow interval at a high level of confidence means we can be fairly certain about the location of our parameter. The confidence intervals was a Neymanian construction, one which Fisher disapproved of. Instead, Fisher wanted to have a full distribution of his uncertainty, in the same fashion as the Bayesian paradigm. But he wanted it without using unwarranted prior information.

For this purpose, Fisher created his fiducial distribution, which aims to do precisely this – obtain a posterior distribution without unwarranted prior distributions. The fiducial argument isn’t found in modern textbooks, and it has been largely forgotten by the statistical mainstream. The reason being that it was surrounded by controversies, most of which had to do with how the resulting distributions should be interpreted, or how they should be constructed. In addition, Fisher kept insisting that he was in the right, even when most of his statistical colleagues thought he was in the wrong. Recently though, Fisher’s original ideas has received some renewed interest, spawned by heavy hitters in the field;
1.1 Probability

The *modus operandi* of many statisticians is not to think too hard about what probability *really* is. Often though, this can lead to misunderstandings, especially when communicating results to the public. During the 2016 US presidential election, Nate Silver’s blog, *FiveThirtyEight* ran a daily updated forecast of the election and each candidate’s probability of winning. On the election day, the probabilities where 71.4% in favour of Hillary Clinton winning the election, with Donald Trump estimated only at a 28.6% chance of winning. Several other media channels had similar results in favour of Hillary Clinton. We all know that Donald Trump won the election, but what followed, was an interesting debate from a statistical point of view. “How could the statisticians be so wrong?” was a commonly asked question. How could hundreds of polls, and people whose job it is to predict the outcome, be so utterly wrong?

I think the big question to ask here is: “Were they wrong, or is there a gap between the technicalities of mathematical probability and the common-sense interpretation of it?”

The modern mathematical construction of probability is set within measure theory. We start by defining a set, $\Omega$, a $\sigma$-algebra, $\mathcal{A}$, of measurable subsets of $\Omega$, and a measure $P$ that assigns a numerical value to elements $E \in \mathcal{A}$. We call the triple $\{\Omega, \mathcal{A}, P\}$ a probability space, if the measure $P$ adheres to the axioms laid down by Kolmogorov;

1. **Non-negativity:** For all subsets $E \in \mathcal{A}$, $P(E) \geq 0$.

2. **Unitarity:** $P(\Omega) = 1$.

3. **Countable additivity:** If $E_1, E_2, \ldots$ are mutually disjoint, then $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

As an example, consider rolling two dice. The set $\Omega$ is our sample space, the values our dice can take, we can denote this as pairs $\{1,1\}, \{1,2\}, \{1,3\}, \ldots, \{6,6\}$, representing the dice faces. The $\sigma$-algebra of measurable subsets, $\mathcal{A}$, denotes all events we may want to know the probability of. One event could be “the sum of the two faces equals 3”, another could be “the product of the two faces equals 9”. These seems like natural things to want to know the probability of, but the axioms from above gives no clear answer as to *how* these values should be assigned by the measure $P$. It may seem natural to assign probabilities according to the relative frequency of which they would occur in the space of all possible outcomes. Take the event “the sum of the two faces equals 3”. If we were to evaluate all the sums that we can possibly attain by rolling two dice, we can see that of all 36 combinations of die faces possible, only in two cases will the sum be three. If the first die rolls a one, and the second a two; or the first die rolls a two, and the second a one. We may then want to assign the probability $2/36 \approx 0.0556$ to this event. This is the relative frequency interpretation of probability, and one may check that it behaves according to the axioms above.

As the name indicates, this kind of probability is at the heart of the frequentist paradigm of statistics. It is sometimes also called *aleatory* probability, from the Latin noun *āleā*, translating roughly to “A game with dice, and in gen., a game of hazard

or chance. This kind of probability is intrinsically linked with that of games of chance, naturally occurring random variations and the like. It describes what would happen in the long run, if a process was repeated several times, and we took note of how often our event happened.

If we were to roll our dice $N$ times, and take note of how many times, $n$, the sum of their faces equals three, in the long run we would have

$$\frac{n}{N} \to P_F(\text{sum of faces equals three}) = \frac{2}{36} \text{ as } N \to \infty,$$

where the subscript $F$ indicates that the probability in question is one of frequency.

Going back to the example of the US election, employing this kind of probability interpretation, we would have $P_F(\text{Trump wins election}) = 28.6\%$. Following the interpretations given above, this should mean that if the election was repeated one thousand times, we should expect Trump to win 286 of these repeated elections—which doesn’t seem so improbable. It’s more probable than rolling a die and having it come up as one, $1/6 \approx 16.6\%$, which seems to happen way too often. But elections aren’t dice rolls, and they certainly cannot be repeated a thousand times under equal settings. Furthermore, I don’t think this is the interpretation most people had in mind when viewing the number 28.6\%.

I expect most people interpret the above probability as a measure of certainty, or at least that it should say something related to how certain one can be of the outcome. In everyday conversation, it is not uncommon to use expressions such as “it is likely that ...”, or “I will probably ...”. Clearly, these aren’t statements of long-run frequency. If I’m asked whether or not I’ll attend a party, and my response is “I will probably swing by”, the thought process behind this statement isn’t considering what would happen if the evening of the party was repeated numerous times. It is a qualification as to how certain my attendance is.

There seems to be some kind of duality in our notions of probability. On the one hand, we have our frequency interpretation of probability, connected to games of chance and random variations, but on the other hand, the way we use the language in daily life seem to represent degrees of certainty. This interpretation of probability is often called epistemic probability, after the philosophical term episteme, meaning knowledge, or understanding. The statement $P_E(\text{Trump wins election}) = 28.6\%$, where the subscript $E$ denotes epistemic probability, is a statement much closer to what the average person has in mind when using probability in his or her daily life. Namely a statement of how certain we can be, given all the evidence, that Trump would win the 2016 US presidential election.

Hacking (2006) traces the origins of probability theory back to its earliest inception in Europe in the 17th century, and finds that this duality has always been present. Probability as a concept started emerging through the works of Pascal, Fermat, Huygens and Leibniz in the decade around 1660. Many problems considered by these authors were aleatory in nature, concerning outcomes and strategies of certain games of chance, others were epistemic, concerning what could reasonably be learned from evidence—Leibniz, for example, wanted to apply probability in law, to measure degrees of proof.

What about the epistemic interpretation of probability, does it conform to the axioms from above? Well, we haven’t been given an instruction manual for how to

assign numerical values for a given outcome, so it is difficult to check – without a
mathematical recipe for how to assign these numerical values, how can we check
that the recipe conforms to the axioms?

We don’t yet have a rule for constructing the numerical values in the first place,
like we did in the relative frequency interpretation with dice. – simply counting the
outcomes. But there is a rule that tells us how to update our values, given new
information. And crucially, this rule will conform to the axioms of Kolmogorov, in
fact, it is a pretty direct consequence.

\textbf{Theorem 1.1 (Bayes’ Theorem).} Let \( A \) and \( B \) be events, and \( P(B) \neq 0 \). Then

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)},
\]

where \( P(A|B) \) denotes the conditional probability of \( A \), given the event \( B \). Or in
words, the probability of \( A \) given \( B \).

The important part now, is to always do our calculations within the system, using
Bayes’ theorem to update our prior beliefs in light of new information. In this way,
we can ensure that our epistemic probabilities always conform to the Kolmogorov
axioms, and are in fact valid probabilities in the technical sense. Note that there are
no subscript on the probabilities in the theorem. The reason is that Bayes’ theorem
will hold in any interpretation of probability, as long as it conforms to the axioms –
of course the interpretations will be different though.

Returning a last time, to the example of the US election. I think that most people
have an epistemic notion of probability in mind when presented with these quanti-
tative measures, and I do believe that Nate Silver, who is a well-known Bayesian,
also has this interpretation in mind. The question then remains; Were the polls
wrong?

To flip the question; at what numerical value would people feel assured that the
polls were right? Surely if the predictions where 90% in favour of Donald Trump,
and he won – they would be assured that the underlying techniques were good. At
50%, people might still find it reasonable, chalking it up to a “close-call” situation.
What about 40%, or 35%, or 28.6%? My point is that probability is hard, and to
reason under uncertainty is not always intuitive. The truth is probably somewhere in
between the two extremes, the polls might have been a bit off, but so is the general
public’s notion of how to reason about probability.

1.1.1 Paradigms of statistical inference

The two interpretations of probability has given rise to two distinct schools of
thought in the subject of statistical inference; commonly referred to as frequentism,
and Bayesianism – the names indicating which probabilities are underlying.

Modern statistical inference typically start with defining a statistical model for
the underlying phenomena we want to discuss. This model if often contingent on
the value of some underlying parameter \( \theta \), that tweaks some aspect of the the data-
generating process, \( f(x|\theta) \). Interest is typically on making inferences about this
parameter, based on observations \( \mathbf{x} = (x_1, \ldots, x_n) \) from the model in question.

In the frequentist paradigm of statistics, as outlined by Fisher, Neyman and Wald
during the first half of the 20th century, the probabilities are aleatory\footnote{We will see later that Fisher wasn’t clear on the distinction} representing
long-run frequencies. The underlying parameters in the model are considered fixed, unknown quantities, to be discovered from the data.

The way inference often proceeds in the frequentist paradigm is to find a statistic of the parameter, say $S(\mathbf{X})$, whose sampling distribution, under repeated replications, can be derived. Ideally, the sampling distribution is independent of the parameter in the model, and we can use it to formulate a test concerning some hypothesis $H_0$ we might have about $\theta$. Typically, the statistic is formed in such a way that, assuming $H_0$ is true, we should expect smaller values of the observed statistic $S(\mathbf{x})$. A larger value of the observed statistic should give us some evidence that $H_0$ may in fact be false. That is, given an outcome, we can calculate the probability, under $H_0$, that this, or something more extreme, happened simply by chance, and if this probability is small, our assumption that $H_0$ is true, should come into question.

This is a very strong and cogent logical argument, and one that resonates well with Karl Popper’s empirical falsification principle. Another strong point of the argument is that it is completely “objective”, there is no notions of “degree of certainty”, or prior beliefs, the discrediting of $H_0$ is a matter of how probable an outcome is. A weakness is that new techniques, test statistics, and sampling distributions, must be derived in each new case of consideration, making it a less cohesive theory.

In the Bayesian paradigm of statistics, the probability in question is epistemic – where it represents degrees of certainty about the parameter. The question of inference is still one of estimating the value of some true underlying parameter, but before we collect data, there is uncertainty about its value. Since probabilities now are epistemic, we can represent our knowledge, or uncertainty, about the parameter, prior to data collection, by a probability distribution. This is what is known as the prior probability distribution, and it is also what is typically meant when people say that the parameters are ‘random’ in the Bayesian paradigm. Once we have this, we can collect data, and update our prior beliefs in light of new information, to obtain a posterior distribution on the parameter space using Bayes’ theorem. The resulting posterior distribution will also be one of epistemic probability, representing our degree of certainty about the location of the parameter, in light of the new evidence we just observed.

The scheme is simpler than the frequentist paradigm, and it is a more coherent one. The same technique can be applied in each and every case, and the result is an (epistemic) probability distribution on the parameter space. It is also easy (theoretically) to include any preceding knowledge one might have about $\theta$, into the analysis – simply by changing the prior distribution to represent this.

If it is so simple and coherent, why was it rejected by the frequentists, who delivered “a nearly lethal blow” as Zabell put it? There is a question we haven’t tackled yet. I stated that we didn’t have a recipe for assigning numerical values for epistemic probability, but that we had a rule to update them, in light of new evidence. And that this rule, when used correctly, would provide probabilities that obeyed the Kolmogorov axioms. What came under attack by the frequentists was the question of how exactly the prior distribution should be assigned in the first place. Especially in cases where one might not have much prior knowledge to build upon.

Laplace suggested using uniform prior distributions for parameters that one had little or no knowledge about, and had in such an ‘objective’ Bayesian theory. There are some troubling consequences of the uncritical use of such priors, which was pointed out by many authors in the late 19th and early 20th century, and it even-
1.2. OUTLINE OF THE THESIS

The theory was eventually put back on a philosophically sound framework through the works of de Finetti and Savage\(^5\) amongst others – building on a more subjective notion of epistemic probability. In essence, a version where the probability distributions utilized are meant to represent a certain individual’s representation of knowledge. These distributions may vary significantly from individual to individual, even in light of the same data, depending on each individual prior knowledge ex-ante. While philosophically sound, it has no intentions of being an objective measure of uncertainty, in the sense of Laplace. The theory of objective Bayesianism was also put on a more solid footing, notably through the works of Harold Jeffreys whom we will get back to. The notion of (epistemic) probability here is one of impersonal degree of belief, as Cox (2006, p. 73) calls it, where the resulting distributions are to be interpreted as how a rational agent would assign probabilities, given the available information, or lack thereof.

Between these two large paradigms of statistical inference, Bayesian and frequentist – Fisher suggested his fiducial distribution as somewhat of a compromise, yielding what he felt was an “objective” epistemic probability distribution on the parameter space, on the basis of aleatory sampling distributions in the sample space, a sort of frequentist-Bayesian fusion. Alas, largely forgotten and in ill repute.

1.2 Outline of the thesis

As previously mentioned, the subject of the thesis is to study the interplay between Fisher’s fiducial argument (and modern variations of it), and the more “objective” forms of Bayesian inference. The focus is more on the history, and the underlying logic rather than on practical applications. Though statistics is a mathematical subject, its concern is an epistemological one. It is of interest to think twice about why we reason as we do, and what underlies our techniques and methods.

Chapter two gives an outline of Fisher’s original fiducial argument, as it was presented in Fisher (1930, 1935). Further, it gives an outline to its history, and the controversies surrounding it. A notable reference in this regard is Zabell (1992). The fiducial argument has been revived in the last few years, with modern approaches coming into play. The main focus here will be on confidence distributions (CDs), a purely frequentist take on the problem as exemplified by Schweder & Hjort (2016), Schweder & Hjort (2017), and Xie & Singh (2013), where certain optimality results in the spirit of Neyman-Pearson theory can be reached. I will also touch upon the generalized fiducial inference of for example Hannig (2009) and Hannig et al. (2016), and highlight a first connection to Bayesian theory.

In chapter three, the theory of objective Bayesian inference is studied. Laplace used uniform priors to represent ignorance, a principle known as the principle of insufficient reason, or the principle of indifference, which can go very wrong if not used carefully. I examine the history of objective Bayes, and proceed in a semi-chronological fashion, looking at proposed solutions for objective prior distribution, and examine how, when and why they go wrong. The focus is on the uniform priors of Laplace, Jeffreys’ invariant prior distribution, and the natural extension to the reference prior theory outlined in for example Bernardo (1979), Berger & Bernardo

\(^5\)cf. for example de Finetti (1937) and Savage (1954).
The underlying logic of these is founded on the information theoretical concept of entropy, which is introduced and discussed through the works of E. T. Jaynes, summarized in his book from 2003.

In chapter four, the two methods of obtaining a distribution on the parameter space, fiducial and objective Bayes, are compares across some examples. All these examples are focused, meaning that there is a single, scalar, parameter $\psi$ of interest. We wish our inference to be as good as possible for this single parameter, treating all other parameters in the model as nuisance parameters. In the context of the natural exponential family, there are optimal solutions available from the CD approach, and it is of interest to see whether or not these correspond to some Bayesian solutions, and for which prior distributions. Some classical problems are revisited, such as the Behrens-Fisher problem, and the Fieller-Creasy problem.

While the question of numerical agreement between fiducial and Bayesian posteriors is an old one, it hasn’t to my knowledge been studied in connection with uniformly optimal confidence distributions.

In the fifth, and final chapter, I give some concluding remarks and outline a few natural extensions to the topics covered.

1.3 A note on notation

Instead of including a full glossary, I will simply outline some rules of thumb for the notation in the thesis – most will be familiar. Sample spaces are denoted by calligraphic letters, $\mathcal{X}, \mathcal{Y}$, while parameter spaces by large greek letters, $\Theta, \Lambda$, etc., a notable exception being $\Phi$ which denotes the cumulative density of a standard normal distribution. Large letters $X, Y,$ and $Z$ denote random variables, while lower-case letters denote actual fixed observed values, sometimes with the subscript “obs”, like $x_{\text{obs}}$. Parameters are denoted by greek letters, $\alpha, \beta$, and a subscript zero, $\alpha_0, \beta_0$ denote the actual true, underlying value of parameters used to generate the data at hand. Bold versions of the above indicates vectors, i.e. $\mathbf{X} = (X_1, \ldots, X_n)$ is an $n$-dimensional vector of random variables. For prior distributions, $\pi_J$ will denote Jeffreys’ prior, while $\pi_R$ denotes a reference prior. The function $1_A(x)$ denotes the set function for the set $A$. 
Chapter 2

Frequentist distribution estimators

2.1 Fiducial probability and the Bayesian omelette

The concept of fiducial probability was first introduced by Fisher in a 1930 paper titled “Inverse Probability”. In it he criticizes the use of inverse probability methods, commonly known as Bayesian methods, when one has insufficient prior knowledge. Especially, he criticizes the use of flat priors to represent ignorance about a parameter; a practice he deems “fundamentally false and devoid of foundation” (Fisher 1930, p. 528). As an alternative he proposes what has come to be known as the "fiducial argument" to obtain a distribution function on the parameter space, like the Bayesian posterior distribution, but without the specification of a prior distribution. In the words of Savage (1961); Fisher attempts to “make the Bayesian omelet without breaking the Bayesian egg”. Below follows a short introduction to the initial fiducial argument as it was presented in Fisher’s 1930 paper, for a more thorough exposition of the rise (and fall) of fiducial inference see Zabell (1992).

The argument in the 1930 paper goes something like this: If $T$ is a continuous statistic and $p$ is the probability that $T \leq t$, for some value $t$, there is a relationship of the form:

$$p = F(t, \theta) = P_{\theta}(T \leq t).$$

(2.1)

If the exact value of $\theta$ is known, then for a fixed $p \in [0, 1]$, say 0.95, the equation above states that $t = t_{0.95}(\theta)$ is the 95th percentile of the sampling distribution of $T$. Fisher (1930, p. 533) writes:

this relationship implies the perfectly objective fact that in 5 per cent. of samples $T$ will exceed the 95 per cent. value corresponding to the actual value of $\theta$ in the population from which it is drawn.

What Fisher now realized was that, instead of viewing the parameter as fixed and finding percentiles of the sampling distribution of $T$ for each $p$; he could consider the observed value of the statistic $T = t_{\text{obs}}$ as fixed, and look for the values of $\theta$ solving (2.1) for each $p$. In the case where $t_p(\theta)$ is increasing in $\theta$, Fisher called this the fiducial $100(1 - p)$ percent value of $\theta$ corresponding to $t_{\text{obs}}$. He gives the following interpretation:

\footnote{Fisher kept making changes to his initial argument over the years as the theory’s shortcomings were pointed out.}
the true value of $\theta$ will be less than the fiducial 5 per cent. value corresponding to the observed value of $T$ in exactly 5 trials in 100.

This process of transferring the uncertainty from the statistic $T$ to the parameter $\theta$ is what constitutes Fisher’s fiducial argument. Note the purely frequentist interpretation Fisher gives; under repeated sampling, the true value of $\theta$ will be less than the (data dependent) fiducial 5% value in exactly 5% per cent. of the samples.

If the function $F$ in (2.1) is decreasing in $\theta$, then for an observed outcome of the statistic $T = t_{obs}$, Fisher termed the function

$$H(\theta) = 1 - F(t_{obs}, \theta)$$

(2.2)

the fiducial distribution for the parameter $\theta$. While the language may seem cryptic, the idea should be a relatively familiar one. Fisher uses the sampling distribution of the statistic $T$, to make a precise probability statement about the location of $\theta$ in the parameter space. Below is an example in modern notation

**Example 2.1 (Fiducial distribution for a Normal mean).** Consider $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, and put $T = n^{-1} \sum_i X_i \sim N(\theta, 1/n)$. Then equation (2.1) becomes

$$p = F(t, \theta) = \Phi\left(\sqrt{n}(t - \theta)\right),$$

(2.3)

where $\Phi$ is the standard normal cdf, and $t_p(\theta) = \theta + \Phi^{-1}(p)/\sqrt{n}$ is increasing in $\theta$. Then the fiducial argument amounts to inverting the equation and setting

$$\theta_p(t_{obs}) = t_{obs} - \frac{\Phi^{-1}(p)}{\sqrt{n}},$$

now viewing $\theta$ as a function of the data. Then Fisher notes that

$$P(\theta \leq \theta_p(T_{obs})) = P\left(\theta \leq t_{obs} - \frac{\Phi^{-1}(p)}{\sqrt{n}}\right) = P\left(\Phi^{-1}(p) \leq \frac{\sqrt{n}(T_{obs} - \theta)}{\sqrt{n}}\right) = 1 - \Phi\left(\Phi^{-1}(p)\right) = 1 - p$$

where $T_{obs}$ has been capitalized to make it clear that under repeated sampling, this is the random element. That is, under repeated sampling, the true parameter $\theta_0$ will be less than $\theta_p(T_{obs})$ in exactly $1 - p$ proportion of samples. This is the same as saying that the interval $(-\infty, \theta_p(T_{obs})]$ is a 100(1 − $p$)% confidence interval, to put it on more familiar terms.

Further then, since the function $F$ in (2.3) is decreasing in $\theta$,

$$H(\theta) = 1 - \Phi\left(\sqrt{n}(t_{obs} - \theta)\right)$$

is a fiducial distribution for $\theta$.

The fiducial distribution, $H(\theta)$ is a distribution function on the parameter space $\Theta$ in the mathematical sense that it is a right-continuous, strictly increasing function with;

$$\lim_{\theta \to \inf \Theta} H(\theta) = 0 \text{ and } \lim_{\theta \to \sup \Theta} H(\theta) = 1$$

(2.4)
The exact coverage property of its quantiles, i.e. that quantiles are one-sided confidence intervals, is what ensure that the resulting distribution function can be used to make inferences about the unknown parameter. This *exactness* property was crucial in an age where one could not simply run simulations to check that the coverage was approximate. Back then, without this property, the distribution function would be meaningless.

From this, Fisher derived the *fiducial density* as

$$h(\theta) = -\frac{\partial F(t, \theta)}{\partial \theta}.$$  

(2.5)

Note that if $F(t_{obs}, \theta)$ is *increasing* in $\theta$, the fiducial distribution is simply $H(\theta) = F(t_{obs}, \theta)$ and its density the derivative as before.

If we differentiate the fiducial distribution from the above example, we will see that the fiducial density of $\theta$ is that of a Normal distribution, with mean $\bar{x}_{obs}$ and variance $1/n!$ By a slight abuse of notation, I will denote this as

$$\theta^{\text{fid}} \sim N(\bar{x}_{obs}, n^{-1}),$$

meaning that $\theta$ is *fiducially distributed* as a Normal distribution with the stated mean and variance. But keep in mind that, being frequentists, $\theta$ is not a random variable. It is a fixed, but unknown real number. The only stochastic element here is $\bar{X}_{obs}$ that will change for every new dataset, depending on the samples selected.

With his argument, Fisher managed to transfer the *randomness* from the sample space into the parameter space, while maintaining the probability structure, without appealing to Bayes’ theorem! He had in fact been able to make the Bayesian omelette without breaking any eggs.

With modern eyes, Fisher’s 1930 paper contains nothing controversial, and as [Neyman (1934)](10.2307/1267198) pointed out, the approach is identical to what is now the pivotal quantity approach of constructing confidence intervals. In fact, the whole argument hinges on the existence of an underlying pivotal quantity in equation 2.1.

Before moving on to discuss the controversies, let’s do another example, one that avoids using a pivotal quantity.

**Example 2.2 (Fiducial distribution without a pivotal quantity).** While the initial argument needed a pivotal quantity to work, later techniques were derived to extend the Fiducial argument to cases where such a quantity doesn’t exist, or isn’t easily available. Consider an observation $x$ from a model with probability distribution $f(x; \theta_0)$ for a fixed $\theta_0$. The probability of observing this, or something smaller is captured in the cumulative density

$$F(x; \theta_0) = \int_{-\infty}^{x} f(t; \theta_0)dt.$$  

Next, consider observing $K$ new observations from the same model $(x_1, \ldots, x_K)$. Then, there exists $\theta_i \in \Theta$ corresponding to each $x_i$ such that $F(x_i; \theta_0) = F(x_i; \theta)$ for all $i = 1, \ldots, K$, irrespective of the true value $\theta_0$. Implicitly, the samples $x_1, \ldots, x_K$ generates samples $\theta_1, \ldots, \theta_K$ in the parameter space while preserving the probability

\[A pivotal quantity is a function of the data and the unknown parameter, whose distribution is independent of the parameter.\]
structure of the sample space. Then, if we take infinitely many samples \( x_1, x_2, \ldots \)
we implicitly define a distribution on \( \Theta \) through the above relationship.

The technique is due to Sprott (1963), and the idea is very much in sync with Fisher’s original idea, utilizing the sampling distribution of our data and transferring the randomness to the parameter space through a function \( F(x; \theta_0) \).

Consider observing \( x_{\text{obs}} \) from the binomial distribution, \( \text{Bin}(n, p) \). The cumulative density function can be written as

\[
P(X \leq x_{\text{obs}} | p) = \sum_{x=0}^{\text{obs}} \binom{n}{x} p^x (1 - p)^{n-x},
\]
or, in terms of the regularized incomplete beta function;

\[
P(X \leq x_{\text{obs}} | p) = (n - x_{\text{obs}}) \binom{n}{x_{\text{obs}}} \int_0^{1-p} u^{n-x_{\text{obs}}-1} (1 - u)^{x_{\text{obs}}} \, du.
\]

But, when \( x_{\text{obs}} \) is considered fixed and \( p \) random, this expression is also the cumulative distribution function of a Beta distribution with parameters \( n - x_{\text{obs}} \) and \( x_{\text{obs}} + 1 \). Thus the fiducial distribution for \( p \) given the observation \( x_{\text{obs}} \) is simply;

\[
p_{\text{fid}} \sim \text{Beta}(x_{\text{obs}} + 1, n - x_{\text{obs}}).
\]

Fisher would not have approved of this distribution, as he did not like the idea of applying his Fiducial argument on discrete distributions. The reason is that exact matching can only be obtained at certain levels of significance. We will return to this problem in section 4.2 □

The controversies associated with Fisher’s fiducial inference started in the years following the 1930 publication, and was either related to how the fiducial distribution should be interpreted, or issues in connection with multiparameter fiducial distributions.

### 2.1.1 Interpretation of fiducial distributions

In his 1930 paper, Fisher stressed that the logical context for fiducial inference was one of repeated sampling from hypothetical population with a fixed underlying parameter. He was careful to interpret the resulting quantiles and intervals in terms of what we may now recognize as coverage probability. At the same time though, even in his 1930 paper, he did regard the resulting distribution as a "definite probability statement about the unknown parameter" (Fisher 1930, p. 533, emphasis added), which surely is problematic given that it is a fixed real number, unless he is invoking an epistemic notion of probability. The reason behind this duality has to do with Fisher’s interpretation of probability, which on the one hand was purely aleatory — representing frequencies in a hypothetical infinite population, but at the same time epistemic — summarizing a rational agent’s degree of belief. He did not make a clear distinction between these, cf. Zabell (1992).

One of the problems we face when using fiducial inference (as well as confidence intervals) is that of utilizing two interpretations of probability at once. We are utilizing frequency argumentation and repeated sampling (aleatory probabilities) to establish what we might deem frequentist properties of fiducial quantiles and intervals. That is, by knowing something about the sampling distribution of \( X \) we can
2.1. FIDUCIAL PROBABILITY AND THE BAYESIAN OMELETTE

construct statements such as $P(\theta \leq T_\alpha(X)) = \alpha$. Now then, once data is collected and we have a $x_{obs}$ available for analysis, it is tempting to interpret the interval $(-\infty, T_\alpha(x_{obs})]$ as having probability $\alpha$ of containing the true parameter. But in the frequentist paradigm, the parameter is considered fixed, so the interval either contains $\theta$, or it doesn’t. Instead, by careful phrasing, we say that there is a $100\alpha\%$ chance that the interval contains the true parameter. An orthodox frequentist will be satisfied by this formulation, and not spend too much time worrying if a particular interval calculated from a single sample, actually contains the parameter or not, or how certain he can be that it does. Worrying about these things is closer to an epistemic interpretation of probability, something that isn’t present in frequentist theory. The problem, stated more generally, is to infer something from the outcome of a single case, when we only know what happens in the long run – philosophers refer to this problem as the problem of the single case.

For fiducial inference, it is clear that the distributions themselves can be understood in an aleatory sense, but that inferences made from a single distribution must be given an epistemic interpretation. In his early writings, Fisher did not spend much time on these issues, and he did not find it troublesome to start the analysis by considering $\theta$ as a fixed parameter under repeated sampling (aleatory), and then switch to an epistemic probability interpretation by the end – now regarding $\theta$ as random in the epistemic sense, with a fiducial distribution on the parameter space. In addition, he was clear that the resulting epistemic distributions were subject to ordinary probability calculus – that a fiducial distribution for $\theta^2$ could be found from that of $\theta$ by the usual rules. This is not in general true, and fails even in the simplest cases – Pedersen (1978) proves that the coverage of $\alpha$-level sets is strictly larger than $\alpha$ for all $\theta$ in a setup similar to that of example 2.1.

2.1.2 Simultaneous fiducial distributions

In Fisher (1935), the fiducial argument is extended to the multiparameter setting by an example. In the paper, the simultaneous fiducial distribution for $(\mu, \sigma)$ in a normal distribution is found from the jointly sufficient statistics $(\bar{x}, s^2)$ using a clever argument via the Student t-distribution.

In the paper, Fisher first derives the fiducial distribution of an additional observation $x$ from the same model, after first having observed $(\bar{x}_{obs}, s^2_{obs})$ from an original sample of size $n_1$. He then considers the more general case of observing $n_2$ new observations from the same model, and deriving a fiducial distribution for $(\bar{x}_{n_2}, s^2_{n_2})$ based off of this new sample. Fisher has in mind the two pivotal quantities,

$$t = \frac{\bar{x}_{obs} - \bar{x}_{n_2} \sqrt{n_1 n_2 (n_1 + n_2 - 2)}}{\sqrt{n_1 + n_2}} \left(\frac{1}{n_1 - 1} s^2_{obs} + \frac{1}{n_2 - 1} s^2_{n_2}\right)$$

and

$$z = \log(s_{obs}) - \log(s_{n_2}),$$

that are functions of the two samples, of which he knows the joint distribution. He then simply substitutes in the expressions of $t$ and $z$ to obtain a joint fiducial distribution for $\bar{x}_{n_2}$ and $s^2_{n_2}$. Then, letting $n_2 \rightarrow \infty$, the statistics $(\bar{x}_{n_2}, s^2_{n_2})$ converge to $(\mu, \sigma^2)$ and he obtains the joint fiducial distribution for $(\mu, \sigma^2)$ given the observed statistics $(\bar{x}_{obs}, s^2_{obs})$.

---

3This is the source of much confusion.

4A predictive fiducial distribution of sorts – analogous to the frequentist predictive intervals.
Furthermore, he shows that the marginal fiducial distributions of $\mu$ and $\sigma$, found simply by integrating out $\sigma$ or $\mu$ from the joint fiducial, are the fiducial distributions we would have arrived at by starting from the familiar pivotal quantities:

$$\frac{\sqrt{n_1}(\mu - \bar{x}_{\text{obs}})}{s_{\text{obs}}} \sim t_{n_1 - 1} \quad \text{and} \quad \frac{(n_1 - 1)s_{\text{obs}}^2}{\sigma^2} \sim \chi^2_{n_1 - 1}.$$ 

From this nicely behaved example he concludes boldly that

...it appears that if statistics $T_1$, $T_2$, $T_3$, ... contain jointly the whole of the information available respecting parameters $\theta_1$, $\theta_2$, $\theta_3$, ... and if functions $t_1$, $t_2$, $t_3$, ... of the $T$’s and $\theta$’s can be found, the simultaneous distribution of which is independent of $\theta_1$, $\theta_2$, $\theta_3$, ... then the fiducial distribution of $\theta_1$, $\theta_2$, $\theta_3$, ... simultaneously may be found by substitution. (Fisher 1935, p. 395)

...an extrapolation that in general isn’t true.

Again, Fisher regards the joint fiducial distribution as a regular probability distribution for the unknown parameters. This would entail that the distribution could be reparametrized into a parameter of interest $\psi(\mu, \sigma)$ and a nuisance parameter $\lambda(\mu, \sigma)$ to obtain the joint fiducial distribution for $(\psi, \lambda)$. Then $\lambda$ could be integrated out to obtain the marginal fiducial distribution for $\psi$. It turns out that this is not true in general, as there is no guarantee that the resulting marginal distributions will have the correct coverage. Even in this simple case, Pedersen (1978) proved that exact coverage is only obtained for interest parameters of the form $\psi = a\mu + b\sigma$, where $a$ and $b$ are constants.

Another example due to Stein (1959) illustrates that, when treating fiducial distributions as regular probability distributions, the exact coverage property could be lost.

Example 2.3 (The length problem). This slightly artificial example clearly illustrates that the fiducial distributions cannot be treated as general probability distributions, obeying standard probability calculus. Let $X_1, \ldots, X_n$ be independent random variables distributed as $N(\mu_i, 1)$ for $i = 1, \ldots, n$. Let the quantity of interest be $\sum_{i=1}^n \mu_i^2$, that we wish to obtain a fiducial distribution for. The fiducial distribution of $\mu_i$ from a single observation $X_i$ is simply $N(X_i, 1)$ as $\mu_i - X_i \sim N(0, 1)$. From standard probability calculus we then have that $\mu_i^2 \sim \chi^2(X_i^2)$ where $X_i^2$ is a non-centrality parameter. Further, by independence of the observations we have,

$$\sum_{i=1}^n \mu_i^2 \sim \chi_n^2 \left( \sum_{i=1}^n X_i^2 \right).$$

Now then, letting $\Gamma_n(\cdot, \lambda)$ denote the cdf of a non-central chi-squared distribution with $n$ degrees of freedom and non-centrality parameter $\lambda$, and $\Gamma_n^{-1}(\cdot, \lambda)$ its inverse, a natural one-sided $\alpha$-level fiducial interval is simply

$$\left[ \Gamma_n^{-1}(1 - \alpha; \sum_{i=1}^n X_i^2), \infty \right).$$

Stein (1959) proved that the coverage probability of this interval, for any value of $\alpha$ can be made smaller than any $\epsilon > 0$ simply by choosing a sufficiently large $n$. 

Stein compares this fiducial solution to a typical confidence interval for the same parameter, based directly on the distribution of $\sum X_i^2$, which is a non-central chi squared distribution with $n$ degrees of freedom and non-centrality parameter $\sum \mu_i^2$. For large $n$, this distribution can be approximated by $N(n + \sum \mu_i^2, 2n + 4 \sum \mu_i^2)$, and a confidence interval, analogue to the fiducial interval, is the set

$$\left\{ \sum \mu_i^2 : \sum X_i^2 > \sum \mu_i^2 + n + z_{1-\alpha} \sqrt{2n + 4 \sum \mu_i^2} \right\}.$$

Stein notes that if we were to base the fiducial distribution solely upon $\sum X_i^2$, ignoring the individual observations, the confidence and fiducial intervals would have been equal. It seems to matter when, in the derivation, the fiducial argument is utilized, and that once a fiducial distribution has been obtained, it cannot be manipulated further using regular probability calculus – contrary to Fisher’s claims.

The controversies associated with fiducial inference started with an example Fisher gave at the end of his 1935 paper. In the example he finds the marginal fiducial distribution for the difference of two normal means, in the case where the variances are unequal and unknown – the so-called Behrens-Fisher problem. First he finds the individual fiducial t-distributions for $\mu_1$ and $\mu_2$ from the regular pivotal quantities. He then finds the fiducial distribution of $\delta = \mu_2 - \mu_1$ as a convolution of these two independent fiducial distributions, and notes that his solution corresponds to the solution found by Behrens in 1929, and it provides an exact test for $H_0 : \delta = 0$. The problem with Fisher’s solution (and Behrens’) is that it isn’t exact. The resulting interval estimator for $\delta = 0$, gives rise to a hypothesis test with the wrong level of significance – something that was pointed out by Bartlett (1936). We will revisit this problem in chapter 4, and find that the critiques may not be so damning after all.

A back-and-forth between Fisher and Bartlett in the years 1936-1941, as well as vocal disagreements with Neyman, led Fisher to change much of his initial fiducial argument. By giving up the criterion of exact coverage, he could construct marginal fiducial distributions from joint fiducial distributions – viewing the resulting distributions simply as distributions of epistemic probability. He also gave up the unconditional nature of his argument, insisting instead that joint fiducial distributions should be based on sufficient statistics through a process of stepwise conditioning – in line with the emerging frequentist school of thought lead by Neyman and Wald. He insisted, until his death in 1962, that joint fiducial distributions constructed in this fashion are unique, and that regular probability calculus applies. Conjectures that, in their most unrestricted form, were repeatedly proved wrong.

While Fisher viewed his theory of fiducial inference as, "the jewel in the crown of 'ideas and nomenclature' for which he was responsible" (Zabell 1992 p. 370); after his death the theory has largely vanished from the statistical tool belt, and is by many considered as Fisher’s "one great failure" (Zabell 1992 p. 369).
2.2 Generalized Fiducial Inference and fiducial revival

In recent years, there has been a resurgence of ideas in the spirit of Fisher’s fiducial argument. The defining feature of these approaches is in line with Fisher’s goal: obtaining inferentially meaningful probability statements on the parameter space, often in the shape of a distribution, without introducing subjective prior information. These approaches include the Dempster–Shafer theory of belief-functions, confidence distributions (that will be introduced later), and the inferential model approach by [Martin & Liu (2015)].

A particularly well-developed approach is that of generalized fiducial inference (GFI) [Hannig (2009), Hannig et al. (2016)], which in the one-parameter case is identical to Fisher’s initial argument, but extends and generalizes to multi-parameter problems.

The starting point of the GFI approach is to define a data-generating equation, say

$$X = G(U, \theta),$$

(2.6)

where $\theta$ is the unknown parameter and $U$ is a random element with a completely known distribution which is independent of $\theta$. We imagine that the data at hand is created by drawing $U$ at random from its distribution, and plugging it into the data generating equation. In the easiest example, where $X$ is iid. normally distributed with mean $\mu$ and variance $\sigma^2$, we can write $X = G(U, \theta) = \mu + \sigma U$ where $U$ is a vector of iid. $N(0, 1)$ random variables, and $\theta^T = (\mu, \sigma)$.

Now, assuming the data-generating equation has an inverse $Q_y(u) = \theta$ for any observed $y$ and any arbitrary $u$ – Fisher’s initial fiducial argument simply amounts to finding the distribution of $Q_y(U^*)$ for an independent copy $U^*$ of the original $U$ used to generate the data. Samples from the fiducial distribution can be obtained by simulating $u_1^*, \ldots, u_n^*$ and plugging them into the equation. Notice the similarities to example 2.2 here. The existence of this inverse means that transferring the randomness from the sample space, into the parameter space can be done in a nice way, which was always the case in Fisher’s examples. But in order to generalize and extend Fisher’s initial argument, one needs to consider the case where this might not be as easy.

To obtain a more general solution, there are some technical difficulties one needs to sort out. First off, there is no guarantee that the inverse function $Q_y(u) = \theta$ even exists. The point being that the set, $\{\theta : y = G(u, \theta)\}$, for a given $y$ and $u$, could be either empty or contain more than one value of $\theta$. If the set is empty, the suggested solution is to restrict the distribution of $U$ to a subset where solutions do exist, and renormalize the distribution of $U$ on this set. The rationale behind this method is that, the data must have come from some value of $U$ for a given, but unknown, $\theta_0$. However, as [Hannig et al. (2016)] notes, the set where solutions do exist, will typically have measure zero, and conditioning on a set of measure zero can lead to strange results. In any case, care must be taken.

In the case where there are more than one value of the parameter that solves equation 2.6 for given values of $u$ and $y$, [Hannig et al. (2016)] suggests choosing

---

5Due to the Borel-Kolmogorov paradox, see e.g. [Jaynes (2003, sec. 15.7)]
one of the values, possibly with some sort of random mechanism. They give some examples that show that the extra uncertainty introduced by this mechanism doesn’t disturb the final inference much.

Delving too far into these technical details is beyond the scope of this subsection, but it’s worth noting that extending Fisher’s initially simple argument, is far from easy. It turns out though, that under fairly mild regularity assumptions, a simple expression can be found for the generalized fiducial density.

**Definition 2.1 (Generalized Fiducial Density (GFD)).** If \( \theta \in \Theta \subset \mathbb{R}^p \) and \( x \in \mathbb{R}^n \), then under mild regularity assumptions, cf. [Hannig et al. (2016, Appendix A)], the generalized fiducial density for \( \theta \) is of the form

\[
r(\theta|x) = \frac{f(x, \theta) J(x, \theta)}{\int_\Theta f(x, \theta') J(x, \theta') d\theta'},
\]

where \( f(x, \theta) \) is the likelihood, and

\[
J(x, \theta) = D\left( \frac{\partial}{\partial \theta} G(u, \theta) \bigg|_{u = G^{-1}(x, \theta)} \right).
\]

If (i) \( n = p \) then \( D(A) = |\det A| \), otherwise the function will depend on the which norm is used; (ii) the \( L_\infty \)-norm yields \( D(A) = \sum_{i=(i_1, \ldots, i_p)} |\det A_i| \); (iii) under the \( L_2 \)-norm, if the entries, \( \frac{\partial}{\partial \theta} G(u, \theta) \), all have continuous partial derivatives for all \( \theta \) and all \( u \), then \( D(A) = \left( |\det A^T A| \right)^{1/2} \).

The statement in equation [2.7] is, in a way a normalized likelihood, but with an added Jacobian to make sure that the resulting distribution function is proper over the parameter space – it has a certain Bayesian flavour, as illustrated in the following example from [Hannig et al. (2016, p. 1350)]

**Example 2.4 (GFD for the linear regression model).** Consider the simple linear regression model where \( Y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon_i \) where \( \epsilon_i \sim N(0, \sigma^2) \) for \( i = 1, \ldots, n \). Expressed on matrix form, this can be viewed as a data-generating function à la equation [2.6]

\[
Y = X \beta + \sigma U \text{ where } U \sim N_n(0, I),
\]

where \( X \) is the design matrix, assumed with full rank, and \( U \) is the vector of normally distributed errors. Now then,

\[
\frac{\partial}{\partial \theta} G(U, \theta) \bigg|_{u = G^{-1}(y, \theta)} = (X, \sigma^{-1}(y - X \beta)),
\]

and utilizing the \( L_2 \) norm we have that

\[
J(y, \theta) = \left( \det [(X, \sigma^{-1}(y - X \beta))^T (X, \sigma^{-1}(y - X \beta))] \right)^{1/2}.
\]

The matrix product inside the determinant can be written as a block matrix,

\[
\begin{bmatrix}
X^T X & \sigma^{-1} X^T (y - X \beta) \\
\sigma^{-1}(y - X \beta)^T X & \sigma^{-2} (y - X \beta)^T (y - X \beta)
\end{bmatrix}.
\]
and since $X^T X$ is invertible the determinant can be expressed as
\[
\det[X^T X] \sigma^{-2} \det [(y - X\beta)^T (y - X\beta) - (y - X\beta)^T X (X^T X)^{-1} X^T (y - X\beta)].
\]

After some manipulation we obtain that the determinant is equal to
\[
\det[X^T X] \sigma^{-2} \det \left[ y^T y - y^T X (X^T X)^{-1} X^T y \right] = \det[X^T X] \sigma^{-2} \text{RSS},
\]
where RSS denotes the residual sum of squares if $\hat{y}$ was estimated by regular OLS, i.e. using $\hat{\beta} = (X^T X)^{-1} X^T y$. Then we arrive at,
\[
J(y, \theta) = \sigma^{-1} \left| \det X^T X \right|^{1/2} \text{RSS}^{1/2}.
\]

Finally, the generalized fiducial density for $(\beta, \sigma)$ given the observed data is
\[
r(\beta, \sigma|X, y) \propto \sigma^{-n-1} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right\}.
\]

The Bayesian reader will recognize this as the Bayesian posterior distribution of $(\beta, \sigma)$ when utilizing Jeffreys’ prior $^6 \pi_J(\beta, \sigma) \propto \sigma^{-1}$. \hfill \Box

While the generalized fiducial approach extends Fisher’s original argument, it is subject to the same critiques pointed out in the 1930s and 40s – the resulting distributions often don’t have the exact coverage property, and they are not unique. However, with the dawn of computers and ease of simulations, it is easy to check that the resulting distributions often have close to the intended coverage, even for small $n$. For generalized fiducial inference, there is even a Bernstein-von Mises type theorem available that guarantees the asymptotic coverage of generalized fiducial intervals, see Hannig et al. (2016, Theorem 3). Uniqueness is not obtained however, since the density will depend both on the data-generating sequence as well as the norm employed in finding the Jacobian. This is a point that don’t seem to bother the modern writers as much as it did Fisher, who insisted on the coherency of his theory. We will later see that non-uniqueness gives rise to some natural questions of optimality, that in some cases can be answered.

Lastly there is the question of interpretation of generalized fiducial densities. Generalized fiducial inference is nestled within the frequentist paradigm, so the parameters are considered fixed, but unknown. As there is no room for epistemic probabilities in this paradigm, the distributions obtained are simply viewed as distribution estimators $^7$ for the parameters in the model. Distributions that summarize what can reasonably be learned from the observed data, without invoking probabilistic statements about inherently non-random quantities. Inferential validity is ensured through the properties mentioned above; approximate coverage for small $n$, and exact coverage asymptotically.

Again, it seems that Fisher’s initial idea can be a fruitful one. In my opinion, there is a tendency in the literature, that if simulation tools where around in Fisher’s day, and had he not insisted that his fiducial distributions could do more than they could, the fiducial argument would not have been shot down so unequivocally.

---

$^6$see section 3.3
$^7$see also section 2.3.2
2.3 The Confidence Distribution

Another approach to modern fiducial inference is that of confidence distributions (CDs), as laid out in the review paper by [4] Xie & Singh (2013), the book by Schweder & Hjort (2016), or Schweder & Hjort (2017). This approach follows the strictly Neymanian interpretation of the fiducial argument; as a method for constructing confidence intervals, and the resulting distributions are interpreted as distributions of coverage probability. That is, if \( C(\theta) \) is a CD for \( \theta \), any data-dependent set \( K_\alpha(x) \subset \Theta \) satisfying

\[
P(\theta \in K_\alpha(x)) = \int_{\Theta} 1_{K_\alpha(x)}(\theta) dC(\theta) = \alpha,
\]

(2.8)

carries the interpretation that, under repeated sampling, the set \( K_\alpha(x) \) will contain the true parameter value in approximately \( 100\alpha \% \) of samples. The probability in this statement is over the sample space, and so far we are in line with Fisher’s fiducial argument. The difference between CDs and fiducial distributions is mostly in their interpretation. As previously discussed, Fisher would start his analysis by treating the parameter as a fixed, unknown quantity – his probability was frequentist, and the stochasticity was in the sample space. Once the data was collected, and the fiducial distribution in place, he would now consider the parameter as being random in the epistemic sense, and regard his distribution as an epistemic probability distribution of the parameter. Through this argumentation he obtained a proper probability distribution on the parameter space, without needing to invoke subjective prior information. We’ve seen that this does not always work as intended. The CDs, on the other hand, does not have this final interpretation. They are instead considered simply as a collection of confidence statements about the unknown parameter, given the collected data – not as a distribution of the parameter itself. A useful distinction is made by [Xie & Singh (2013) p. 7]:

a confidence distribution is viewed as an estimator for the parameter of interest, instead of an inherent distribution of the parameter.

This places some restrictions on the theory, notably that CDs are one-dimensional. A general theory for defining multi-parameter confidence sets with exact coverage, is as far as I know still an open problem in the world of statistics. 

Cox (1958) was the first to invoke the term confidence distribution when comparing Fisher’s fiducial distribution to the Neymanian confidence intervals. He noted that the difference between the two is mostly due to presentation, and there is no reason to limit the Neymanian approach to only intervals on the parameter space. Instead he suggested constructing the set of all confidence intervals at each level of probability \( \alpha \) to obtain a distribution on the parameter space, and he called this a confidence distribution.

One method of creating a confidence distribution is by inverting the upper limits of one-sided confidence intervals. That is, given the outcome of an experiment \( x \), if \((-\infty, K(\alpha, x)]\) is a valid one-sided \( \alpha \)-level confidence interval for \( \theta \), and \( K(\alpha, x) \) is strictly increasing in \( \alpha \) for any sample \( x \), one could invert the upper endpoints to obtain the confidence distribution \( F(\theta) = K^{-1}(\theta) \) keeping \( x \) fixed. The resulting function is in fact a distribution function on the parameter space, obtained by carefully shifting the uncertainty from the sample space to the parameter space in the
same fashion as Fisher’s initial fiducial argument. In addition, since it is constructed from exact confidence intervals, all subsets will have the specified coverage probability — again exactly like Fisher’s 1930 argument. But there is no fiducial reasoning involved; Cox stressed that the resulting distribution is inherently frequentist. And while it is tempting to interpret statements such as \( P(\theta \leq K(\alpha, x)) = \alpha \) as a definite probability statement about the parameter, this is logically wrong because the statement \( \theta \leq c \) for a given constant \( c \), is either true or false.

A more modern definition of confidence distributions, and one not relying on inverting upper endpoints of confidence intervals is found in Xie & Singh (2013).

**Definition 2.2 (Confidence Distribution).** A function \( C_n(\cdot) = C_n(x, \cdot) \) on \( \mathcal{X} \times \Theta \rightarrow [0, 1] \) is called a confidence distribution (CD) for a parameter \( \theta \) if

i. For each \( x \in \mathcal{X} \), \( C_n(\cdot) \) is a cumulative distribution function on \( \Theta \).

ii. At the true parameter value \( \theta = \theta_0 \), \( C_n(\theta_0) := C_n(x, \theta_0) \), as a function of the sample \( x \), follows the uniform distribution \( U(0, 1) \).

If the uniformity requirement only holds asymptotically, \( C(\cdot) \) is an asymptotic confidence distribution (aCD).

The crux in the definition is requirement (ii.) that ensure exact coverage of the specified subsets. Analogue to Fisher in his 1930 paper, we will call the derivative (when it exists) the **confidence density** and denote it in lower-case: \( c_n(\theta) = C'_n(\theta) \). Sometimes the CDs will have the subscript \( n \) to indicate their dependence on the number of observations, sometimes it is omitted.

As noted before, one of the reasons why fiducial inference was considered a failure in its time was due to the non-uniqueness of the distributions and marginalization issues, but also because Fisher insisted that the theory could do more than it actually could. The CD theory does not fix any of these issues, but the goal isn’t to "derive a new fiducial theory that is paradox free" (Xie & Singh 2013, p. 4), but instead work within the limitations of the theory. It is easy to check that the fiducial distribution in example 2.1 is an exact confidence distribution, most one-dimensional fiducial distributions are. As for marginalization issues, the marginal distributions found from joint fiducial distributions are often approximate confidence distributions, see for example the Behrens-Fisher problem in chapter 4. Non-uniqueness isn’t really a concern either, but leads instead to some natural optimality considerations, see section 2.3.3.

### 2.3.1 Constructing CDs

The re-framing of Fisher’s fiducial argument strictly in terms of confidence opens up a bunch of tools for the theory, allowing it to become a natural part of the frequentist inferential toolbox, and to draw on techniques developed in the last 100 years of frequentist theory. In this section I will outline two ways to a confidence distribution.

**By way of a pivotal quantity:**

The typical, and somewhat canonical way of constructing a CD, is by a pivotal quantity. That is, a function of the data and parameter of interest \( \psi = a(\theta) \) whose distribution is independent of the underlying parameters in the model, \( \theta \). We’ve
already done this in example 2.1 without being explicit about where the pivotal quantity came into play. The technique goes as follows:

Let \( \text{piv}(\psi, X) \) be a pivotal quantity that is increasing in \( \psi \), and let \( G(\cdot) \) denote the cumulative density function of the pivotal quantity’s sampling distribution. Then, a natural CD for \( \psi \) is simply \( C(\psi) = G(\text{piv}(\psi, X)) \). Analogous, if the pivot is decreasing in \( \psi \), the natural CD is \( C(\psi) = 1 - G(\text{piv}(\psi, X)) \). It is quite easy to check that the resulting distribution has a uniform distribution when evaluated at the true value \( \psi_0 \) – due to the probability transform.

\[ Z = \sqrt{n}(\theta - \bar{X}), \]

which has a standard normal distribution. This means that a confidence distribution for \( \theta \), given observed data \( \bar{x}_{obs} \) is available as

\[ C(\theta) = \Phi \left( \sqrt{n}(\theta - \bar{x}_{obs}) \right). \]

Which is also the fiducial distribution we would have arrived at.

\[ \square \]

By way of ML-theory: Via maximum likelihood theory, there is often an easy way to find an approximate confidence distribution for a parameter of interest.

\[ Z = \frac{\sqrt{n}(p - \hat{p})}{\sqrt{\hat{p}(1 - \hat{p})}} \to N(0,1), \]

i.e. \( Z \) is asymptotically a pivotal quantity, increasing in \( p \), and thus an approximate CD is directly available as

\[ C(p) = \Phi \left( \frac{\sqrt{n}(p - \hat{p})}{\sqrt{\hat{p}(1 - \hat{p})}} \right). \]

\[ \square \]

This technique opens up a wide spectre of possibilities for the CD theory. It is for example of interest to study how fast the convergence to normality is, and if first order approximations can be improved upon. Endeavors in this direction is outlined in Schweder & Hjort (2016, ch. 7).

2.3.2 Inference with CDs

As discussed in the introduction, having the uncertainties fully represented by a distribution is quite nice. Distributions contain a wealth of information about the
CHAPTER 2. FREQUENTIST DISTRIBUTION ESTIMATORS

parameter, and can be a useful visual aid in presenting results. Typically, in frequentist analysis, it is common to give a point estimate, a confidence interval and perhaps a p-value for a pre-specified null hypothesis. For example, in a regression setting, one would typically give the maximum likelihood estimate $\hat{\beta}$, a 95% equal-tailed confidence interval around it, and a p-value for testing $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. In the CD setting, all of these are easily available, either directly visible from plots or calculable from the distribution function itself. Figure 2.1 shows a plot of a confidence density (borrowed from Xie & Singh (2013)) where some natural point estimators (mode, median and mean), a 95% confidence interval, and a one-sided p-value are annotated.

First off; point estimators. The natural point estimators from a confidence distribution are the mode; \( \tilde{\theta} = \arg \max_c c(\theta) \), the median; \( M_n(\theta) = C^{-1}(0.5) \), and the mean; \( \bar{\theta} = \int_\Theta \theta dC(\theta) \). Under mild assumptions all three can be shown to be consistent estimators of the true value \( \theta_0 \), see Xie & Singh (2013, Thm. 1). In many applications, when the distributions are unimodal, the estimators will all turn out to correspond with the maximum likelihood estimator, e.g. in example 2.7.

The distribution can also be used for hypothesis testing due to the intimate connection between confidence intervals and p-values. Consider testing the hypothesis $H_0 : \theta \geq c$ for some pre-specified value $c$. We would like to, under the null hypothesis, measure the amount of support that our data lends to $H_0$. In the CD setup, the amount of confidence we have in the statement amounts simply to

$$\int_{\Theta} 1_{\{\theta \geq c\}} dC(\theta) = 1 - C(c),$$

and in typical frequentist fashion, we should reject the null hypothesis if our confidence in the statement is low. Essentially, the confidence distribution is a p-value function corresponding to a certain test statistic.

While the confidence density of figure 2.1 is a useful graphical summary of our confidence, the preferred way to visualize CDs will be either the confidence distribution itself, $C(\theta)$, or the confidence curve (cc) as introduced by Birnbaum (1961); $cc(\theta) = |1 - 2C(\theta)|$.

In figure 2.2 a confidence distribution and a confidence curve is displayed. The confidence curve, in the right panel, gives a quick visual summary of equal tailed, confidence intervals. By fixing a level of confidence, say $\alpha = 0.8$, the corresponding interval $(a, b)$ can be read of by where the lines intersect. Similarly, the confidence
2.3. THE CONFIDENCE DISTRIBUTION

Figure 2.2: Confidence distribution, $C(\theta)$ in the left panel, and the corresponding confidence curve, $cc(\theta)$, in the right panel.

quantiles can be found by looking at the confidence distribution itself. Here we can see that our interval $(a, b)$ corresponds to the 10th and 90th percentile of $C(\theta)$, as they should.

Let’s do a quick visual comparison of two versions of the situation in example 2.5

Example 2.7 (Non-uniqueness). In example 2.5, we had the data $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, and we employed the pivot $\sqrt{n}(\theta - \bar{X})$ which is distributed as $N(0, 1)$. This pivot is increasing in $\theta$, so the natural confidence distribution is

$$C^1(\theta) = \Phi(\sqrt{n}(\theta - \bar{x})),$$

Another, perfectly valid way to a confidence distribution for $\theta$, is by estimating the standard deviation $s^2 = (n - 1)^{-1} \sum_i (x_i - \bar{x})^2$ and utilizing the pivot $\sqrt{n}(\theta - \bar{x})/s$, which has a Student’s t-distribution with $(n - 1)$ degrees of freedom. The natural CD based on this pivot will be

$$C^2(\theta) = T_{n-1} \left( \frac{\sqrt{n}(\theta - \bar{x})}{s} \right),$$

where $T_{n-1}$ is the cdf of the Student’s t-distribution with $n - 1$ degrees of freedom. Now we are in a situation where there are two perfectly valid, exact confidence distributions available for $\theta$. Confidence intervals from both distributions will have the desired coverage, but intuition suggests that the estimates from $C^2(\theta)$ will be more uncertain due to the variance being estimated instead of using its known, exact value. A quick look at the corresponding confidence curves in figure 2.3 will confirm this. The confidence intervals arising from $C^2(\theta)$ is wider than those arising from $C^1(\theta)$ at all levels of confidence.

The example above naturally leads to the question; "When several confidence distribution for the parameter of interest is available, which one is the best?"

2.3.3 Optimality

Due to the intimate connection between confidence intervals and hypothesis testing, the CD theory can fit fairly well into the Neyman-Pearson paradigm of hypothesis testing. Given a confidence distribution for a parameter of interest, $\psi$, the value
Figure 2.3: Comparison of $C^1(\theta)$ and $C^2(\theta)$ in example 2.7 for $\bar{x} = 3.67$, $s^2 = 2.21$ and $n = 8$

$C(\psi_0)$ corresponds to the p-value of testing $H_0 : \psi \leq \psi_0$ against $H_1 : \psi > \psi_0$, for any value of $\psi_0$. In this sense, the confidence distribution is nothing more than a p-value function, providing a measure of evidence of $H_0$. If the evidence provided in the data for $H_0$ is too low, the hypothesis is rejected.

Before we can state the major theorems that connect CDs to traditional frequentist optimality theory, there are a few definitions that need to be in place. First we need to discuss what we mean by optimal in the context of confidence distributions. In the example given in the preceding subsection, two CDs for the same parameter were constructed, one based on the jointly sufficient statistics $(\bar{X}, S)$ and the t-distributed pivotal quantity, and the other assuming $\sigma$ was known using the standard Z pivotal quantity. The two curves meet in the same point, i.e. have the same point estimate for the median, but confidence intervals read of at any level will be shorter for the CD assuming $\sigma$ is known. Again using the nomenclature of hypothesis testing, it is clear that utilizing the information that $\sigma$ is known gives more statistical power to the underlying test, as there is less uncertainty in the test procedure. Intuitively, we would like our CDs to be (i) centered around the true value $\psi_0$, and (ii) that the distributions are as tight as possible given the observed data – still preserving exact coverage.

Following the approach by Schweder & Hjort (2016, ch. 5), we start by formalizing these thoughts into a decision theoretic framework, working our way towards a statement of optimality. The first step is by defining confidence loss and risk – analogous to loss and risk for point estimators. In the following, it is handy to rep-
represent the confidence distribution by a random variable \( \psi_{CD} \), that might be viewed as a random draw from the CD after data has been collected – analogous to a draw from the Bayesian posterior distribution.

**Definition 2.3 (Confidence loss and risk).** Let \( \Gamma \) be a, non-negative function over \( \mathbb{R} \), with \( \Gamma(0) = 0 \). The confidence loss at \( \theta \) of the CD \( C(\psi, y) \), with respect to the penalty function \( \Gamma \), where \( \psi = a(\theta) \) is a focus parameter, is

\[
\text{lo}(\theta, C) = \int \Gamma(\psi' - \psi)dC(\psi', y) = E\Gamma(\psi_{CD} - \psi).
\]

The risk is then the expectation of this taken over the data \( Y \) under parameter value \( \theta \)

\[
R(\theta, C) = E_\theta \text{lo}(\theta, C) = E_\theta E\Gamma(\psi_{CD} - \psi).
\]

That is, we extend the notion of 'loss' for a single point by taking its expectation with regards to the confidence distribution. Then, we define risk by taking the expectation again, this time with respect to the data.

Under squared error loss, i.e. \( \Gamma(z) = z^2 \), the confidence loss become

\[
\text{lo}(\theta, C) = \text{Var } \psi_{CD} + (E\psi_{CD} - \psi)^2
\]

i.e the sum of the variance of \( C(\psi) \) and the squared estimation error. We recognize now that minimizing confidence loss is connected to our two wishes for an optimal CD – that it be tight (variance term small), and centered near, or at, the true value of \( \psi \) (estimation error small, preferably zero).

Now consider two confidence distributions for the same interest parameter, based on two different statistics \( S \) and \( T \), denote these CDs by \( C(\psi, S) = C^S \) and \( C(\psi, T) = C^T \) respectively. By comparing loss and risks of the two CDs, we are able to answer the question of which distribution is best – and also extend some optimality theorems from the uniformly most powerful (UMP) testing theory. We need two more definitions before the main theorems can be stated.

The first definition concerns itself with a notion of uniform optimality, i.e. that one confidence distribution is better than another, at all values of the true parameter.

**Definition 2.4 (Uniform optimality).** A confidence distribution \( C^S \) is uniformly optimal if for each penalty function \( \Gamma \), non-increasing on the negative half-axis and non-decreasing on the positive, with \( \Gamma(0) = 0 \), and for any other confidence distribution \( C^T \),

\[
\text{lo}(\psi_0, C^S) \leq \text{lo}(\psi_0, C^T)
\]

almost surely, for every value of \( \psi_0 \). If, on the other hand, we only have

\[
R(\psi_0, C^S) \leq R(\psi_0, C^T),
\]

for every value of \( \psi_0 \), we say that \( C^S \) is uniformly optimal in the mean. □
Note that the confidence loss is stochastic, as it will depend on the specific data point, $y_{\text{obs}}$, we are constructing the CD on – hence the almost surely statement in the definition.

We also need the notion of monotone likelihood ratio (MLR) in order for our CDs to be connected to regular UMP theory, see for example Casella & Berger (2002, ch. 8).

**Definition 2.5 (Monotone Likelihood Ratio (MLR)).** Let $T$ be a statistic with a distribution indexed by a real-valued parameter $\theta$, e.g. $p(t|\theta)$. This distribution has a monotone likelihood ratio (MLR) of, for every $\theta_2 > \theta_1$, $g(t|\theta_2)/g(t|\theta_1)$ is a non-decreasing function of $t$ on the set $\{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$. □

We are now ready to state the first of two theorems, this one concerning the use of sufficient statistics and confidence distributions. The theorem confirms our intuition from UMP testing theory, that CDs constructed on the basis of a sufficient statistic will perform better than CDs based on any other statistic.

**Theorem 2.1 (Rao-Blackwell for confidence distributions).** If the pivotal quantity has the same functional form in a statistic $T$, as in a sufficient statistic $S$, and if both the pivot and confidence quantile are convex in the statistic, the confidence distribution based on $S$ is uniformly as good in the mean. □

As it stands, the theorem is somewhat limited, needing convexity and only comparing pivotal quantities of the same functional form. Intuition tells us that this should hold in wider generality, but the theory is complicated by the fact that CDs do not obey regular probability calculus.

Finally, the main optimality theorem will be stated, which is a Neyman-Pearson type theorem for confidence distributions.

**Theorem 2.2 (Neyman-Pearson for confidence distributions).** A confidence distribution based on a sufficient statistic in which the likelihood ratio is everywhere increasing, is uniformly optimal. □

This statement is far more powerful than theorem 2.1, since it is a statement of uniform optimality, not simply optimality in the mean – in addition it is stated in full generality. The proof of this theorem, as well as the proof for Rao-Blackwell is found in Schweder & Hjort (2016, ch. 5), and I will not go into the details of them, but a heuristic explanation can be done for the Neyman-Pearson result.

Recall that the Neyman-Pearson lemma for composite hypothesis states that, if we are using a sufficient statistic $S$, whose distribution has an MLR, to test the hypothesis $H_0 : \psi \leq \psi_0$ versus $H_1 : \psi > \psi_0$. Then, any test that rejects $H_0$ if, and only if, $S > s_0$, is a UMP level $\alpha$ test, where $\alpha = P_{\psi_0}(S > s_0)$, i.e. the probability of this happening under the null. The p-value function of this test, is precisely the confidence distribution created on the basis of the same sufficient statistic, using the observed value of $s_{\text{obs}}$.
2.3.4 Uniform Optimality in the exponential family

The exponential family of distributions has certain properties that make it especially amenable to the optimality theorem in the past section. Proofs of the statements in the following paragraph can be found in for example Lehmann & Romano (2005, ch. 2-4).

A distribution is a member of the exponential family if its density can be written as,

\[ f(x|\theta) = h(x) c(\theta) \exp \left\{ \sum_{i=1}^{k} w_i(\theta) t_i(x) \right\}. \]  

(2.9)

The set of parameters \((\eta_1, \ldots, \eta_k) = (w_1(\theta), \ldots, w_k(\theta))\) are called the natural parameters of the family, and a characterization of the exponential family in terms of these yield the same structure as above, except for a change in the \(c(\theta)\) term.

If \(X_1, \ldots, X_n\) is a sample from a distribution with density \(f(x|\theta)\), the joint distribution of the \(X_i\)’s is also a member of the exponential family, and the statistics \(\left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)\) are jointly sufficient for the natural parameters.

If we have an exponential family of the form

\[ f(S, A|\psi, \lambda) = h(S, A) c(\psi, \lambda) \exp \left\{ \psi S(X) + \sum_{i=1}^{p} \lambda_i A_i(X) - k(\psi, \lambda_1, \ldots, \lambda_p) \right\}, \]  

(2.10)

then the conditional distribution of \(S\) given \(A\), is also an exponential family – whose distribution is independent of \(\lambda\). The statistics \((S, A)\) are jointly sufficient for \((\psi, \lambda)\), and \(S\) is sufficient for \(\psi\) for any fixed value of \(A\), i.e. \(A\) is ancillary. Lastly, all members of the exponential family has the MLR property, cf. definition 2.5.

Then, there exists UMP tests for testing hypothesis of the form \(H_0 : \psi \leq \psi_0\) vs. \(H_1 : \psi > \psi_0\), and by the theory from the last section, there also exists an optimal confidence distribution. We have the following theorem from Schweder & Hjort (2016, Thm. 5.11)

**Theorem 2.3 (Optimal CDs for exponential families).** Let \(\psi\) be a scalar parameter of interest, and \(\lambda\) a \(p\)-dimensional vector of nuisance parameters in a continuous exponential model, with density

\[ f(x, \psi, \lambda) = h(x) \exp \left\{ \psi S(x) + \sum_{i=1}^{p} \lambda_i A_i(x) - k(\psi, \lambda_1, \ldots, \lambda_p) \right\}, \]  

for data vector \(x\) in a sample space independent of the parameters. Assume \((\psi, \lambda)\) is contained in an open \((p + 1)\)-dimensional parameter set. Then

\[ C_{S|A}(\psi) = P_\psi(S > s_{obs}|A = a_{obs}) \]  

is the uniformly most powerful confidence distribution for \(\psi\).

This theorem provides a quick and easy recipe for finding optimal CDs for a parameter of interest, and an effective way of dealing with nuisance parameters – in a large family of distributions.

Sometimes, the conditional distribution of \(S\) given \(A\) is available in closed form, but often it will need to be simulated by an MCMC type algorithm.

Let’s do two quick examples to illustrate the theorem.
Example 2.8 (Normal distribution). Let’s first check that the example that has followed us through this chapter is optimal. We had data $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$, which makes the joint distribution

$$f(\mathbf{x}|\theta) \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2 \right\}$$

$$\propto \exp \left\{ \theta \sum_{i=1}^{n} X_i - n\theta^2/2 \right\},$$

which is of the form from Theorem 2.3 – with $S(\mathbf{x}) = \sum X_i$. Hence, an optimal confidence distribution is by

$$P(\sum X_i > \sum x_{i,obs}) = P(\bar{X} > \bar{x}_{obs}),$$

which is course is the same distribution we based our original CD on – since $\bar{X} \sim N(\theta, 1/n)$. □

Of course, the power in this theorem first becomes visible when there are nuisance parameters present in the model.

Example 2.9 (Normal with nuisance parameter). Let $X \sim N(\mu, 1)$ and $Y \sim N(\mu + \psi, 1)$, where $X$ and $Y$ are independent. The joint distribution is of the form

$$f(x, y|\mu, \psi) \propto \exp \left\{ -\frac{1}{2} \left[ (x - \mu)^2 + (y - (\mu + \psi))^2 \right] \right\}$$

$$\propto \exp \left\{ y\psi + (x + y)\mu - k(\mu, \psi) \right\},$$

which is of the form of the theorem. A uniformly optimal CD for $\psi$ is simply by

$$C(\psi) = P(Y > y_{obs}|X + Y = x_{obs} + y_{obs})$$

$$= P(Y - X > y_{obs} - x_{obs}|X + Y = x_{obs} + y_{obs})$$

$$= P(Y - X > y_{obs} - x_{obs}),$$

since $Y - X$ is independent of $X + Y$ – this is a well-known property of the normal distribution. Then $Y - X \sim N(\psi, 2)$ and a confidence distribution is directly available from the standard pivotal quantity approach. □

We will see other examples as well of the theorem in chapter 4, where we look at connections between optimal confidence distributions and Bayes’ procedures using objective prior distributions – which is the topic of the next chapter.
Chapter 3

Objective Bayesian Inference

In this section, the theory of objective Bayesian inference is reviewed. With 'objective', I mean theories developed within the Bayesian framework that attempt to develop prior distributions, not on the basis of subjective pre-existing knowledge, but by some formal set of rules. The use of such rules has a long history, dating back to the works of Bayes himself, and of Laplace in the early 1800s. We’ve seen that the use of flat priors to represent ignorance was an idea Fisher felt was "fundamentally false", and many other influential figures in the early 1900s agreed. Keynes (1921), for example, devotes an entire chapter to the subject in his treatise on probability.

There are many approaches one can take at this point, but they all try to answer the same question in some sense. Namely, if a state of complete ignorance can be represented by a function on the parameter space, what should it look like? How does one parametrize ignorance? This is more of an epistemological question than a mathematical one, and it forces us to think hard about questions concerning the logic of inference. Also, from a mathematical point of view, the implementation once a prior is found, is often straight forward. Care must be taken to ensure that the posterior distribution is proper, but other than that, the concerns have mainly been of a philosophical nature.

Due to the large amount of possible approaches one can take, and the limitations of a thesis, only a few theories are reviewed in the following – for a fuller review see for example Kass & Wasserman (1996). The goal is to compare some of the objective Bayesian theories with the fiducial inference approaches outlined in the previous chapter. The main focus will be on Jeffreys’ prior and the reference prior approach, as developed by Bernardo (1979) and many others.

But first, why should we even care about objectivity?

3.1 The case for objectivity

It has been argued that the Bayesian paradigm of statistics is a subjective one. The probability in question is an epistemic, representing various states of individual knowledge, and the prior distribution is meant to capture any initial state of knowledge before new data is collected. Once new data is obtained, the prior state of knowledge is updated to a posterior state of knowledge, in light of the new evidence. While this subjectivist view of Bayesian inference is logically appealing, a natural question to ask is whose knowledge is being represented, and on an even deeper level, how is knowledge represented in a prior distribution?
Not including pre-existing prior information into the analysis could be regarded as a mortal sin of plausible reasoning, but there are several reasons why a more objective version of Bayesian inference should be desired. The first, and most obvious reason, is in the situation where no prior knowledge really exists. One could argue that some prior information always exist, as there often are physical constraints on the underlying parameters in question. But in more complex models, such as hierarchical models often encountered in modern applications, the parameters can lose a physical interpretation, or at least be hard to interpret. We still need to specify a prior distribution on the parameter, hopefully one that doesn’t disturb final inferences too much.

Another main reason is that the "appearance of objectivity", as Berger (2006) calls it, is often required of our results. Researchers in a wide range of fields look to statistics to provide them with an objective verification of their findings – letting the numbers speak for themselves. Very often this involves frequentist methods such as hypothesis testing and confidence intervals, where the analysis depends exclusively on the data, and the methods properly calibrated. For many, the specification of a prior distribution based on previous knowledge seems shady, making inference unreliable.

Even when a suitable prior distribution is available, such as a clinical drug trials where expert opinions can be incorporated, one often choose to use frequentist methods instead. The reason being as stated earlier, the results can feel somewhat tampered with by the ambiguity of how the prior was selected.

Based on these reasons, we should want a more objective version of the Bayesian paradigm. We should also hope that such a theory is coherent, logically consistent and easy to use for researchers also outside of mathematical community. As we will see in this chapter, and in examples later, the theory of objective Bayesian inference is none of these things.

3.2 The principle of indifference

We will start from the earliest application of what we would call objective Bayesian methods. Suppose we have a model with a parameter of interest, of which nothing is previously known. How do we represent our complete ignorance in this setup?

The tradition, going back to Bayes and Laplace, was to appeal to the principle of insufficient reason as a general recipe for creating a prior, Keynes called it the principle of indifference. The principle states that when there is no a priori evidence available to the researcher to indicate that some values of \( \theta \) are more likely than others, one must assume that they are all equally likely, i.e utilizing a 'flat' prior: \( \pi(\theta) \propto 1 \). In the case of a discrete and finite parameter space, \( \theta \in \{u_1, \ldots, u_N\} \), the principle leads to the appropriately normalized prior \( \pi(\theta) = 1/N \). Which intuitively seems correct; if one has absolutely no reason to believe that heads are more probable than tails when flipping a coin, the only way to quantify ones prior beliefs is to give the two outcomes equal probability. Deviating from this simple situation the principle quickly runs into trouble. If the parameter space is infinite, like \( (-\infty, \infty) \) or \( (0, \infty) \), the principle will lead to an improper prior, i.e. one that cannot be normalized. Philosophically this entails that it is not a proper representation of

\footnote{Whatever that is supposed to mean... I find the notion that “numbers are objective” to be nonsensical.}
knowledge, and it is not obvious that Bayes theorem even applies in this case. Even more worrying, what are we applying the principle of indifference to? Can we always pick elements from an infinite set? And if so, if they receive equal weight, the probability of each one will be zero so that only intervals of the parameter space will have a non-zero probability. Pragmatically, let’s not worry about the axiom of choice, and simply deem an improper prior as permissible if it leads to a proper Bayesian posterior distribution.

Another logical flaw, and more worrying, that appears in continuous parameter spaces, is that the flat prior isn’t invariant under monotone transformations. If the researcher is truly ignorant about the parameter $\theta$, and uniformity really represents this ignorance, he should be equally ignorant about a one-to-one transformation of it, $\phi = a(\theta)$, and the transformed prior

$$
\pi_\phi(\phi) = \pi_\theta(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| 
$$

(3.1)

should carry the same information, or lack thereof.

**Example 3.1 (The two indifferent statisticians).** As an example, consider the case where we have data, $X$, from a binomial distribution with parameters $n$ and $p$, and we wish to estimate the probability of success, $p$. Let’s assume that the data at hand is given to two different statisticians for analysis. One of the statisticians is interested the quantity, $p$ directly, while the other is more interested in estimating the odds, $R(p) = p/(1-p)$. There is no prior knowledge available for the probability $p$, so each statistician chooses to appeal to the principle of indifference.

For the first statistician, the prior distribution is $\pi_1(p) \propto 1$ on the interval $(0, 1)$ which leads directly to the posterior distribution

$$
\pi_1(p|x) \propto p^x(1-p)^{n-x} \sim \text{Beta}(x + 1, n - x + 1).
$$

The second statistician, who is more interested in estimating the odds, rewrites the likelihood of the distribution as

$$
L(R) \propto \left( \frac{R}{1+R} \right)^x \left( \frac{1}{1+R} \right)^{n-x},
$$

and then appeals to the principle of insufficient reason to obtain a prior $\pi_2(R) \propto 1$ over the interval $(0, \infty)$. The posterior distribution he obtains is thus directly proportional to the likelihood above.

Later, the two statisticians meet to compare their analysis, and in order to do so, the second statistician transforms his posterior distribution back to the parameter $p$. Since, in the posterior distribution, the parameter $p$ is considered random, he multiplies by the Jacobian $dR/dp = (1-p)^{-2}$ and obtain the posterior distribution

$$
\pi_2(p|x) \propto p^x(1-p)^{n-x-2} \sim \text{Beta}(x + 1, n - x - 1).
$$

Thus, starting from the same data, and applying the same logical principle for representing ignorance, the two statisticians have obtained different posterior distributions.

The example above should lead us to believe that the principle of indifference is badly flawed, and it is for this reason that it has been scrapped as a general rule
for objective priors. Over the years, several new principles have been developed for finding objective prior distributions. These efforts can mainly be divided into two broad camps, one which aims at fixing invariance issues like the one above, and the other using information theoretic approaches to best describe states of ignorance.

### 3.3 Jeffreys’ Prior – the invariance path

Based mainly on invariance arguments, Jeffreys (1946) proposed a general rule for the construction of a non-informative prior. Letting $I(\theta)$ denote the Fisher information matrix, with elements

$$I(\theta)_{ij} = E \left( -\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right),$$

the prior proposed is;

$$\pi_J(\theta) \propto \det \left( I(\theta) \right)^{1/2}. \quad (3.2)$$

In the one-dimensional case, checking that this prior has the invariance property of (3.1) is straight-forward. Let $X$ be distributed according to some model $f(x|\theta)$ and let $\phi(\theta)$ be a monotone transformation, and denote its inverse by $\theta(\phi)$. Then

$$\pi(\phi) \propto \sqrt{I_\phi(\phi)} = \sqrt{E \left[ \left( \frac{\partial f(x|\phi)}{\partial \phi} \right)^2 \right]}$$

by the chain-rule of differentiation and the definition of the Fisher information.

Let’s revisit the binomial example from the last section, and check that this prior really is invariant.

**Example 3.2 (The two invariant statisticians).** Assume now, that the two statisticians from the previous example had used Jeffreys’ prior to parametrize their ignorance. Again, let $X \sim \text{Bin}(n,p)$ with interest on the probability of success, $p$. The log-likelihood of this model can be written as

$$\ell(p) = \text{const.} + x \log p + (n - x) \log (1 - p),$$

from which the Fisher information is easily derived

$$I(p) = -E \left( \frac{\partial^2 \ell}{\partial p^2} \ell(p) \right) = \frac{1}{p(1-p)}.$$
Then, Jeffreys’ prior is simply $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$ in the original parametrization. The second statistician, who works with the rewritten likelihood obtains the log-likelihood

$$\ell(R) = \text{const.} + x \log R - n \log (1 + R),$$

which leads the the Fisher information $I(R) = n/(R(1 + R)^2)$ and thus Jeffreys’ prior as the square root. The two marginal posteriors then become

$$\pi_1(p|x) \propto p^{x-1/2}(1-p)^{n-x-1/2} \sim \text{Beta}(x + 1/2, n - x + 1/2)$$

$$\pi_2(p|x) \propto \pi_2(R(p)|x) \frac{1}{(1-p)^2} = p^{x-1/2}(1-p)^{n-x-1/2} \sim \text{Beta}(x + 1/2, n - x + 1/2).$$

Now the two statisticians can compare results, and no seemingly strange paradoxes occur. By using Jeffreys’ prior, the amount of information, or lack thereof, is equal in each parametrization. □

It is a bit difficult to give a good explanation as to why Jeffreys’ prior in a sense represents ignorance. Many authors don’t like the idea of functions representing ignorance, and prefer to think of it solely in terms of the invariance property – it makes sense that a rule for objective inference should be invariant to monotone transformations.

In figure 3.1, Jeffreys’ prior for the binomial example is plotted from 0 to 1. At least the principle of indifference had some intuition behind it – it is hard to see that this function in a sense represents ignorance about the parameter $p$ since it goes to infinity near 0 and 1. I’ve seen some possible analogies in the literature that might help explain it. Consider a scientist having discovered some new chemical compound, and he is interested to see whether it is soluble in water. He takes a small amount of the compound, dilutes it in water, and observes that it does indeed dissolve. Based on this single observation, and due to the nature of the hard sciences, he should be pretty certain that the compound is in fact soluble in water. A flat prior would not nearly give as much posterior certainty to this statement as Jeffreys’ prior would, pushing the posterior towards one. In situations where the true $p_0$ is close to the edges of the parameter space, Jeffreys’ prior ensures that evidence in favour of such a hypothesis is given assistance.
Jeffreys’ prior is not used in multi-parameter settings, as it can have some pretty strong inconsistencies. Consider the following example, which is of the same flavour as example 2.3.

**Example 3.3** (*Neyman-Scott example*). Consider data $X_{ij} \sim N(\mu_j, \sigma^2)$ for $i = 1, 2$ and $j = 1, \ldots, m$. That is, we have pairs of data $(X_{1j}, X_{2j})$ from $m$ different populations with common variance, but different means. Suppose we are interested in estimating the common variance in the data, $\sigma^2$, and we wish to use Jeffreys’ prior to aid us.

The likelihood of a single pair from the model can be written as

$$L(\sigma^2, \mu_j) \propto \frac{1}{\sigma^2} \exp \left\{ -\frac{1}{\sigma^2} \left[ \frac{S_j^2}{4} + (\bar{x}_j - \mu_j)^2 \right] \right\},$$

where $S_j^2 = (x_{1j} - x_{2j})^2$ and $\bar{x}_j$ is the within-group mean. Since $\mu_j$ only crops up in the full likelihood through the above expression, the contribution to the Fisher information matrix in the $\mu_j$ term is by

$$-E \left[ \frac{\partial^2}{\partial \mu_j^2} \ell(\mu_j, \sigma^2) \right] = \frac{2}{\sigma^2},$$

and the cross terms are zero. The full likelihood can be written as

$$L(\mu_1, \ldots, \mu_m, \sigma^2) \propto \frac{1}{\sigma^{2m}} \prod_{j=1}^{m} \exp \left\{ -\frac{1}{\sigma^2} \left[ \frac{S_j^2}{4} + (\bar{x}_j - \mu_j)^2 \right] \right\},$$

and it is easy enough to derive the Fisher information matrix as

$$I(\mu_1, \ldots, \mu_m, \sigma^2) = \begin{bmatrix}
2/\sigma^2 & 0 & \cdots & 0 & 0 \\
0 & 2/\sigma^2 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 2/\sigma^2 & 0 \\
0 & 0 & \cdots & 0 & m/\sigma^4
\end{bmatrix},$$

which makes Jeffreys’ prior $\pi_j(\mu_1, \ldots, \mu_m, \sigma^2) \propto \sigma^{-(m+2)}$.

Now, the marginal posterior distribution of $\sigma^2$, found by integrating out all the $\mu_j$ from the posterior can be found to be

$$\pi(\sigma^2 | x) \propto (\sigma^2)^{-(m+1)} \exp \left\{ -\frac{1}{\sigma^2} \frac{S^2}{4} \right\},$$

where $S^2 = \sum_{j=1}^{m} S_j^2$. This distribution is in fact an inverse-gamma distribution with parameters $(m, S^2/4)$, with expectation $S^2/(4(m-1))$. The problem now is that, while $\hat{\sigma}^2 = S^2/(2(m-1))$ is a consistent estimator of the true $\sigma_0^2$, the marginal posterior distribution for $\sigma^2$ will concentrate around $\sigma_0/2$! Even more so for large $m$. \(\square\)

Since Jeffreys’ prior can lead to some very inconsistent results in higher dimensions, it is rarely used in these situations. The underlying issue is that the nuisance parameters in the model was not dealt with properly. We will see later that the reference prior approach by Berger and Bernardo is a nice extension of Jeffreys’ prior when nuisance parameters are present in the model, but to understand their foundations, we need a little detour.

\[^2\text{see for example Kim et al. (2006).}\]
3.4 E. T. Jaynes – the entropy path

The second line of argument in attempts to parametrize ignorance, is connected to the information theoretical concept of entropy. The leading figure behind this approach is E.T. Jaynes, who had a fascinating take on statistical inference. The following section is based on the exposition in Jaynes (2003, ch. 2,11,12) and is kept relatively loose. The main point is to provide some kind of information-theoretical foundation for the reference priors.

Jaynes was a proponent of the "probability as extended logic" school of thought. What this means is that he viewed the modern concept of probability as a direct extension of formal logic, thinking of statistical inference abstractly, as reasoning under incomplete information. He is interested in the underlying logic of this process, and how it can be made quantitative.

The classical way of deductive reasoning is to first state a causal connection between two propositions $A$ and $B$. We could say that if $A$ is true, then $B$ must also be true – or that $A$ implies $B$. From this statement, we can draw the following conclusions; If we observe $A$ to be true, then $B$ must also be true. Or, if we observe $B$ to be false, then surely $A$ too must be false. These kinds of deductions are usually too strong in practical science, where the causal relationships are not so straightforward. Instead, we use what may be deemed plausible reasoning.

Plausible reasoning abstractly takes the form of a much weaker syllogism;

If $A$ is true, then $B$ becomes more plausible.

$B$ is true, therefore, $A$ becomes more plausible.

In practical applications of statistics, we are often interested in proving that $A$ is true, on the basis of something that is easier to observe, $B$. Consider a doctor trying to figure out whether or not his patient has some kind of disease. The disease itself is not directly observable, but the anti-bodies that the immune system produces are observable in the patient’s blood. The presence of anti-bodies in the blood should make the doctor more certain that the patient has the disease, and the amount of anti-bodies in the blood should be directly connected to how certain the doctor should be. A tiny, almost undetectable, amount of antibodies could be ascribed to a measurement error, and the doctor should maybe run an extra test in a few days for verification. While, on the other hand, a large amount of anti-bodies in the blood should make the doctor quite confident that the patient has the disease in question.

Jaynes is interested in how these things should best be quantified, and find a unifying framework of how to do it. The reason why this is interesting for finding non-informative priors distributions, is that we need a measure of certainty before we can know how to be least certain.

Jaynes outlines three basic desiderata of a theory for representing certainty. First off, one needs to assume that degrees of plausibility can be represented by real numbers, in some way. Further, one would need that this representation is somehow in sync with ‘common sense’, i.e. that strong evidence leads to a stronger degree of certainty, as represented in real numbers. Lastly, we should like our theory to be consistent. By this, it is meant that if a conclusion can be worked out from different angles, every possible angle should lead to the same result.

\[3\]The statement "$B$ is true, therefore $A$ must be true", is a false one – famously demonstrated by Erasmus Montanus.
Working from the simplest case of a finite and discrete parameter space \{\theta_1, \ldots, \theta_n\}, Jaynes (2003, p. 347) constructs a measure of uncertainty using the above principles. First off, we need to assume that some numerical measure \(H_n(p_1, \ldots, p_n)\) exists that can assign degrees of certainty \(p_i\) to each possibility \(\theta_i\). It also seems natural to demand that such a measure is continuous in its arguments, \(p_i\), such that a small increase in \(p_i\) does not lead to a large leap in the overall uncertainty. Next, the measure should be in sync with ‘common sense’, which Jaynes bases on the principle of insufficient reason. He argues that we should be more uncertain when there are many possibilities to choose from, than when there are few, i.e. that the function

\[
h(n) = H_n \left( \frac{1}{n}, \ldots, \frac{1}{n} \right)
\]

is monotonically increasing in \(n\). Lastly, we should demand consistency of the measure. If there are many ways to compute the measure in a given situation, each way should yield the same answer.

After some technicalities, Jaynes finds that the only function \(H_n\) satisfying his conditions imposed on a reasonable measure of uncertainty, is the function

\[
H_n(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i, \quad (3.3)
\]

which is the Shannon entropy.

The Shannon entropy dates back to Shannon (1948), and his works on information theory. Originally, it was meant as a measure of the information content in English text, useful for data compression schemes. It has many interpretations, but the one that is most useful for our purpose is that it can be considered as a measure of unpredictability. Assume a data sequence is generated from some stochastic process, \(f\) taking values in the set \{\theta_1, \ldots, \theta_n\}. If the corresponding probabilities are highly skewed, say that \(p_1 = 0.9\) while \(p_2, \ldots, p_n = 0.1/(n-1)\), this is a sequence that is highly predictable. Constantly guessing that the next value will be \(\theta_1\), is on average a very good guess – this sequence has a low entropy, or high predictability. If instead the probabilities assigned where according to the principle of insufficient reason, guessing the next value in the sequence would be hard, since they are all equally probable – this sequence will have high entropy. In cryptography, one of the main desiderata is that the cipher utilized has a uniform distribution over the language, i.e. the alphabet A-Z, for English text – making it difficult to predict.

Jaynes thus argues that the only non-informative prior distribution we should assign in the case of a finite and discrete parameter space, is the one that maximizes the entropy – namely the principle of insufficient reason.

In the case where the parameter space is continuous, there are some difficulties that must be overcome. A natural extension of the discrete entropy in (3.3) would be to simply move from summation to integration and define

\[
H' = - \int_{\Theta} p(\theta) \log p(\theta) d\theta \quad (3.4)
\]

which is the so-called differential entropy. The problem with this is that it is not invariant under monotone transformation, so that the measure of uncertainty will
be different depending on the parametrization. A property that we’ve seen should really be demanded of a good prior distribution. Another, slightly more worrying aspect of this definition, is that it is not the result of some well defined limiting process.

Jaynes argues that passing from discrete parameter spaces to continuous should only be done in cases where the continuous spaces can be derived through a natural limiting process. Starting from a finite discrete set of points \( \{ \theta_1, \ldots, \theta_n \} \), and the corresponding Shannon entropy \( H_n \), letting \( n \to \infty \) such that the parameter space becomes denser, he finds that the continuous measure of uncertainty becomes

\[
H^c = -\int_{\Theta} p(\theta) \log \frac{p(\theta)}{m(\theta)} \, d\theta,
\]

for some unknown reference measure \( m(\theta) \). Keeping with the technique in the discrete case, we should choose our prior distribution \( p(\theta) \) in such a way that it maximizes the quantity (3.5), which will turn out to be when \( p(\theta) \propto m(\theta) \). But the problem still remains, which reference measure \( m \) should we choose in the above expression?

### 3.5 Reference Priors

#### 3.5.1 Motivation and Definition

Reference priors were first introduced in 1979 by [Bernardo] and uses information-theoretical arguments to derive a prior that in some sense is least-informative, to be used as a reference, or baseline, when doing Bayesian inference. This is very much in line with the thoughts of E.T. Jaynes from the past section, and we will see that the two positions are highly connected.

Since the posterior distribution is a compromise between prior information and the observed data, the main idea is that we should choose a prior distribution that influence the posterior as little as possible, while still facilitating a proper posterior. One measure of this influence is the Kullback-Leibler divergence.

**Definition 3.1 (Kullback-Leibler divergence).** The Kullback-Leibler divergence from the prior distribution \( \pi(\theta) \) to the posterior \( p(\theta|x) \) is defined by:

\[
\kappa(p|\pi) = \int_{\Theta} p(\theta|x) \log \frac{p(\theta|x)}{\pi(\theta)} \, d\theta,
\]

where \( \log \) is the natural logarithm. In the discrete case, integrals are summations.

Loosely stated, the KL-divergence from \( q \) to \( p \) can be viewed as the distance between the two distributions, but there are several interpretations. In this section, we will view it as the information gain from prior to posterior distribution. It gives a measure of how much information is obtained when updating our beliefs from the prior distribution to the posterior distribution. A sharp, localized prior distribution will give less weight to the data, resulting in a smaller information gain from observing the data. Intuitively we should want a prior distribution that, when updated with information, moves as much as possible – one that gains as much information as possible from the data.
CHAPTER 3. OBJECTIVE BAYESIAN INFERENCE

Figure 3.2: The KL-divergence of the flat and Jeffreys’ priors, for a range of observed values X from a binomial model, with n = 100.

An important property of the KL-divergence is that it is always non-negative and equal to zero if and only if q = p almost everywhere – that is, if the posterior distribution equals the prior, i.e. that no new information was obtained from the new data.

Notice the direct similarity to Jaynes’ continuous measure of uncertainty (3.5), which is maximized when p(θ) = m(θ) almost everywhere. For Jaynes, this implied that m(θ) was in fact the prior distribution that represented "complete ignorance" (Jaynes 2003, p. 377). The reference prior approach is not looking for the one that represents "complete ignorance", but rather one that gets the closest.

It might be interesting to look at how the KL-divergence actually ranks the different priors distributions. In figure 3.2, the KL-divergence from prior to posterior distribution in the binomial setting is visualized. The setup is a binomial experiment with n = 100, and the KL-divergence found, by numerical integration, for both the flat prior, and Jeffreys’ prior on p for different observed values X ranging from zero to 100.

We see that the KL-divergence from prior to posterior is heavily dependent on the observed data, with Jeffreys’ prior having the largest divergence for most of the sample space, while the flat prior "outperforms" Jeffreys’ at the edges of the sample space. This makes sense, since Jeffreys’ prior already puts a lot of weight there (a large X is more likely if p0 is large), and having data confirm it simply means that the posterior distribution don’t change as much as it would if the prior was flat.

If the true underlying p0 in an experiment is somewhere around 0.5 is is easy to see that Jeffreys’ prior would give more weight to the data on average, say if the example was repeated many times. To formalize this thought, consider removing the dependency on the observed data by averaging the KL-divergence over the sample space. Then, we obtain what is called the expected information gain.

---

4It is also relatively uniform, which looks to me as a general trend with Jeffreys’ prior when looking at the KL-divergence.

5at this point, we are breaking the likelihood principle, which some orthodox Bayesians might object to.
Definition 3.2 (Expected information gain). Let $x$ be an observation from the model $\mathcal{M} := \{p(x|\theta), x \in \mathcal{X}, \theta \in \Theta\}$. When the prior distribution is $\pi(\theta)$, the expected information gain from one observation is

$$ I(\pi|x) = \int_{\mathcal{X}} p(x) \kappa(p|\pi) \, dx $$

$$ = \int_{\mathcal{X}} \int_{\Theta} p(x|\theta) \pi(\theta) \log \frac{p(\theta|x)}{\pi(\theta)} \, d\theta \, dx, $$

since $p(\theta|x) = p(x|\theta) \pi(\theta)/p(x)$. In the discrete case, integrals are summations. □

Choosing the prior that will maximize the expected information gain from a single experiment intuitively seems like a good idea, but we should want something much stronger. A reference prior should be a baseline for the entire model $\mathcal{M}$, not just on the basis of what we can expect to learn about $\theta$ from a single experiment. We should aim for a setting where the reference prior maximize the information gain from the prior to what is theoretically possible to learn from the model, perhaps by repeating the experiment many times.

Consider $k$ independent hypothetical replications of the experiment at hand, leading to a sequence of realizations $x^k = (x_1, \ldots, x_k)$ and a measure of information gain based on this entire vector, $I(\pi|x^k)$. As the number of replications increase, the experiment will eventually provide the all missing information about $\theta$, until it is completely known. The quantity $I(\pi|x^k)$ then, will, as $k \to \infty$, begin to represent the amount of missing information about $\theta$ contained in the prior, $\pi$.

This requires some clarification. Say that we are comparing two prior distributions $\pi_1$ and $\pi_2$. Suppose also that the two quantities $I(\pi_1|x^k)$ and $I(\pi_2|x^k)$ both converge to some real numbers, $N$ and $M$ respectively, as $k \to \infty$. Then, if $N > M$, this means that the expected amount of information needed to know the value of $\theta$ with certainty, is larger when starting with $\pi_1$ than compared to when starting from $\pi_2$. This then means that $\pi_1$ is less informative about $\theta$ than what $\pi_2$ is, and the quantities $N$ and $M$ in a sense also represents the amount of missing information associated with each prior.

Now then, a reference prior is, loosely stated, a permissible prior distribution that maximizes the missing information within a class of candidate priors, $\mathcal{P}$. This class of candidate priors is one that agrees with any a priori knowledge the researcher might have, such as natural parameter boundaries, continuity or non-negativity.

This is all incredibly abstract, and there are several concerns that should come to mind when reading it. The biggest concern should be if the missing information even exists, I mean, is the limit $\lim_{k \to \infty} I(\pi|x^k)$ even defined? Or what about the expected information gain, what is stopping the integrals in definition 3.2 from being infinite? Nothing we’ve done so far has put any restrictions in place that makes the above equations and definitions hold.

Let’s start with the one-parameter case, for which the theory is on a pretty solid mathematical foundation, and follow the way it is laid out in [Berger et al.] (2009). For the definitions to work, we actually need a stricter notion of permissible prior distributions – it is not enough to demand that they lead to a proper posterior distribution, we need the reference prior to be the result of a well-defined limiting process, again in line with Jaynes. Consider the following definition;
Definition 3.3 (Expected logarithmic convergence of posteriors). Let \( \{\Theta_i\}_{i=1}^{\infty} \) be an increasing sequence of compact subsets of \( \Theta \) such that \( \bigcup_{i=1}^{\infty} \Theta_i = \Theta \). Let \( \{\pi_i(\theta|x)\}_{i=1}^{\infty} \) denote the corresponding sequence of posteriors, i.e.
\[
\pi_i(\theta|x) \propto \int_{\Theta_i} p(x|\theta) \pi_i(\theta) d\theta
\]
where \( \pi_i(\theta) \) is the natural restriction of the prior to the subset \( \Theta_i \). Then the sequence of posteriors is expected logarithmically convergent to the formal posterior \( \pi(\theta|x) \) if
\[
\lim_{i \to \infty} \int_{\mathcal{X}} \kappa (\pi_i(\cdot|x)\pi_i(\cdot|x)) p_i(x) dx = 0,
\]
where \( p_i(x) = \int_{\Theta_i} p(x|\theta) \pi_i(\theta) d\theta \).

A stronger notion of permissible is then that, (i) the prior should yield a proper posterior distribution, and that (ii), for some approximating compact sequence, the corresponding posterior distributions are expected logarithmically convergent. That is, the prior distribution itself is derived implicitly, through a limiting sequence of posteriors.

Next, we need to tackle the question raised earlier, about the existence of the expected information gain, and if it really converges to anything. Berger et al. (2009) notes that a bounded parameter space is typically enough to ensure existence of \( I(\pi|x^k) \) for a given \( k \), but if the parameter space is continuous, it will typically diverge as \( k \to \infty \) – would it not take an infinite amount of information to know the value of a real number with certainty? The following definition deals with these issues,

Definition 3.4 (Maximizing missing information (MMI) property). Let \( \mathcal{M} \) be a model with one continuous parameter, and let \( \mathcal{P} \) be the class of prior distributions yielding a proper posterior. The function \( \pi(\theta) \) is said to have the MMI property for model \( \mathcal{M} \) and prior class \( \mathcal{P} \) if, for any compact subset \( \Theta_0 \subset \Theta \) and any \( p \in \mathcal{P} \),
\[
\lim_{k \to \infty} \left\{ I(\pi_0|x^k) - I(p_0|x^k) \right\} \geq 0,
\]
where \( \pi_0 \) and \( p_0 \) are renormalized restrictions of \( \pi \) and \( p \) to \( \Theta_0 \).

The fact that \( I(\pi|x^k) \) can diverge as \( k \to \infty \) is taken care of, simply by insisting that it will be larger than the missing information of any other qualified prior \( p \) (which also diverges).

Then the formal definition of a reference prior in the one-dimensional case is the following;

Definition 3.5 (Reference Prior). A function \( \pi_R(\theta) = \pi(\theta|M, \mathcal{P}) \) is a reference prior for model \( \mathcal{M} \), given a class of potential priors \( \mathcal{P} \), if it is permissible, in the stronger sense, and has the MMI property.

This definition gives no real clue as to how the problem at hand should be solved, i.e. finding the function \( \pi(\theta) \) that maximizes the missing information about \( \theta \). But there are explicit forms or algorithms available for finding them.
3.5. REFERENCE PRIORS

3.5.2 Explicit Forms of the Reference Prior

Finding the explicit form of the reference prior in a given model is not always an easy task, and the definition does not offer many hints on how to go about the problem. To motivate the formal solution to this problem consider the following reformulation of the expected information gain under conditions regular enough to allow the operations:

\[
I(\pi|x_k) = \int_{X^k} \int_{\Theta} p(x^k|\theta) \pi(\theta) \log \frac{p(\theta|x^k)}{\pi(\theta)} \, d\theta \, dx^k
\]

\[
= \int_{\Theta} \pi(\theta) \int_{X^k} p(x^k|\theta) \log \frac{p(\theta|x^k)}{\pi(\theta)} \, dx^k \, d\theta
\]

\[
= \int_{\Theta} \pi(\theta) \log \frac{f_k(\theta)}{\pi(\theta)} \, d\theta
\]

\[
= -\int_{\Theta} \pi(\theta) \log \frac{\pi(\theta)}{f_k(\theta)} \, d\theta,
\]

where

\[
f_k(\theta) = \exp \left\{ \int_{X^k} p(x^k|\theta) \log p(\theta|x^k) \, dx^k \right\}.
\]

Equation 3.7 is in fact \(-\kappa(\pi|f_k)\), the (negative) KL-divergence from \(f_k(\theta)\) to \(\pi(\theta)\). And since the KL-divergence is always greater than or equal to zero, with equality if and only if \(\pi(\theta) = f_k(\theta)\) almost everywhere, this equation is maximized when \(\pi(\theta) \propto f_k(\theta)\). Typically though, as we’ve discussed, in continuous parameter cases, or with an unbounded parameter space, the expected information gain either isn’t defined, or it diverges as \(k \to \infty\). The following theorem, which is proved in Berger et al. (2009), circumvents these issues for the one-parameter case.

**Theorem 3.1 (Explicit form of the reference prior).** Assume a standard model \(\mathcal{M} := \{p(x|\theta), x \in X, \theta \in \Theta \subset \mathbb{R}\}\) and a class \(\mathcal{P}\) of continuous priors on \(\Theta\) that lead to a proper posterior distribution. Let \(\pi^*(\theta)\) be a strictly positive continuous prior such that the corresponding posterior, \(\pi^*(\theta|x^k)\), is proper and asymptotically consistent, i.e. concentrating around the true value \(\theta\) as \(k \to \infty\). Define, for any interior point \(\theta_0\) of \(\Theta\),

\[
f_k(\theta) = \exp \left\{ \int_{X^k} p(x^k|\theta) \log [\pi^*(\theta|x^k)] \, dx^k \right\}
\]

and

\[
f(\theta) = \lim_{k \to \infty} f_k(\theta_0).
\]

If (i) each \(f_k(\theta)\) is continuous and, for any fixed \(\theta\) and sufficiently large \(k\), \([f_k(\theta)/f_k(\theta_0)]\) is either monotonic in \(k\) or bounded above by some \(h(\theta)\) integrable on any compact set, and (ii) \(f(\theta)\) is a permissible prior function, then \(\pi_R(\theta) \propto f(\theta)\) is a reference prior for model \(\mathcal{M}\) and prior class \(\mathcal{P}\). \(\square\)

---

6A bounded parameter space will ensure the existence of \(I(\pi|x)\) and further conditions are needed to apply Fubini’s theorem.

7It also provides an answer to Jaynes question of which reference measure \(m(\theta)\) to use in equation 3.5.
CHAPTER 3. OBJECTIVE BAYESIAN INFERENCE

The choice of \( \pi^*(\theta) \) and \( \theta_0 \) in the theorem is arbitrary, and may be chosen to make the computations easier. In addition, the theorem still holds, and yields the same reference prior, when computed on the basis of a sufficient statistic \( t^k = t(x^k) \), instead of the full data vector \( x^k \), see theorem 5 in Bernardo (2005).

The theorem is a general recipe for constructing reference priors in the one-parameter case, and it puts the theory of reference priors on a solid mathematical foundation. However, the theorem is rarely used in this form, since under regularity conditions which often are satisfied, there are simpler closed form expressions available.

Consider the following example with a rather loose argumentation:

Example 3.4 (A difficult path to Jeffreys’ prior). Let’s assume that we have an experiment set up, the result of which, is a binomial distributed random variable \( X \), with known \( n \) and unknown \( p \), i.e. \( X \sim \text{Bin}(n, p) \). We wish to derive the reference prior for the parameter \( p \) in this setup. To do this, we assume that we have \( k \) iid. replications of the experiment, \( X_1, \ldots, X_k \) available to us, and go through the calculations outlined in the preceding paragraphs.

A natural sufficient statistic to consider is \( T^k = \sum_{i=1}^{k} X_i \) whose distribution is \( \text{Bin}(kn, p) \). The expression from equation 3.9 becomes

\[
f_k(p) = \exp \left\{ \sum_{T^k} \left( \frac{nK}{t_k} \right) p^{t_k} (1-p)^{nk-t_k} \log \left[ \pi^* (p|t_k) \right] \right\}. \tag{3.11}
\]

Now, assuming the true value \( p_0 \) is not at the edge of the parameter space, and that the prior \( \pi^*(p) \) is nice, this model is regular enough so that asymptotic posterior normality is obtained for \( \pi^*(p|t^k) \). That is, by the Bernstein-von Mises theorem, for large \( k \), \( \pi^*(p|t^k) \approx N(\hat{p}, I_k^{-1}(\hat{p})) \) where \( \hat{p} \) is a consistent estimator of \( p \), and \( I_k(\hat{p}) \) is the Fisher information evaluated at \( \hat{p} \). By independence and consistency, the density of this distribution, for large \( k \), evaluated at the true value \( p_0 \), can be written,

\[
\pi^*(p_0|t^k) \approx (nk)^{1/2} I^{1/2}(p_0) = (nk)^{1/2} (p_0(1-p_0))^{-1/2},
\]

which makes equation (3.11):

\[
f_k(p) \approx \exp \left\{ \log \left[ (nk)^{1/2} (p(1-p))^{-1/2} \right] \right\} \propto \frac{1}{\sqrt{p(1-p)}},
\]

when switching back to \( p \) instead of \( p_0 \), which is in fact Jeffreys’ prior for this problem. \( \square \)

The above example has very loose argumentation, where several unjustified approximations are made, but it is indicative of something greater, which is true in wide generality, and holds not only in the binomial case above; in continuous one-parameter cases, under conditions that ensure asymptotic posterior normality, the reference prior equals Jeffreys’ prior. A proof of this statement can be found in, for example, Bernardo (2005).

That the reference prior equals Jeffreys’ prior in the one-dimensional case means that Jeffreys’ prior can also be placed on solid entropy-based footing, and not simply justified through invariance.

\[\text{I think it is interesting to have seen the theorem, so that when we use some of the other algorithms for finding reference priors later, the notion of deriving the prior through a limit isn’t so foreign.} \]
3.5. REFERENCE PRIORS

Properties

The reference prior has many desirable properties that one should hope a non-informative prior has. Most desirably, since it agrees with Jeffreys’ prior, it is invariant under monotone transformations in the one-dimensional case. In the multi-parameter case, the reference prior no longer equal Jeffreys’, which we will see later, and this property is lost. For multi-parameter models, the reference prior is invariant only under certain transformations, see Datta & Ghosh (1996). This also entails that the reference prior will depend on the parameter of interest! The reference prior for \((\mu, \sigma^2)\) may be completely different from the reference prior for \((\sigma/\mu, \sigma)\).

The reference prior is independent of the sample size in the model. From a Fisher-information viewpoint this is because \(I_n(\theta) = nI(\theta)\), but note that it may depend on the sampling design in cases where \(n\) cannot be pulled out of the information matrix – we will see an example of this situation later. From a Bayesian viewpoint, the dependence on sampling design might seem strange, but it stems from the "averaging over the sample space" we did in definition 3.2. Again, we are violating the likelihood principle here.

A last important property of the reference priors is that, since it is often a part of the construction to approximate the full parameter space \(\Theta\) with an expanding sequence of compact subsets \(\{\Theta_i\}_{i=1}^{\infty}\); the reference prior of the full space \(\Theta\), restricted to a subset \(\Theta_i\) will also be the reference prior found by only working on \(\Theta_i\) from the beginning.

Nuisance parameters

When there are nuisance parameters present in our model, we’ve seen that Jeffreys’ prior can provide less than optimal inference. The reference prior paradigm provides a scheme for dealing with nuisance parameters by ordering parameters by their inferential importance, and using a stepwise-conditioning argument. When there is only a single nuisance parameter present in the model, the algorithm takes on a simple form.

Consider the problem where \(\mathcal{M} := \{x \in X, \psi \in \Psi \subset \mathbb{R}, \lambda \in \Lambda \subset \mathbb{R}\}\) where \(\psi\) is the parameter of interest, and \(\lambda\) is a nuisance parameter. In this case, one would need a joint reference prior for the set of parameters \((\psi, \lambda)\), to obtain a joint posterior for the pair, and then a marginal posterior for the parameter of interest, \(\psi\). Bernardo (1979) suggested the following scheme to produce the joint prior distribution for \((\psi, \lambda)\).

**Step 1** Consider \(\psi\) fixed and derive the one-dimensional conditional reference prior \(\pi_R(\lambda|\psi)\). In regular cases, this amounts to computing Jeffreys’ prior.

**Step 2** If the the conditional prior \(\pi_R(\lambda|\psi)\) is proper, i.e. \(\int_{\Lambda} \pi_R(\lambda|\psi) d\lambda < \infty\), the nuisance parameter may be integrated out to find the one-parameter integrated model \(p(x|\psi) = \int_{\Lambda} p(x|\psi, \lambda) \pi(\lambda|\psi) d\lambda\).

**Step 3** Now, derive the one-dimensional reference prior based on \(p(x|\psi)\), \(\pi_R(\psi)\). In regular cases this again amounts to Jeffreys’ prior.

**Step 4** Finally, the joint reference prior is by \(\pi_R(\psi, \lambda) = \pi_R(\psi) \pi_R(\lambda|\psi)\).
This is a nice scheme, but often it does not work out as intended, the soft spot is in step 2, where the conditional Jeffreys’ prior often isn’t proper. Berger & Bernardo (1989) produce a more general scheme for the one-dimensional nuisance parameter setup. Let $I(\psi, \lambda)$ denote the Fisher information matrix of this model, and denote the parameter space by $\Theta = \Psi \times \Lambda$. Step 1 is exactly as before, and typically amounts to finding Jeffreys’ prior for $\lambda$, while keeping $\psi$ fixed. The rest of the procedure continues as

**Step 2** Choose an expanding sequence of compact subsets of $\Theta$, $\{\Theta_i\}_{i=1}^{\infty}$ such that $\bigcup_i \Theta_i = \Theta$ and $\pi(\lambda|\psi)$ is integrable on the slice $\Omega_i,\psi = \{\lambda : (\psi, \lambda) \in \Theta_i\}$ for all $\psi$. Then normalize on each slice to obtain

$$
\pi_i(\lambda|\psi) = K_i(\psi)\pi(\lambda|\psi)\mathbb{I}_{\Omega_i,\psi}(\lambda)
$$

where

$$
K_i^{-1}(\psi) = \int_{\Omega_i,\psi} \pi(\lambda|\psi)d\lambda
$$

**Step 3** Compute

$$
\pi_i(\psi) = \exp \left\{ \frac{1}{2} \int_{\Omega_i,\psi} \pi_i(\lambda|\psi) \log \frac{|I(\psi, \lambda)|}{|I(\psi, \lambda)_{22}|} d\lambda \right\},
$$

where $I$ denotes the full Fisher information matrix, and $I_{22}$ denotes the second diagonal element.

**Step 4** Obtain the joint reference prior as:

$$
\pi(\psi, \lambda) = \lim_{i \to \infty} \frac{K_i(\psi)\pi_i(\psi)}{K_i(\psi_0)\pi_i(\psi_0)}\pi(\lambda|\psi)
$$

where $\psi_0$ is any fixed, interior point of $\Psi$.

We will use this algorithm later to derive a reference prior in a relatively simple setup.

This algorithm is complex, and the integrals one needs to solve often take some ugly forms. Luckily, since the nuisance parameter is integrated out before finding the reference prior for $\psi$, the choice of nuisance parameter won’t affect $\pi_R(\psi)$. This means that we can often reparametrize the problem into one that is easier to work with.

If we have multiple nuisance parameters in our model, we must order the parameters in the model after inferential importance, and apply the above algorithm sequentially. Say the model has parameters $\{\psi, \lambda\}$ where $\psi$ is a parameter of interest, while $\lambda$ is a vector of nuisance parameters, an ordering might be $\{\psi, \lambda_1, \ldots, \lambda_n\}$ where $\psi$ and $\lambda_n$ are most and least important, respectively. Then, we hold $\{\psi, \lambda_1, \ldots, \lambda_{n-1}\}$ fixed and derive the reference prior for $\lambda_n$, $\pi_R(\lambda_n|\psi, \lambda_1, \ldots, \lambda_{n-1})$. We then integrate this parameter out of the model and continue, now holding $\{\psi, \lambda_1, \ldots, \lambda_{n-2}\}$ fixed and finding the reference prior $\pi_R(\lambda_{n-1}|\psi, \lambda_1, \ldots, \lambda_{n-2})$. Doing this sequentially, we construct the final reference prior as

$$
\pi_R(\psi, \lambda_1, \ldots, \lambda_n) \propto \pi_R(\psi)\pi_R(\lambda_1|\psi)\cdots\pi_R(\lambda_n|\psi, \lambda_1, \ldots, \lambda_{n-1})
$$

Luckily, several shortcuts have been developed for dealing with nuisance parameters in the models so that we don’t actually have to solve the reference prior in this manner every time.
3.5. REFERENCE PRIORS

3.5.3 Shortcuts to a reference prior

Over the years, numerous shortcut formulas have been derived to make the reference prior approach easier in practice. Many of these shortcuts are possible if the Fisher information matrix of the model is on a certain form, and that the model in question is asymptotically normal. In this section, two such shortcut theorems are stated.

The first theorem can be used if the parametrization is orthogonal in the sense of Cox & Reid (1987), i.e. that the Fisher information is diagonal. It was first proved by Datta & Ghosh (1995b).

Suppose that the parameters in the model, \( \theta = (\theta_1, \ldots, \theta_p)^T \) can be ordered into \( m \) groups \( \theta = (\theta^{(1)}, \ldots, \theta^{(m)}) \) of descending inferential interest, i.e. that \( \theta^{(1)} \) is more important than \( \theta^{(2)} \) and so on. Within the grouping, the parameters are considered of equal inferential importance. Assume that (1) the Fisher information matrix \( I(\theta) \) is block diagonal, with elements \( \{h_1(\theta), \ldots, h_m(\theta)\} \). Let \( \theta_j^C = \{\theta^{(1)}, \ldots, \theta^{(j-1)}, \theta^{(j+1)}, \ldots, \theta^{(m)}\} \), and assume (2) that \( \det h_j(\theta) = h_{j1}(\theta_j^C)h_{j2}(\theta_j^C) \) for non-negative functions \( h_{j1} \) and \( h_{j2} \).

Finally (3), choose the sequence of rectangular compact subsets to be

\[ \Theta_i = A_1^i \times \cdots \times A_m^i, \]

where each \( A_j^i = \{\theta_j : \theta_j \in A_j^i\} \).

Then, we have the following theorem.

**Theorem 3.2 (Reference prior shortcut I).** Assume (1)-(3), then the reference prior is simply

\[ \pi^R(\theta) \propto m \prod_{j=1}^m h_{j1}(\theta_j)^{1/2}. \]

(3.16)

This is a really powerful theorem when there are multiple nuisance parameters present in the model. The difficulty often lies in tracking down a block diagonal parametrization for the parameters in the model, and in choosing how the parameters should be grouped together. The reference prior is very sensitive to how the order of inferential importance is chosen, and we must check in each case which posterior performs the best. We will see a clear example of this in section 4.3.

**Example 3.5 (Neyman-Scott continued).** Consider again the setup in example 3.3. The Fisher-information matrix for this problem is diagonal, and the natural grouping of the parameters would be \( \{\sigma^2, \mu_1, \ldots, \mu_n\} \), using the above theorem then directly yields the reference prior as \( \pi^R(\sigma^2, \mu_1, \ldots, \mu_n) \propto \sigma^{-2} \). The marginal posterior distribution of this distribution becomes \( \text{Inv-Ga}(m/2, S^2/4) \) which concentrates around \( S^2/(2(m-2)) \) as the number of groups \( m \) becomes large. But this statistic is actually a consistent estimator of the true underlying variance \( \sigma_0^2 \) so that the posterior distribution concentrates around \( \sigma_0^2 \) as \( m \to \infty \)! In the setup where Jeffreys' prior failed due to the amount of nuisance parameters in the model, the reference prior approach works fine!

The second theorem is due to Bernardo (2005), and doesn’t require diagonality of the Fisher information, but it does have some other restrictions. I will only state the
Theorem for the two-dimensional case with an interest parameter $\psi$, and a nuisance parameter $\lambda$. The ‘one-nuisance parameter case’ is the only case we will use later, but it extends easily to the multi-parameter case.

Let the Fisher information matrix, and its inverse be written as

$$ I(\psi, \lambda) \begin{bmatrix} h_{\psi,\psi} & h_{\psi,\lambda} \\ h_{\lambda,\psi} & h_{\lambda,\lambda} \end{bmatrix} \quad I^{-1}(\psi, \lambda) = \begin{bmatrix} v_{\psi,\psi} & v_{\psi,\lambda} \\ v_{\lambda,\psi} & v_{\lambda,\lambda} \end{bmatrix}. \quad (3.17) $$

Then, we have the following theorem;

**Theorem 3.3** (Reference prior shortcut II). If the functions $v^{-1/2}_{\psi,\psi}$ and $h^{1/2}_{\lambda,\lambda}$ in the matrices $(3.17)$ can be factorized as

$$ v^{-1/2}_{\psi,\psi} \propto f_{\psi}(\psi) g_{\psi}(\lambda) \quad h^{1/2}_{\lambda,\lambda} \propto f_{\lambda}(\psi) g_{\lambda}(\lambda), $$

and the parameters are variationally independent, i.e. that $\Lambda_i$ does not depend on $\psi$. Then, the reference prior is simply

$$ \pi_R(\psi) \propto f_{\psi}(\psi) g_{\lambda}(\lambda). \quad (3.18) $$

The "variationally independent" part might seem a bit strange. In the algorithm laid out for finding reference prior in the previous section, we required some integrals to be taken over the entire parameter space $\Lambda$, or over compact subsets $\Lambda_i$. This requirement simply states that the area of integration we are working on when doing these integrals cannot depend on the parameter of interest.

There are many other shortcuts that exist, but these are the only two we will make use of later, in addition to the algorithm from the previous section.

### 3.5.4 On compact subsets and uniqueness

During the theoretic outline of the reference prior, we’ve often relied on an approximating sequence of compact subsets to make the mathematics go the way we want. Specifically we needed them to work with improper prior distributions, that can only be normalized only smaller subsets. But they also serve another purpose, one that is slightly more subtle. In all real applications of statistical theory, the parameter space is always bounded in some way. By this I mean that parameters often have a physical interpretation, and that this physicality impose some restrictions on how large or small their numerical values can be. The actual boundaries may be pretty large, and so it seems natural to use an unbounded set as an approximation to the unknown large, bounded space we are actually working inside of. This is in line with Bernardo’s own views

I believe one should always consider a probability model endowed with an appropriate compact approximation to its parameter space, which should then be kept fixed, via the appropriate transformations, for all inference problems considered within that model. \cite{Bernardo1997} p. 187

The question then becomes, how should we choose this sequence of subsets? Is there a default choice, or does it even matter? It turns out that it can in fact matter,
and that the reference prior might actually depend on how the sequence was chosen. This is demonstrated in for example [Berger & Bernardo (1989)] and [Sun & Ye (1995)] when deriving the reference prior for the product of normal means. Depending on how the sequence is chosen, the reference prior can change quite a bit, which is slightly unsettling. Berger and Bernardo provide the following pragmatic view:

Choosing the \( \Theta_i \) to be natural sets in the original parametrization has always worked well in our experience. Indeed, the way we think of the \( \Theta_i \) is that there is some large compact set on which we are really noninformative, but we are unable to specify the size of this set. We might, however, be able to specify a shape, \( \Omega \), for this set, and would then choose \( \Theta_i = i\Omega \cap \Theta \), where \( i\Omega \) consists of all points in \( \Omega \) multiplied by \( i \). (Bernardo & Berger 1992, p. 42)

Taking this literally, it means we need to think about which shapes in our parameter space it is more natural to be non-informative over. Being pragmatic, in cases where multiple reference priors are available, we should evaluate their performance, perhaps by some frequentist criteria.

### 3.6 Frequentist properties of the Bayesian posterior

When Bayes theorem is employed with a subjective prior distribution, the resulting posterior distribution has a clear, tangible meaning. Your prior knowledge, as it is expressed in the prior distribution is updated through Bayes theorem, to yield Your current state of knowledge. And the resulting posterior distributions can be used to make definite (epistemic) probability statements about the value of the parameters. But when the prior distribution is not chosen to represent the researchers state of knowledge before the experiment, what should we make of the posterior distribution? Does it even make sense to do inference in this way?

As the clear logic of subjective Bayesian analysis is unavailable for us, one might want to consider some more frequentist properties of the posterior distributions – though most objective Bayesians still interpret their posterior distributions the same way as subjective Bayesians. Sound frequentist properties would also ensure that the Bayesian procedure does well on average, something that can be valuable to both subjective and objective Bayesians.

One way to analyze the frequentist properties of the Bayesian posterior is to look at the coverage of posterior sets. If \( K_\alpha \subset \Theta \) is a region with posterior probability \( \alpha \), i.e. \( \int_{K_\alpha} p(\theta|X) \, d\theta = \alpha \) then what is \( P(\theta_0 \in K_\alpha) \) in the frequentist sense?

In a few select cases, we can get exact matching, i.e. that any \( \alpha \) level credible sets have exactly \( \alpha \)-level coverage probability. This amounts to the (marginal) Bayesian posterior distribution being an exact confidence distribution. Situations where this happens is studied in section [4.2.1].

Outside of the limited frame where Bayesian posterior distributions are also confidence distributions, one might ask how close to the intended coverage can we get? Keeping with the terminology from higher order asymptotics, we will say that a prior distribution is \( k \)th-order matching if

\[
P(\theta_0 \in K_\alpha) = \alpha + O\left(n^{-k/2}\right),
\]
where $O(g(n))$ denotes terms that are bounded by the function $g(n)$.

There are many routes to go here, and the theory is well developed. But let’s initially focus on the coverage probability of credible quantiles, i.e. letting $K_\alpha$ denote the $\alpha$-level posterior quantile of $\theta$. The classical reference in this direction is Welch & Peers (1963) who proved several interesting cases. Notably, quite generally, for any continuous prior distribution, $K_\alpha$ is a first-order matching set. That is,

$$P(\theta_0 \leq K_\alpha) = \alpha + O\left(n^{-1/2}\right),$$

which, in a sense is comforting. When we have enough data, asymptotically our credible intervals also becomes confidence intervals, which is also a consequence of the Bernstein-von Mises theorem.

The main result in the paper, is that second-order matching can be attained if, and only if, the prior distribution utilized is Jeffreys’ prior. And further, third-order matching can be obtained by Jeffreys’ prior, only in special cases. Again, we see that Jeffreys’ prior, initially suggested based on its invariance properties, has some really nice properties!

3.6.1 Probability matching priors

As a last side-note on the theory of objective Bayesian inference, note that there are approaches to prior construction that aim solely for good frequentist coverage. These are called Probability matching priors (PMPs), and, while being developed by many authors, the go-to reference seems to be Datta & Mukerjee (2004).

The results in the previous section concerned itself only with one-sided posterior quantiles. But it turns out that stronger results can be obtained if we consider other credible regions, such as highest posterior regions, or if we are only interested in frequentist properties for a single, scalar, parameter of interest, $\psi$.

These priors often have excellent frequentist properties, even more so than the reference prior. It seems natural that a prior distribution, developed solely on the basis of frequentist criteria, would be an ideal candidate for a non-informative prior. There is literally no prior knowledge that goes into the construction, except the same model and sampling considerations that went into the reference priors.

It also seems that these priors would make an ideal candidate for comparison with the confidence distributions of the previous chapter. Some of the reference priors we are considering in the following chapter are actually first- or second-order probability matching priors; particularly those that come from theorem 3.2. But it has not been the focus of the thesis to pursue these ideas.
Chapter 4
Comparisons and Examples

In this chapter, some examples are worked through using the theory in the previous chapters. The examples are not intended to be interesting simply on their own merit, but are used to highlight various practical and theoretical aspects of the various theories. Furthermore, as an objective goal, to compare the optimal confidence distributions, viewed somewhat as a gold standard of what can be learned from data, to the objective Bayesian theories.

4.1 Difference of exponential parameters

The purpose of this example is to perform a pretty standard objective Bayesian analysis, deriving objective priors and analyzing their frequentist properties. In addition we compare posterior distributions to the optimal CDs, trying to shed some light on how ‘objective’ our posteriors really are. The example illustrates the difficulties that one encounters trying to derive a reference prior, and contrasts it to the quick and easy CD approach. Lastly, we look at boundary cases where the frequentists have it easy, but the Bayesians might have it hard.

Let \( Expo(\theta) \) denote the exponential distribution with density
\[
f(x) = \theta \exp \{-\theta x\}, \quad x > 0, \quad \theta > 0
\]
and consider independent exponentially distributed random variables \( X_i, Y_i \) where \( X_i \sim Expo(\theta) \) and \( Y_i \sim Expo(\theta + \delta) \). Suppose the parameter of interest is \( \delta \), the difference between the exponential rates, and that we observe \( n \) independent pairs \((X_i, Y_i)\).

4.1.1 An optimal confidence distribution

Starting from the simplest case, with a single pair being observed, the joint density of \( X \) and \( Y \) can be written as
\[
p(x, y|\theta, \delta) = \theta(\theta + \delta) \exp \{-\theta(x + y) - \delta y\},
\]
which we can recognize as being on the form of Theorem 2.3. Then an optimal confidence distribution for \( \delta \) is available by
\[
C_1^*(\delta) = P(Y < y_{obs}|(X + Y) = (x_{obs} + y_{obs})).
\]
Putting $Z = X + Y$, the joint density of $(Y, Z)$ is easily available, and from this one finds the conditional distribution of $Y$ given $Z = z_{\text{obs}}$ as

$$p_{Y|Z}(y|z) = \frac{\delta \exp\{-\delta y\}}{1 - \exp\{-\delta z\}} \quad \text{for} \quad 0 < y < z,$$

which can be viewed as a truncated exponential distribution. This immediately yields the optimal CD for $\delta$ as

$$C^*_1(\delta) = 1 - \exp\{-\delta y_{\text{obs}}\} \quad \text{for} \quad 0 < y_{\text{obs}} < z_{\text{obs}},$$

(4.4)

Note that, in the case where $\delta = 0$ this expression is not defined, but the limit at this point exists and equals $y_{\text{obs}}/z_{\text{obs}}$, so we just put $C^*_1(0) = y_{\text{obs}}/z_{\text{obs}}$ in this case.

For the case where $n > 1$, the joint distribution will again be of the form in Theorem 2.3, and the optimal CD is available as a direct analogue of equation 4.2,

$$C^*_n(\delta) = p(S < s_{\text{obs}} | A = a_{\text{obs}}),$$

(4.5)

where $S = \sum_{i=1}^n Y_i$, $A = \sum_{i=1}^n (X_i + Y_i)$ and $(s_{\text{obs}}, a_{\text{obs}})$ the actual values of these statistics observed in the data. This distribution has no easily available closed form parametrization, but can be simulated from using a simple MCMC scheme.

For a single value of our interest parameter, $\delta_i$, we can simulate $N$ observations of $S$ from the distribution of $S$ given $A = a_{\text{obs}}$. Then

$$\frac{1}{N} \sum_{j=1}^N I(S_j < s_{\text{obs}}) \to C_n(\delta_i) \quad \text{as} \quad N \to \infty.$$

(4.6)

To obtain the samples $S_i$ from the conditional distribution, a Metropolis style algorithm can be employed – keeping in mind that the MCMC chains are in the sample space, and not the parameter space like in the typical Bayesian procedures. Typically we perform the MCMC for a range of fixed $\delta_i \in \Delta$, and then use a smoothing mechanism to derive the full confidence distribution. See also remark 4.2 later in the section.

### 4.1.2 Objective Bayesian analysis

Trying to be an objective Bayesian in this problem poses some difficulty. From the joint distribution (4.1) one finds the Fisher information matrix as

$$I(\delta, \theta) = \begin{bmatrix}
\frac{1}{(\theta + \delta)^2} & \frac{1}{\theta^2 + (\theta + \delta)^2} \\
\frac{1}{(\theta + \delta)^2} & \frac{1}{\theta^2 + (\theta + \delta)^2}
\end{bmatrix},$$

(4.7)

which makes Jeffreys’ overall prior:

$$\pi_J(\delta, \theta) = \frac{1}{\theta(\theta + \delta)},$$

(4.8)

As discussed in section 3.3 and illustrated in example 3.3 Jeffreys’ overall prior could have some issues, and we would much rather like to utilize a reference prior
approach to deal with our nuisance parameter $\theta$. Following the recipe for reference priors regarding $\theta$ as a nuisance parameter, we get:

**Step 1:**
Jeffreys’ conditional prior, when keeping $\delta$ fixed, is available directly from the Fisher information matrix in equation (4.7) as:

$$
\pi(\theta|\delta) = \left(\frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2}\right)^{1/2}.
$$  \hspace{1cm} (4.9)

**Step 2:**
Next, we need to obtain the marginal distribution $p(x|\delta) = \int_{\Theta} p(x|\theta,\delta)\pi(\theta|\delta)d\theta$. This integral becomes

$$
p(x|\delta) = \exp\{-\delta y\} \int_0^\infty \sqrt{\theta + \delta}^2 + \theta^2 \exp\{-\theta(x + y)\}d\theta,
$$

which doesn’t really have a closed form solution – or at least it is difficult to solve. One might then try the second algorithm given in section 3.5.2 and normalize the expression over an expanding sequence of compact subsets. This entails computing the integral from equation (3.12):

$$
K_i^{-1}(\delta) = \int_{A_i} \left(\frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2}\right)^{1/2} \theta^2 d\theta,
$$

for some compact set $A_i$. Now the difficulty lies in first choosing the sequence of compact sets in a good way since for a fixed $\delta_0$, this integral will blow up near $\theta = 0$, and assuming $\delta_0 < 0$ (which one should allow for), it will blow up near $\theta = -\delta_0$ as well. The second, and larger issue is again that there aren’t really any closed form solutions available for this integral. We can however go further in the problem if we enforce a restriction on the parameter space.

**Enforcing a restriction**
If the restriction $\delta > 0$ is enforced, a reparametrization to a nicer nuisance parameter might help. Consider the transformation

$$
\psi = \delta \text{ and } \lambda = (\theta + \delta)/\theta
$$

which is 1-1 as long as $\delta \neq 0$. The parameter space now becomes $\Psi \times \Lambda = (0, \infty) \times (1, \infty)$ and the Jacobian of this transformation is

$$
J = \begin{bmatrix}
1 & 0 \\
1 & \psi \\
\lambda - 1 & -(\lambda - 1)^2
\end{bmatrix},
$$

which leads to the reparametrized information matrix:

$$
I(\psi, \lambda) = J^T I(\delta(\psi, \lambda), \theta(\psi, \lambda))J = \begin{bmatrix}
\frac{(\lambda - 1)^2}{\lambda^2 \psi^2} & 0 \\
0 & \frac{1 + \lambda^2}{(\lambda - 1)^2 \lambda^2}
\end{bmatrix}.
$$  \hspace{1cm} (4.11)
This turns out to be an orthogonal parametrization. From this point we could aim directly for Theorem 3.2 using rectangles in \((\psi, \lambda)\)-space, but will choose our rectangles in \((\delta, \theta)\)-space instead. We will come back to this issue later in the example.

The conditional Jeffreys prior becomes

\[
\pi(\lambda|\psi) = \sqrt{1 + \lambda^2} \frac{(\lambda - 1)}{\lambda}.
\]

A natural sequence of compact sets to consider in the original parametrization are the rectangles \((\delta, \theta) \in [l_i, k_i] \times [l_i, k_i]\), where \(l_i \to 0\) and \(k_i \to \infty\). This leads to the slices \(\Omega_{i,\psi} = (1 + \psi/k_i, 1 + \psi/l_i)\) in \(\Lambda\).

**Step 3 (restricted case)**

Working further with the restricted case, equation 3.14 becomes

\[
\pi_i(\psi) = \exp \left\{ \frac{1}{2} \int_{\Omega_{i,\psi}} K_i(\psi) \pi(\lambda|\psi) \log \left[ \frac{1}{\lambda} \frac{(\lambda - 1)^2}{\lambda} \right] d\lambda \right\}
\]

\[
= \exp \left\{ -K_i \log(\psi) \int_{\Omega_{i,\psi}} \pi(\lambda|\psi) d\lambda \right. \\
\left. + \int_{\Omega_{i,\psi}} K_i(\psi) \pi(\lambda|\psi) \log \left[ \frac{(\lambda - 1)}{\lambda} \right] d\lambda \right\}
\]

\[
= \frac{1}{\psi} \exp \left\{ \int_{\Omega_{i,\psi}} K_i(\psi) \pi(\lambda|\psi) \log \left[ \frac{(\lambda - 1)}{\lambda} \right] d\lambda \right\}
\]

\[
= \frac{1}{\psi} \exp \left\{ K_i(\psi) I_1^i(\psi) \right\},
\]

where

\[
I_1^i(\psi) = \int_{\Omega_{i,\psi}} \sqrt{1 + \lambda^2} \frac{\lambda - 1}{\lambda(\lambda - 1)} d\lambda.
\]

Now, we’ve narrowed the problem down to computing the integral \(I_1^i(\psi)\), and the normalizing constant from equation 3.13, i.e.:

\[
K_i^{-1}(\psi) = \int_{1+\psi/l_i}^{1+\psi/k_i} \frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} d\lambda.
\]

The following Lemma is proved in appendix A.

**Lemma 4.1.** Let

\[
I_1^i(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} \log \left[ \frac{\lambda - 1}{\lambda} \right] d\lambda \quad (4.12)
\]

and,

\[
K_i^{-1}(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} d\lambda. \quad (4.13)
\]
These integrals equals,

\[ I_i^1(\psi) = \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) + \frac{\psi}{k_i} - \log^2 \left( \frac{\psi}{k_i} \right) \right) + B_7, \]  
\[ K_i^{-1}(\psi) = \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} - 2 \log \left( \frac{\psi}{k_i} \right) + \sqrt{2} \log (1 + \psi/l_i) \right) + B_{10}, \]  

where \( B_7 \) and \( B_{10} \) are bounded functions of \( k_i \) and \( l_i \).

Then we have the following expression for \( \pi_i(\psi) \):

\[ \pi_i(\psi) = \frac{1}{\psi} \exp \left\{ K_i(\psi) I_i^1(\psi) \right\} = \frac{1}{\psi} \exp \left\{ \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) + \frac{\psi}{k_i} - \log^2 \left( \frac{\psi}{k_i} \right) \right) + B_7 \right\}. \]  

**Step 4 (restricted case)**

Now we have the expression for \( \pi_i(\psi) \), and we need to take the limit as in equation 3.15, i.e

\[ \lim_{i \to \infty} \frac{K_i(\psi) \pi_i(\psi)}{K_i(\psi_0) \pi_i(\psi_0)}. \]

To make the calculations easier, we choose \( l_i = 1/k_i \) leading to the limit being taken over compact sets \([k_i^{-1}, k_i] \times [k_i^{-1}, k_i]\) in the original parametrization, and we set \( \psi_0 = 1 \). The following lemma is proved in appendix [A]

**Lemma 4.2.** Let \( K_i(\psi) \) and \( \pi_i(\psi) \) be as defined above. Let \( \{k_i\}_{i=1}^{\infty} \) be a sequence diverging to infinity and put \( l_i = k_i^{-1} \). Then, putting \( \psi_0 = 1 \) we have that

\[ \lim_{i \to \infty} \frac{K_i(\psi) \pi_i(\psi)}{K_i(1) \pi_i(1)} = \psi^{-4+\frac{5}{\sqrt{2}}} \]  

Then, the reference prior for \((\psi, \lambda)\) becomes:

\[ \pi_R(\psi, \lambda) \propto \psi^{-4+\frac{5}{\sqrt{2}}} \sqrt{(1+\lambda^2)} \]  
\[ \frac{1}{(\lambda-1)\lambda}, \]

or, transformed back into \((\delta, \theta)\) space,

\[ \pi_R(\delta, \theta) \propto \delta^{-4+\frac{5}{\sqrt{2}}} \left( \frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2} \right)^{1/2}. \]  

So now we have an expression for the reference prior in this problem under the restriction that \( \delta > 0 \). It would of course be nice to have a reference prior without this restriction, but I have not been able to find it. There are several complicating factors that make utilizing the algorithms outlined in section 3.5.2 difficult.

First and foremost, look at the parametrization that was chosen in expression 4.10. It might seem arbitrary, but it is chosen to yield a diagonal Fisher information.
matrix, in which the terms factor into separate function of $\psi$ or $\lambda$. This means that
the integrals are simplified since the parameter of interest only crops up at the edges
of the integration area, making the technique utilized in the proof of Lemma 4.1
possible.

We might have been able to use the same technique to solve the reference prior
for the case where $\delta < 0$, since the parametrization is still 1-1 in this case, but it
would leave a gap in the parameter space at $\delta = 0$. Another complicating factor we
would have to deal with in this case, is that the parameter space of $\delta$ will depend
on $\theta$ due to the restriction that $\theta + \delta > 0$.

A second reference prior

Like we’ve discussed, how the compact sequence of subsets is chosen might affect
the reference prior, and papers working with reference priors often try to prove an
independence result of the reference prior and the compact sequence. We chose
the sets $[k^{-1}, k] \in \Delta \times \Theta$, and stated that it led to slices of the form $\Omega_{i, \psi} = (1 + \psi/k, 1 + k\psi)$ in $\Lambda$, for which we then applied the algorithms sketched out earlier.

If we instead had chosen our rectangles directly in $\Psi \times \Lambda$-space, we could have utilized
Theorem 3.2 directly, now working with rectangles in another parametrization. Since $\psi = \delta$ the reparametrization really only amounts to choosing a different nuisance parameter, so we should be fine. But, in this case it turns out that, working with the rectangles in $(\psi, \lambda)$-space and applying theorem 3.2 for the parameter grouping $\{\psi, \lambda\}$; we immediately have

$$\pi_R^*(\psi, \lambda) = \psi^{-1} \sqrt{1 + \lambda^2} / \lambda(\lambda - 1),$$

or, in the original $(\delta, \theta)$ parametrization;

$$\pi_R^* (\delta, \theta) \propto \delta^{-1} \left( \frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2} \right)^{1/2}. \quad (4.19)$$

This is also a valid way to construct a reference prior for this situation, which
leaves us with two competing reference priors. The problem now is to choose the
'best' one, by some metric.

To summarize, we are in the situation where we have three competing priors,
Jeffreys’ overall prior $\pi_J$, and the two reference priors, $\pi_R$ and $\pi_R^*$. In terms of how
the marginal posterior behaves, $\pi_J$ and $\pi_R$ seem to yield proper marginal posteriors,
even for $n = 1$. I do not have a proof of this statement, but it is a conjecture based on
thousands of simulations across a wide range of parameter values and samples. The
reference prior $\pi_R^*$ is less clear, for some samples it yields nice marginal posteriors
for $\delta$, but other times it appears to not have a proper posterior distribution at all.
Due to the large amount of weight it places near zero, MCMC must be done with
care to avoid the chains getting stuck, perhaps by a transformation to $\kappa = \log \delta$.

4.1.3 Comparisons

There are a few reasons why the prior in $(4.19)$ is appealing. Since we are enforcing
the restriction that $\delta > 0$ a priori, the prior $\delta^{-1}$ amounts to a flat prior on $\log \delta$,
which is quite a standard non-informative prior under this restriction. Another
reason is that of symmetry; it looks nice. There is a sense that when something this simple tumbles out of the theory – there must be something to it. Especially when the other candidate contains the strange looking $\delta^{-4+5/\sqrt{2}}$.

A third reason has to do with the underlying logic of the construction procedure. Recall the heuristic explanation for the compact sets from section 3.5.4; their shapes are meant to represent shapes in the parameter space one might be indifferent over. In this case, where the parameters are orthogonal – it seems natural to work with rectangles in $(\Psi \times \Lambda)$ space, since the value of one will not affect the parameter space of the other. On the one hand, considering rectangles in $\Delta \times \Theta$ space, like we did for the prior in (4.18), leads to oddly shaped ‘wedges’ in the $(\psi, \lambda)$ parametrization, see figure 4.1. It is difficult to give a justification to why we should be indifferent over these shapes, instead of rectangles.\footnote{One might argue that an even more intuitive way to construct the sequence is on the basis the parameters $\phi_1 = \theta$ and $\phi_2 = (\theta + \delta)$, and consider rectangles in this parametrization.} On the other hand, since $\lambda = (\theta + \delta)/\theta = 1 + \delta/\theta$, as $\theta$ becomes arbitrarily small, $\lambda$ becomes arbitrarily large, no matter the value of $\delta$. This squeezes the parameter space into a point near $\theta = 0$, which might also be unnatural in a sense.

It turns out though, when we look at the marginal posteriors, the reference prior $\pi^\ast_R(\delta, \theta)$ often performs poorly compared to the one derived earlier. The reason being that $\delta^{-1}$ places too much weight near zero, much more so than $\delta^{-4+5/\sqrt{2}} \approx \delta^{-0.47}$. This is especially undesirable for low $n$, where the likelihood is not powerful enough to pull weight away from zero. In figure 4.2, the marginal posterior distributions for $\delta$, based on Jeffreys’ overall prior and the two reference priors, are displayed for two different samples at $n = 3$. In both samples, it is easy to see that the marginal posterior for $\delta$, obtained by using $\pi^\ast_R(\delta, \theta)$, is pulled far to the left compared to both the overall Jeffreys’ prior and the reference prior in (4.18).

This problem, where different sequences lead to different reference priors was first encountered and studied in Berger & Bernardo (1989). In this paper, a natural reference prior comes from choosing sequences of subsets in the original parameter space, and going through the same proofs as I did for lemmas 4.1 and 4.2. They are able to prove that, for general choices of sets in the original parametrization (rectangles, circles etc.), the reference prior is unchanged. But, following the same argumentation we did, where it might be more natural to consider rectangles in $(\psi, \lambda)$-space, they derive a second reference prior based on this sequence. After some comparisons, considering the logic of the underlying procedure and looking at the marginal posteriors, they choose the reference prior based on subsets in the

---

Figure 4.1: Compact rectangles in $(\Delta \times \Theta)$-space transform to wedges in $(\Psi \times \Lambda)$-space
CHAPTER 4. COMPARISONS AND EXAMPLES

Figure 4.2: Marginal posterior distributions for $\delta$ based on the two reference priors $\pi_R$, $\pi^*_R$ and the overall Jeffreys’ prior $\pi_J$. The left panel has $(\sum y_i = 1, \sum x_i + y_i = 2)$ while the right panel has $(\sum y_i = 0.1, \sum x_i + y_i = 1.1)$.

original parametrization. This seems to have become the standard way to derive reference priors in the literature, cf. the quote by Berger and Bernardo in section 3.5.3.

Berger & Bernardo (1989) also recommended using frequentist criteria to evaluate the different priors, focusing on the coverage of credible intervals. I would argue that a nice way of comparing reference priors in this example is to compare the marginal posterior distributions to the uniformly optimal confidence distribution. The uniformly optimal confidence distribution achieves one of the large goals that the reference prior theory is trying to achieve, namely a posterior distribution as independent as possible from the prior.

Remark 4.1 (Comparing CDs and Bayesian posteriors). When we are comparing the optimal confidence distribution to our Bayesian posterior, we run into some issues. There are many ways to construct credible intervals from a marginal Bayesian posterior, and most often, interest is on the highest posterior regions credible intervals (HPD intervals). An $\alpha$-level HPD region, is a set $A \in \Theta$ in which, the probability content of the marginal posterior is exactly $\alpha$, and the marginal posterior density within $A$ is always higher than outside. In the traditional (subjective) Bayesian paradigm, these sets represent regions of the parameter space that that our combined experience (prior + likelihood) should view as the most likely location of the underlying parameter of interest. An appealing property is that for any given level of confidence $\alpha$, the HPD regions are always the shortest intervals (or unions of shortest intervals), for our parameter of interest. This could make them the ideal comparison tool for our optimal CD. If the marginal posterior is unimodal, the highest posterior density is also an equal tailed credible interval.

In the objective Bayesian paradigm, we’ve discussed the underlying logical con-
tent of posterior distributions based on objective priors. It is unclear to me if HPD regions would be ideal under this paradigm since we’ve lost the original justification; namely providing a combined representation of our knowledge about parameter, based on the prior and the data. Most Bayesians however, continue to regard their posterior distribution as one of epistemic probability, even when using a non-informative prior. Fisher felt that once the data had been collected, his Fiducial distributions also represented an epistemic probability distribution over the parameter space. This is also the view of some authors within the CD framework, and I’m inclined to make it my own. Without worrying to much about the interpretations I will convert marginal posterior distributions into posterior cumulative distributions and cc-style plots in order to compare them with the uniformly optimal CDs. Within the context of probability matching, if these curves are close to the optimal CD, we will have good matching of one-sided quantiles, something we know should be true when Jeffreys’ prior is employed, cf. the references in section 3.6.

Figure 4.3 shows marginal posterior distributions based on $\pi^*_R$ and $\pi_R$ contrasted to the optimal confidence distribution. The marginal distributions have been converted to cdfs and credible-curves\(^2\) to compare them to the CD. The CD itself is calculated at 100 evenly spaced points in the parameter space, and smoothing splines have been used to create a smooth final distribution. The figure illustrates a general trend when comparing the two priors – $\pi^*_R$ either performs really well (as in the bottom two plots), or horribly (as in the top two plots). It all depends on which samples have been selected, and the problem persists even for moderately large $n$. The prior distribution $\pi_R$, on the other hand, performs on average quite well – for most samples. This makes sense, as it is part of the construction to average over the sample space, cf. definition 3.2. Even when the marginal distribution misses its target, like in the bottom panels, it is never too far off, at least not in terms of how confidently it places weight. The situation in the top panels is pretty common, and seems heavily dependent on the samples selected, and for this reason I will for now, dismiss the reference prior $\pi^*_R$ in favour of $\pi_R$.

**Frequentist properties**

It is of interest to study the frequentist properties of non-informative priors in this example. Jeffreys’ prior is known to have good properties, but no general theory exists for the reference priors, and they must be studied in each case. The properties of interest are, the coverage probability of quantiles and credible intervals, as well as the average length of credible intervals. This is especially interesting when contrasted to the optimal confidence distribution, which have exact coverage, but also the shortest intervals no matter the sample.

These properties are studied for three different prior distributions; Jeffreys’ overall prior, $\pi_J$, the reference prior from equation \((4.18)\), $\pi_R$, and the flat prior $\pi(\delta, \theta) \propto 1$. For all three priors, simulations where done on a range of parameter values $(\delta_0, \theta_0)$ and different number of observations $n$, using a Metropolis-Hastings type algorithm. For given values $(\delta_0, \theta_0, n)$, and a given prior distribution, $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are drawn from the true distribution, and samples from the joint posterior obtained by the Metropolis-Hastings algorithm.

\(^2\)The bayesian version of the confidence curves.
Figure 4.3: Marginal posterior distributions based on $\pi^*_R$ and $\pi_R$ compared to the optimal confidence distribution for two different samples.
4.1. DIFFERENCE OF EXPONENTIAL PARAMETERS

Then coverage is checked by comparing quantiles and intervals to the known true values of the parameter \( \delta_0 \).

The algorithm was run with multiple starting values, while monitoring convergence statistics and acceptance probabilities, per the recommendations in [Gelman et al.] (2014, ch. 11, 12). The proposal distribution in the simulations was the multivariate normal distribution centered at the previous sample, with covariance matrix

\[
\Sigma = \begin{bmatrix}
\frac{1}{\sum y_i} & 0 \\
0 & \frac{1}{\sum (x_i + y_i)}
\end{bmatrix}.
\]

This proposal distribution performed well across the range of parameters, number of observations, and the three different priors. The underlying logic is that; since the expectation of an exponentially distributed random variable with rate \( \lambda \) is \( \lambda^{-1} \), small values of \( \sum y_i \) correspond to large values of \( \delta \), thus needing larger jumps to explore the entire posterior distribution.

In figure 4.4, the coverage probabilities for \( \delta \) are displayed for the three priors, at the quantiles \{2.5\%, 25\%, 50\%, 75\%, 97.5\%\}. At each point in the boxplot the coverage probability has been calculated by 1000 simulation from samples at a specific combination of \((\delta_0, \theta_0, n)\) for the three priors. The large outliers are due to the simulations being run for small values of \( \delta_0 \) at low \( n \), meaning that the resulting posterior distributions often will cover by default – especially for the larger quantiles. What is interesting about the figure is that it appears that Jeffreys’ overall prior is more sensitive to this phenomena than the reference prior, which is slightly counter-intuitive since Jeffreys’ doesn’t have a singularity at \( \delta = 0 \) like \( \pi_R \) does. The singularity might have the effect of pulling the quantiles just enough towards zero so that better coverage is obtained in this situation.

Overall though, looking at the median, Jeffrey’s prior performs really well, always staying close to the intended quantile (dotted line). The reference prior doesn’t do bad either, but its median is always slightly larger than the Jeffreys posterior. The flat prior is the worst of the three, which might be expected. Its quantiles consistently

Figure 4.4: Coverage probabilities of quantiles from Jeffreys’, Reference and the ‘Flat’ priors.
overshoot its target due to the fact that there is no mechanism that pulls posterior weight towards zero.

In figure 4.5, the coverage probabilities for equal-tailed 95% credible intervals are displayed for the three priors at different sample sizes. Again we see that for low $n$, Jeffreys’ and the flat prior are sensitive to outliers, while the reference prior performs better. Interestingly, as $n$ increases to 50, Jeffreys’ and the flat priors seem to outperform the reference prior, which has a larger spread around the intended coverage. This might also be coincidental, and more simulations might erase this discrepancy.

The average length of credible intervals is also of interest. A good procedure should have the correct coverage, as well as a tight distribution around the true value – like the optimal CDs.

Remark 4.2 (Simulations from the optimal CD). There is a problem that need to be dealt with in order to have consistent simulation results from the optimal confidence distribution. Unlike the Bayesian MCMC, where the Markov chains are defined directly on the parameter space, the optimal CDs are defined in terms of a probability statement (equation 4.5) on the sample space, for a fixed parameter value. The simulations must be repeated across a range of parameter values in the parameter space in order to approximate the full confidence distribution. What is important then, is that the range is defined in such a way that it covers all the interesting parts of the parameter space, for example so that it contains at least the 1st and 99th percentiles of $C(\delta)$. One of the issues that then arises, is that this range will be data dependent, so that defining one range to fit all samples will lead to a computationally inefficient simulation procedure. In the following, I will describe an algorithm for solving this issue.

First off, we need to find the ranges of interest for the $n$ individual CDs based on a single observation, $C_1(\delta)$. This can be done by applying a root-finding algorithm on the function

\[ C_1(\delta) - \alpha = 0 \]
4.1. DIFFERENCE OF EXPONENTIAL PARAMETERS

for a percentile $\alpha$ of interest, e.g. the 1st and 99th percentile. When these percentiles have been found for each of the $n$ individual CDs, we choose an initial range of interest to be the smallest 1st percentile, and the largest 99th percentile, then calculate all the individual $C_1(\delta)$ on this range. Then we have $n$ individual confidence distributions calculated over the same, wide range of parameter values. Next we find the median CD, $C_m(\delta)$, on this range by calculating the median at each point. We choose the median, and not the mean, because it is less sensitive to outliers. One might argue that a more direct combination such as

$$C(\delta) = \Phi \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Phi^{-1} \left( C_i(\delta) \right) \right),$$

(4.20)

where $C_i()$ denotes the $i$th individual CD, would be better than the median, but I've found it computationally difficult to work with in this problem.

Lastly, we split the interval $(0, 1)$ into 100 equal pieces, $(\alpha_1, \ldots, \alpha_{100})$, and find the corresponding values in the parameter space by $C_m^{-1}(\alpha_i)$. These 100 points in the parameter space constitutes our range of interest, and implicitly samples points closer together where the median distribution changes rapidly. This is an estimate for how the uniformly optimal CD will behave, and the hope is that the resulting sequence will cover the interesting part of the resulting optimal CD.

In figure 4.6, the result of the technique is illustrated. The soft grey dotted lines are the individual CDs, while the black dotted line is the median confidence distribution. The tick marks on the x-axis represent the sequence found by the above described technique, and added is the uniformly optimal confidence distribution calculated at each of these points. Lastly, the black solid line represents a smoothed confidence distribution calculated by smoothing splines using the point estimates. We see that the technique does what it is intended, and can be used to quickly find ranges of interest in the parameter space for the uniformly optimal CD.

The code for this algorithm is in appendix B.

In figure 4.7, the average length of 95% equal tailed credible intervals are compared to the corresponding uniformly optimal confidence interval. At each point in
Figure 4.7: Average length of equal-tailed 95% credible intervals, for Jeffreys’, Reference and ‘Flat’ priors, at \( n = 5, 10, 50 \)

For \( n = 5 \), the confidence intervals are on average a bit shorter than the credible intervals, but there are combinations, of \((\delta_0, \theta_0)\), where the average length of credible intervals dips below the confidence intervals. For Jeffreys’ posterior, and the flat posterior, this must be viewed in connection with the coverage outliers in figure 4.5, but the reference posterior had no such large outliers. The reason is simply that it places much more weight near zero than the two other priors to leading to shorter 95% intervals, which for some true, small \( \delta_0 \) the simulations was run for, still has the correct coverage. For the confidence intervals, keep in mind that they do not have a built in mechanism that take into account the \textit{a priori} knowledge that \( \delta > 0 \). This means that confidence intervals are free to span the entire parameter space. To make them comparable to the Bayesian posteriors, I’ve truncated the intervals at zero. What this means is that a confidence interval such as \((-2, 5)\) would get the length of the interval \((0, 5)\), while the interval \((-10, -1)\) would get completely disregarded.\footnote{This was a very rare event in my simulations.}

For \( n = 10 \) and \( n = 50 \) we start to see some really similar results across the posteriors, and the confidence interval. There is still the trend that the CI is shorter on average, closely followed by Jeffreys’ prior and then the reference prior. In addition, the lower bounds of the boxplots now seem to be similar across the posteriors and the confidence intervals. For \( n = 50 \), the average confidence interval is slightly longer than the credible intervals, this is most likely due to the smaller number of simulations that went into it. Perhaps also the simulation scheme of remark 4.2 is not quite optimal.

Based on these simulations, I’m quite pleased with how the reference prior \( \pi_R \) is performing. For this problem though, it seems that Jeffreys’ prior does equally well, or slightly better. It is a bit more sensitive to outliers in the coverage, but it seems
to me that this is more due to the restriction $\delta > 0$ than the prior itself. Removing this restriction, Jeffreys prior continues to perform well, with inference comparable to the uniformly optimal confidence distribution.

### 4.1.4 Boundary parameters

I had to enforce a restriction on $\delta$ in order to find the reference prior in this problem. Working within the confines of this problem, assuming we know this as a fact, the marginal posterior distributions obey this restriction by only being defined on the set $(0, \infty)$. If we had been interested in the possibility that $\delta$ might actually equal zero, we would need to include a separate prior point mass at $\delta = 0$ and include this in our calculations. The confidence distribution does not have any restrictions of this sort at all, and is defined on the entire parameter space. When a confidence interval contains zero, we will simply say that we cannot confidently infer that the parameter isn’t zero.

Consider the situation in figure 4.8. Plotted is the uniformly optimal cc based on a random sample of five pairs from a setup where $(\delta_0, \theta_0) = (1, 2)$. Also added are the individual cc’s for each pair. Since we assume a priori, that $\delta$ is larger than zero the curves are only plotted on the interval $(0, 10)$. Three of the individual cc’s are centered on $\mathbb{R}^+$ while the two others are actually centered at values below zero. The uniformly optimal cc is centered a little below the true value of $\delta_0 = 1$, and has a tighter distribution than the individual cc’s, which is natural since it is a combination of all the information from the 5 pairs.

What is interesting about the optimal cc in this case, is that for all confidence levels larger than roughly 0.3 the natural confidence intervals will contain $\delta = 0$. Meaning that we cannot confidently exclude this possibility.

What would the Bayesians do in this setup? Both reference priors have a singularity at $\delta = 0$, meaning that they shift a lot of weight towards this possibility, and highest posterior regions are often of the form $(0, \epsilon_\alpha) \cup (f_1, \alpha, f_2, \alpha)$ for some small value $\epsilon_\alpha$ depending on the credible value $\alpha$. One might hope that enough weight is
shifted so that credible curves from the marginal posteriors will resemble the optimal cc in this case.

The results, based on the same sample as before, are displayed in figure 4.9, where the uniformly optimal cc is compared to credible curves from two different marginal posteriors; the reference posterior based on $\pi_R$, and Jeffreys’ posterior based on $\pi_J$. Numerical integration of the reference posterior based on $\pi^*_R$ seems to yield an improper posterior distribution in this case, as it does quite often, see also the concluding remarks.

We see that the reference prior does a good job at pulling weight towards zero, staying closer to $C_5(\delta)$ near the boundary than Jeffreys’. It also has a median that is centered closer to the median CD point estimator. Of course Jeffreys’ prior with no restriction would stay pretty close to the CD here, but the point is to study a case where such a restriction occurs, and how the Bayesian paradigm handles it. I mean, if we know, a priori, that $\psi \geq \psi_0$ for some value $\psi_0$, then the Bayesians seems to have some extra difficulties that the frequentists don’t – the natural confidence intervals will simply contain $\psi_0$ for some level of confidence $\alpha > \alpha_0$.

### 4.2 Unbalanced Poisson pairs

This example allows us to tackle confidence distributions based on a discrete sample space with appropriate continuity corrections. I find that a continuity corrected fiducial distribution coincides exactly with the Bayesian posterior based on Jeffreys’ or the reference prior, and discuss exact matching in the context of fiducial inference and confidence distributions.

Consider the situation where we have Poisson distributed random variables, $X_1, \ldots, X_n \overset{iid}{\sim} \text{Po}(\lambda)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} \text{Po}(\lambda \psi)$. The parameter of interest is the ratio between the two Poisson means, i.e. $\psi$, while $\lambda$ is considered a nuisance parameter. The data can be considered as coming from two different counting processes of length $n$ and $m$, during which the rate is constant at $\lambda$ and $\psi \lambda$, respectively.
These types of models have wide application in a variety of fields, and have been studied from both the frequentist and Bayesian paradigms.

The joint density of this model can be written

\[ p(x, y|\lambda, \psi) \propto \lambda^{\sum_i x_i + \sum_j y_j} \psi^{\sum_j y_j} \exp \left\{ -n\lambda - m\lambda\psi \right\}, \]

where \( A = \sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j \) and \( S = \sum_{j=1}^{m} y_j \). This is immediately recognized as being on the form of Theorem 2.3, and an optimal confidence distribution for \( \psi \) is directly available through

\[ C_n(\psi) = P(S > s_{obs} | A = a_{obs}). \]  
(4.21)

This distribution can be worked out exactly by some well known facts about the Poisson distribution. First, the sums of independent Poisson distributed variables are themselves Poisson distributed, this yields

\[ S \sim \text{Po}(m\psi\lambda), \quad A = \sum_i x_i + \sum_j y_j \sim \text{Po}(n\lambda + m\psi\lambda). \]  
(4.22)

Also, the conditional distribution of \( S \) given \( A \) is binomial;

\[ S|A \sim \text{Bin}\left(A, \frac{m\psi}{n + m\psi}\right). \]  
(4.23)

We haven’t touched much upon discrete data in the material so far, but it has its own quirks that needs to be sorted out. For discrete data, exact confidence intervals cannot be obtained at every level of significance without some sort of randomization mechanism. The reason is of course that the equality \( P(X < x) = P(X \leq x) \) does not hold. One way to deal with this is through some sort of continuity correction. Let \( C^1(\psi) \) denote the CD based on a strict inequality in (4.21) and \( C^2(\psi) \) be based on the non-strict version. A typical way to apply a continuity correction is the take the (arithmetic) average of these two distributions, this leads to the typical ‘half-corrected’ confidence distribution;

\[ C_A(\psi) = \frac{1}{2} \left( C^1(\psi) + C^2(\psi) \right) = P(S > s_{obs} | A = a_{obs}) + \frac{1}{2} P(S = s_{obs} | A = a_{obs}). \]  
(4.24)

Since this expression is a linear combination of confidence distribution, we might have lost the exact coverage property, but the resulting distribution is an approximate confidence distribution. There is a strong line of argument within the context of hypothesis testing that leads to this type of continuity correction, see Stone (1969, Appendix 5).

Another continuity correction to consider is taking the geometric mean of the two confidence densities,

\[ c_G(\psi) \propto \left( c^1(\psi)c^2(\psi) \right)^{1/2}, \]  
(4.25)

and constructing the confidence distribution \( C_G(\psi) \) on the basis of this. Again, the exact coverage property might be lost, so the distribution is only an approximate CD. The reason one might consider this expression instead of \( C_A(\psi) \) is because, when combining distributions \( p_1 \) and \( p_2 \) with the same support, the distribution,
$q$, that minimizes the sum of the Kullback-Leibler divergences $\kappa(q|p_1) + \kappa(q|p_2)$ is precisely the geometric mean of $p_1$ and $p_2$. Under some mild conditions\footnote{see \cite{VeroneseMelilli2017} prop. 2}, that are satisfied in this example, $C_G(\psi)$ 'lies between' $C^1$ and $C^2$ so that it in some sense is a compromise between the two.

In example \footnote{cf. \cite{VeroneseMelilli2017} prop. 3} 2.2 I derived the fiducial distribution of the probability parameter, $p$, in a binomial distribution as a beta distribution. \cite{VeroneseMelilli2017} states that the fiducial distribution of a binomial $p$, with observed data $x$ and fixed $N$, under the continuity correction (4.25) is also a beta distribution; $\text{Be}(x + 1/2, N - x + 1/2)$. Plugging in the expression from equation (4.23) we have that

$$\frac{m\psi}{n + m\psi} \overset{df}{\sim} \text{Be}(S + 1/2, A - S + 1/2)$$

and thus an approximate confidence distribution under the geometric average is

$$c_G(\psi) \propto \frac{\psi^{S+1/2}}{(n + m\psi)^{A+1}},$$

which makes $C_G(\psi)$ a regularized incomplete beta function.

On the Bayesian side of things, the natural place to start is with the Fisher information matrix, which easily follows from the joint distribution:

$$I(\psi, \lambda) = \begin{bmatrix} m\lambda & m \\ \psi & n + m\psi \end{bmatrix}$$

Jeffreys’ overall prior is

$$\pi_J(\psi, \lambda) = (\det I(\psi, \lambda))^{1/2} \propto \frac{1}{\sqrt{\psi}},$$

and the expression for the reference prior, in the case where $\lambda$ is considered a nuisance parameter, is quickly available; from the Fisher information matrix, the conditional Jeffreys’ prior of $\lambda$ given $\psi$ is

$$\pi(\lambda|\psi) = \sqrt{\frac{n + m\psi}{\lambda}},$$

which is improper. But since we are working directly in the original parameter space, and the parameters are variationally independent; Theorem\footnote{Veronese & Melilli (2017) prop. 3} 3.3 can be applied directly since both the Fisher information, $I(\psi, \lambda)$, and its inverse factors;

$$I^{-1}(\psi, \lambda) = \begin{bmatrix} \psi(n + m\psi) & \psi \\ \lambda mn & \psi \end{bmatrix}$$

Then we have the reference prior immediately as,

$$\pi_R(\psi, \lambda) = \frac{1}{\psi(n + m\psi)} \frac{1}{\sqrt{\lambda}} \propto \frac{1}{\sqrt{\psi\lambda\left(1 + \frac{m}{n}\psi\right)}}.$$
The particularly interesting thing here is that the reference prior acknowledges the unbalanced sampling design through the ratio \( m/n \), controlling the level of imbalance. Jeffreys’ overall prior does not have this type of built-in balancing mechanism, but it turns out that the reference prior and Jeffreys’ overall prior leads to the same marginal posterior in this specific example.

\[
p(\psi | \mathbf{x}, \mathbf{y}) \propto \frac{\psi^{s+1/2}}{(n + m\psi)^{A+1}}, \tag{4.30}
\]

which is of course the same distribution as the confidence distribution in expression (4.26). That is, we have an exact match between the optimal confidence distribution, based on a certain continuity correction, and a posterior distribution based either on Jeffreys’ prior, or a reference prior. This isn’t the first time we’ve seen an equality between an, inherently fiducial distribution, and Jeffreys’ prior. We’ve already studied the ordinary linear model in example 2.4. It is, however, the first time we’ve seen equality connected to the reference priors.

### 4.2.1 Exact Matching

The question of exact matching between fiducial and Bayesian inference has a long history, and it is an appealing question; when does the Bayesian posterior coincide with the fiducial distribution? Can we always find a prior distribution such that this happens? In the simplest setup, where we have a one-dimensional continuous model, and a fiducial distribution constructed on the basis of a pivot, Lindley (1958) was the first to answer.

Letting \( F(\mathbf{x} | \theta) \) denote the cumulative distribution function of such a model, under some restrictions discussed in 2.1, the fiducial density of \( \theta \) is simply \(-\partial/\partial \theta F(\mathbf{x} | \theta)\). The Bayesian posterior distribution is given by

\[
p(\theta | \mathbf{x}) = \frac{\pi(\theta)p(\mathbf{x} | \theta)}{p(\mathbf{x})},
\]

where \( p(\mathbf{x}) = \int \pi(\theta)p(\mathbf{x} | \theta) d\theta \) and \( p(\mathbf{x} | \theta) = \partial/\partial \mathbf{x} F(\mathbf{x} | \theta) \). Our question now boils down to solving the differential equation:

\[
-\frac{\partial}{\partial \theta} F(\mathbf{x} | \theta) = \frac{\partial}{\partial \mathbf{x}} F(\mathbf{x} | \theta) \frac{\pi(\theta)}{p(\mathbf{x})} \tag{4.31}
\]

for the prior distribution \( \pi(\theta) \). Finding the solution to this equation, Lindley answers the question; The fiducial distribution corresponds to a Bayesian posterior with uniform prior, if and only if, the parameter is a location parameter, or can be made into a location parameter by reparametrization. For a scale parameter, \( \phi \), this amounts to a uniform prior on \( \log \phi \), i.e. \( \pi(\phi) \propto \phi^{-1} \). In the context of the exponential family, Lindley noted that his criteria was only fulfilled within the Gamma and Normal family of distributions (or distributions that can be transformed into these).

The result by Lindley was later generalized by authors working within the framework of generalized fiducial inference, and answered also in the case of discrete distributions in the natural exponential family. Veronese & Melilli (2015) proved that equality between Bayesian posteriors and fiducial distributions can be obtained,
for the one-parameter case, also in the binomial, poisson, and negative-binomial distributions, if the geometric mean continuity correction in equation 4.25 is employed.

In the above example though, we had a multi-parameter model with a parameter of interest, $\psi$, and a nuisance parameter $\lambda$. But we still obtained an exact match between a continuity corrected fiducial distribution, and the Bayesian posterior based on Jeffreys’ or a reference prior approach. The reason for this is connected to our model being a member of the conditionally reducible natural exponential family, cr-NEF. These are members of the exponential family, whose conditional distributions are also members of the exponential family. That is, if $p(x|\theta)$ is a member of the natural exponential family, with parameters $\theta = (\theta_1, \ldots, \theta_d)$, and the density can be written as

$$p_\theta(x) = p_\theta(T_1)p_\theta(T_2|T_1)p_\theta(T_3|T_1, T_2) \cdots p_\theta(T_d|T_1, \ldots, T_{d-1}),$$

(4.32)

for some statistics of the data $(T_1, \ldots, T_d)$. Then, if each conditional distribution is itself a member of the exponential family, this distribution is conditionally $d$-reducible, or a member of the cr-NEF.

In our example, the model can be written as

$$P(x, y|\psi, \lambda) = P(A)P(S|A)$$

where the distribution of $S|A$ was binomial, and $A$ Poisson distributed. This reduction is important for obtaining exact matching between the Bayesian posterior and the uniformly optimal CD.

Results concerning this, and proofs are found in Veronese & Melilli (2017), where a characterization of the cases where equality holds, also in multi-parameter models, is given. Their findings are essentially that, if the data is distributed according to a cr-NEF, and a joint generalized fiducial distribution is constructed by converting each term in the product (4.32) to a fiducial distribution, and taking the product of these; then, equality between the joint Bayesian posterior and the joint generalized fiducial is obtained if, and only if, each conditional distribution is a transformation of one of the following families: Normal with known variance, Gamma with known shape, binomial, Poisson or negative-binomial. The prior distribution that ensures equality is the reference prior. For discrete cases, the continuity correction must be the geometric mean in equation 4.25.

Veronese & Melilli (2017) further notes that the individual fiducial distributions are also confidence distributions, or approximately so in the discrete case. By establishing an order of inferential importance of the parameters in the model, and applying the procedure of stepwise conditioning such that the last term in the product (4.32) is sufficient for a parameter of interest, $\psi$, the fiducial distribution based on this final term is also an optimal CD for $\psi$ according to Theorem 2.3.

Our example fits into this scheme as both $A$ and $S|A$ is a member of the given distributions, we used the reference prior approach, and the geometric mean correction. This would entail that our joint reference posterior distribution would coincide with the joint fiducial distribution constructed in the above fashion.

However, the theory in the preceding paragraphs establishes conditions under which the joint reference posterior distributions coincide with joint fiducial densities.

---


7That Jeffreys’ overall prior yields the same marginal posterior as the reference prior in the example seems to me a matter of mathematical coincidence.
I’ve argued that one of the marginal fiducial distributions that comprise the joint fiducial is an optimal CD. But I can see no guarantee in the theory that exact matching between the uniformly optimal confidence distribution and the marginal posterior is obtained. It happened in our case, but this seems to me something that must be checked in every example. It is clear however, that the above references are a good starting point for proving something along these lines.

### 4.3 Linear combination of Normal means

The following example looks at confidence and objective Bayesian inference in the context of multiple nuisance parameters, allowing us to fully utilize the ’grouping’ arguments looked at regarding reference priors. It introduces new layers of complexity as we go along, starting from the easiest case where optimal frequentist inference is always available.

Consider the following, ANOVA-style setup where

\[ Y_{ij} \sim N(\mu_j, \sigma^2) \quad i = 1, \ldots, n_j \quad j = 1, \ldots, m, \]

where independence is both within and between groups. Let’s first consider the simplest case where \( \sigma \) is considered known \( \text{a priori} \), and say we are interested in estimating a linear combination of the normal means, e.g. \( \psi = \sum_{j=1}^{m} c_j \mu_j \), for given \( c_j \). From a confidence distribution viewpoint, this is actually quite an easy task, as the whole problem can be reduced to a one-dimensional case. Let \( \bar{y}_j \) denote the within-group means, i.e. \( \bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j} \), then it is an easy task to check that

\[ \hat{\psi} = \sum_{j=1}^{m} c_j \bar{y}_j \sim N \left( \psi, \sigma^2 \sum_{j=1}^{m} \frac{c_j^2}{n_j} \right). \]

Thus, the natural confidence distribution for \( \psi \) is simply by

\[ C(\psi) = \Phi \left( \frac{\psi - \hat{\psi}}{\sigma \sqrt{\sum_{j=1}^{m} \frac{c_j^2}{n_j}}} \right). \quad (4.33) \]

In fact, this confidence distribution is uniformly optimal, cf. theorem 5.12 in Schweder & Hjort (2016).

From a Bayesian point of view, deriving prior distributions begin with the Fisher information, which for the full likelihood is,

\[ I(\mu_1, \ldots, \mu_m) = \frac{1}{\sigma^2} \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_m \end{bmatrix}. \]

Consider the orthogonal parametrization

\[ \theta_1 = \psi = \sum_{i=1}^{m} c_i \mu_i, \quad \theta_i = \frac{(\sum_{i=1}^{m} c_i \mu_i) c_i}{\sum_{j=1}^{m} \frac{c_j^2}{n_j}} - \mu_i \quad \text{for } i = 2, \ldots, m, \]
or, in terms of the original parameters;

\[
\mu_1 = \frac{c_1}{n_1} \theta_1 + \sum_{i=2}^{m} (c_i/c_1) \theta_i \\
\mu_i = \frac{c_i}{n_i} \theta_1 - \theta_i \text{ for } i = 2, \ldots, m.
\]

which yields the reparametrized Fisher information matrix

\[
I(\psi, \theta_2, \ldots, \theta_m) = \frac{1}{\sigma^2} \left[ \frac{c_1^2/n_1}{\sum_{j=1}^{m} c_j^2/n_j} \begin{bmatrix} 0^T \\ A \end{bmatrix} \right],
\]

where \(A\) is a diagonal \((m-1) \times (m-1)\) matrix with elements \(\{n_2, \ldots, n_m\}\). By working with compact sets in the orthogonal parametrization, the reference prior may be found by applying theorem 3.2 for the parameter grouping \(\{\psi_1, \{\theta_2, \ldots, \theta_m\}\}\). Then we immediately get

\[
\pi_R(\psi, \theta_2, \ldots, \theta_m) \propto 1.
\]

which is also Jeffreys’ overall prior for this problem. Notice also that this reference prior is independent of how the parameters are grouped. Kim et al. (2006) inadvertently gives the marginal posterior distribution of \(\psi\) under the flat reference prior, while deriving the marginal posterior for the case where \(\sigma^2\) is unknown.

\[
\pi(\psi|y) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{m} c_i^2/n_i \right]^2 \sum_{j=1}^{m} \frac{c_j^2}{c_j/n_j} \left( \psi - \bar{y}_j \sum_{i=1}^{m} \frac{c_i^2/n_i}{c_j/n_j} \right)^2 \right\}. \tag{4.34}
\]

This expression can be viewed as the product of \(m\) normal density functions, with mean and variance

\[
\mu_j = \bar{y}_j \sum_{i=1}^{m} \frac{c_i^2/n_i}{c_j/n_j}, \quad \sigma_j^2 = \frac{\sigma^2 \left[ \sum_{i=1}^{m} c_i^2/n_i \right]^2}{\sum_{j=1}^{m} \frac{c_j^2}{c_j/n_j}} \quad j = 1, \ldots, m. \tag{4.35}
\]

The product of \(m\) such densities is itself a normal density with mean, \(\mu\), and variance \(\kappa^2\);

\[
\mu = \left( \sum_{j=1}^{m} \frac{\mu_j^2}{\sigma_j^2} \right) \kappa^2, \quad \frac{1}{\kappa^2} = \sum_{j=1}^{m} \frac{1}{\sigma_j^2}.
\]

Plugging in the means and variances from 4.35 we find that the posterior in 4.34 is proportional to a normal distribution with mean \(\bar{\psi} = \bar{y}_j \sum_{i=1}^{m} \frac{c_i^2/n_i}{c_j/n_j}\) and variance \(\kappa^2 = \sigma^2 \sum_{i=1}^{m} \frac{c_i^2/n_i}{c_j/n_j}\). But this is precisely the same distribution as what the uniformly optimal CD (4.33) is based on! So in this case, the objective Bayesian approach based on Jeffreys’ or the reference prior, and the uniformly optimal CD approach yield the same result.

---

8See the proof of Theorem 5 in Kim et al. (2006).

9see for example Bromiley (2003).
4.3. LINEAR COMBINATION OF NORMAL MEANS

4.3.1 Unknown variance

As an interesting extension of the previous section, let’s now assume that the common variance \( \sigma^2 \) is unknown. Starting from our uniformly optimal CD in 4.3.3, a natural modification would be to base our CD on the t-distribution. Letting \( s_p^2 \) denote the pooled sample variance;

\[
s_p^2 = \frac{1}{N - m} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 ,
\]

where \( N = \sum_{j=1}^{m} n_j \). From this, we can construct the traditional t-distributed pivot

\[
\frac{\psi - \hat{\psi}}{s_p \sqrt{\sum_{j=1}^{m} c_j^2 / n_j}} \sim t_{N - m},
\]

and a CD is immediately available by;

\[
C(\psi) = T_{N - m} \left( \frac{\psi - \hat{\psi}}{s_p \sqrt{\sum_{j=1}^{m} c_j^2 / n_j}} \right),
\]

where \( T_{\nu}(\cdot) \) denotes the cdfs of the t-distribution with \( \nu \) degrees of freedom. The t-distribution is outside of the exponential family and has no UMP test, so this confidence distribution is not uniformly optimal either. It is exact though, in terms of coverage, and we can still do comparisons with the Bayesian approaches.

Deriving a reference prior in this case is not difficult, but the calculations are tedious. Luckily, the grunt work has been done for us by [Kim et al. (2006)]. Working in the same fashion as we did in the last paragraph, based on the same orthogonal parametrization, they derive the block-diagonal Fisher information matrix as;

\[
I(\psi, \theta_2, \ldots, \theta_m, \sigma^2) = \begin{bmatrix}
I_{11} & 0^T & 0 \\
0 & I_{22} & 0 \\
0 & 0^T & I_{(m+1)(m+1)}
\end{bmatrix},
\]

where

\[
I_{11} = \frac{1}{\sigma^2 \sum_{j=1}^{m} c_j^2 / n_j}, \quad I_{(m+1)(m+1)} = \frac{N}{2\sigma^4},
\]

and

\[
I_{22} = \frac{1}{\sigma^2} \begin{bmatrix}
a_{22} & a_{23} & \cdots & a_{2m} \\
a_{32} & a_{33} & \cdots & a_{3m} \\
\vdots & \vdots & \ddots & \vdots \\
a_{m2} & a_{m3} & \cdots & a_{mm}
\end{bmatrix}, \text{ with } a_{ii} = n_1 c_i^2 c_i^{-1} + n_i, \ a_{ij} = n_1 c_i c_j c_i^{-1} c_j^{-1}.
\]

From this, we can derive Jeffreys’ overall prior as

\[
\pi_J(\psi, \theta_2, \ldots, \theta_m, \sigma^2) = (\det I)^{1/2} = \left( \det I_{11} \det I_{22} \det I_{(m+1)(m+1)} \right)^{1/2} \propto \left( \frac{1}{\sigma^2} \frac{1}{\sigma^2 (m-1)} \frac{1}{\sigma^4} \right)^{1/2} = \frac{1}{\sigma^{m+2}}.
\]
For the reference priors, using theorem 3.2 will yield different priors depending on the parameter grouping. Keeping with the focused theme we’ve been pursuing, where our parameter of interest is most important; two groupings seem natural to consider:

\[ \{ \psi, \{ \theta_2, \ldots, \theta_m, \sigma^2 \} \} \quad \text{and} \quad \{ \psi, \{ \theta_2, \ldots, \theta_m \}, \sigma^2 \}. \]

Both grouping consider \( \psi \) to be the most important parameter, but they differ in how important the variance is. In the first grouping, the variance is considered of equal inferential importance as the other mean parameters; \( (\theta_2, \ldots, \theta_m) \). The second grouping considers the variance of least importance, less so than the other \( \theta_2, \ldots, \theta_m \).

For the first grouping, Theorem 3.2 yields

\[
\pi_{R1}(\psi, \theta_2, \ldots, \theta_m, \sigma^2) = \prod_{i=j}^2 h_{j1}(\theta_j)^{1/2} = 1^{1/2} \left( \det I_2 \det I_{(m+1)(m+1)} \right)^{1/2} \propto \frac{1}{\sigma^{m+1}},
\]

while the reference prior based on the second grouping is

\[
\pi_{R2}(\psi, \theta_2, \ldots, \theta_m, \sigma^2) \propto \frac{1}{\sigma^2}.
\]

We see that the reference prior changes quite a bit depending on how the parameters are grouped together. The last prior is interesting because, in the grouping where the variance is considered least important, it is independent of the number of populations \( m \). Kim et al. (2006) proves many properties for priors on the form \( \pi(\cdot) \propto \sigma^{-2a} \) for \( a > 0 \) in this problem. They prove that the posterior distribution is proper if \( N + 2a - m - 2 > 0 \), and give a formula for the marginal posterior distribution of \( \psi \) under any such \( a \)-prior. They also prove that this marginal distribution is symmetric and unimodal around \( \hat{\psi} \) – just like the CD in 4.37.

The marginal posterior distribution of \( \psi \), resulting from an \( a \)-prior is of the form;

\[
\pi(\psi | y) \propto \left\{ \sum_{j=1}^m \left[ S_j + n_j \left( \bar{y}_j - \psi \frac{c_j/n_j}{\sum_{i=1}^m c_i/n_i} \right)^2 \right] - B(\psi, n, c) \right\}^{-(N+2a-m-1)/2},
\]

where \( S_j \) is the within-group sum of squares, and \( B(\cdot) \) is a function depending on the constants \( c_j \), the samples sizes \( n_j \) and the data \( y \). R code for computing this marginal posterior is included in appendix B.

In figure 4.10, the confidence curve based on (4.37) and credible curves based on marginal posteriors from Jeffreys’ overall prior and the two reference prior is shown. The true underlying variance is \( \sigma_0 = 2 \), the number of groups \( m = 5 \) and the linear combination is \( c = (1, 1, 1, 1, 1) \). Inference is not particularly sensitive to these choices, but will vary with \( n = (n_1, \ldots, n_5) \). In the left panels, computations are made with a random sample from a balanced design with \( n = (3, \ldots, 3) \). We see that all four curves are close together, indicating good frequentist properties.
4.3. LINEAR COMBINATION OF NORMAL MEANS

4.3.1 Behrens-Fisher

As a final note on the problem at hand, let’s look at the Behrens-Fisher problem that we briefly touched upon while discussing simultaneous fiducial inference. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$, where all $X_i$ are independent of all $Y_i$, and consider the two variances unknown and (possibly) unequal. The goal is to obtain inference for the parameter of interest $\psi = \mu_2 - \mu_1$ — simple enough.

This problem is notoriously hard, and no perfect solution really exists, at least not one that can be universally agreed upon. This is troubling, especially since it is such a common situation in statistics to compare two normal means in this fashion.

Let’s first consider Fisher’s solution to the problem, ignoring the discrepancy pointed out by Bartlett. Fisher works from the familiar pivotal quantities

$$t_1 = \frac{\sqrt{n}(\mu_1 - \bar{x})}{s_1} \quad t_2 = \frac{\sqrt{m}(\mu_2 - \bar{y})}{s_2}$$
which are independent t-distributed random variables with degrees of freedom \((n-1)\) and \((m-1)\), respectively. This means that \(\mu_1\) and \(\mu_2\) can be written as

\[
\begin{align*}
\mu_1 &= \bar{x} + \frac{s_1}{\sqrt{n}}T_1 \\
\mu_2 &= \bar{y} + \frac{s_2}{\sqrt{m}}T_2,
\end{align*}
\]

where \(T_1\) and \(T_2\) are t-distributed random variables with \((n-1)\) and \((m-1)\) degrees of freedom, respectively. This means that \(\mu_1\) and \(\mu_2\) are independently fiducially distributed as non-standard t-distributions with location-scale parameters \((\bar{x}, s_1/\sqrt{n})\) and \((\bar{y}, s_2/\sqrt{m})\), respectively.

Interest now is on the parameter \(\psi = \mu_2 - \mu_1\), which yields

\[
\psi = \mu_2 - \mu_1 = \bar{y} - \bar{x} + \frac{s_2}{\sqrt{m}}T_2 - \frac{s_1}{\sqrt{n}}T_1.
\]

Letting \(\hat{\psi} = \bar{y} - \bar{x}\), and considering \((s_1/\sqrt{n}T_1, s_2/\sqrt{m}T_2)\) as a point in 2-dimensional Euclidean space, we can rewrite it in polar coordinates to obtain the statistic;

\[
\frac{\psi - \hat{\psi}}{\sqrt{s_1^2/n + s_2^2/m}} = T_2 \cos \theta - T_1 \sin \theta,
\]

where \(\tan \theta = (s_1/\sqrt{n})/(s_2/\sqrt{m})\).

We see that the statistic on the left side of the equation is distributed as the convolution of two t-distributions. Such convolutions are hard to express in closed form, in fact closed form expressions only exist when both degrees of freedom are odd \(^{\text{Nadarajah & Dev 2005}}\). \(^{\text{Rahman & Saleh 1974}}\) gives a power-series expression for this distribution, and there are R-packages available that will compute it, usually by numerical integration\(^{10}\). Fittingly, this distribution is called the Behrens-Fisher distribution.

If we let \(B_{n,m,\theta}(\cdot)\) denote the cumulative distribution function of the Behrens-Fisher distribution with parameters \(n, m\) and the angle \(\theta\), we could hope that a construction of the kind

\[
C(\psi) = B_{n,m,\theta} \left( \frac{\psi - \hat{\psi}}{\sqrt{s_1^2/n + s_2^2/m}} \right),
\]

might be an approximate confidence distribution. To test this hypothesis, remember the crux in the definition a CD given back in chapter 2\(^2\). Under the true parameter, \(C(\psi_0)\) should be uniformly distributed. By drawing samples from different combinations of \((n, m, \psi, \sigma_1, \sigma_2)\) and evaluating the CD at the true value, we can obtain samples from \(C(\psi_0)\). We can then estimate a kernel density and determine its uniformity.

The result of such a simulation is displayed in figure 4.11 where density estimates from \(N = 100 000\) samples are displayed, based on the beta-kernel density estimator of \(^{\text{Chen 1999}}\) on the bounded interval [0, 1]. This simulation was repeated for different values of parameters (columns), and different samples sizes (rows), yielding the array in the figure. We immediately recognize a problem; for low \((n, m)\) this

\(^{10}\)for example, the package \texttt{asht} available at https://CRAN.R-project.org/package=asht
confidence distribution is not uniform, meaning that any $\alpha$-level interval will not have $\alpha$ coverage probability. Interestingly, uniformity also seems to depend on the variances $(\sigma_1^2, \sigma_2^2)$, where even for low sample sizes uniformity can be somewhat achieved, if the variances are large and unbalanced.

For larger sample sizes, such as $(20, 20)$ we see that uniformity is achieved, and thus the confidence distribution in (4.42) is a decent approximation to correct coverage for moderate sample sizes. It would seem that the damning evidence provided against Fisher’s fiducial solution to this problem is not so damning after all.

The critique levied against Fisher’s above solution and argumentation, concerned itself with testing the hypothesis $H_0 : \psi = 0$, for which an $\alpha$-level test won’t have type-I error probability at the specified level. With this in mind, Welch (1938) proposed a solution based on the same statistic as Fisher, but using a $t$-distribution instead of the Behrens-Fisher distribution. Welch suggested that the statistic in (4.41) approximately has a $t$-distribution with $\nu$ degrees of freedom, where

$$\nu = \frac{(S_1^2/n + S_2^2/m)^2}{S_1^4/(n^3 - n^2) + S_2^4/(m^3 - m^2)}.$$

This distribution has excellent properties for testing the hypothesis $H_0$, no matter the true underlying parameters, and does a decent job even at low sample sizes, see for example Best & Rayner (1987) for a simulation study.

While the confidence distribution can be viewed as a collection of $p$-values, of which the one from testing this particular $H_0$ is only one in many, one might conjecture that a CD based on Welch’s proposal would do quite well as an approximate
confidence distribution for low samples. As a test, I redid the simulations with this t-distribution instead of the Behrens–Fisher distribution and checked for uniformity. The results are displayed in figure 4.12, and we see that, even for small sample sizes, \(C(\psi_0)\) is relatively uniform, more so than the fiducial solution.

While this is interesting, I feel the results of the two small simulations speak more to the fiducial solution, which achieves approximately uniformity, and hopefully this the correct coverage probability, even for moderately small sample sizes. Both confidence distributions are approximately uniform at equal sample sizes of 20, but the Welch approximation seems to have some peaks in its density at the edges for the unbalanced samples \(n = 2\) and \(m = 20\). This is not a fluke of simulation, as I’ve redone them several times and gotten similar results, for other sample sizes and parameters as well. It would seem that the convergence towards uniformity is not as smooth as with the fiducial distribution, though I’m not sure what this could mean for inference.

Fisher provided an approach to epistemic inference, in which his results often agree approximately with the strict frequentist criteria of his day. Simulation studies by for example Robinson (1976) indicates that approximate coverage is obtained quickly, and that at least conservative coverage is achieved, i.e that an \(\alpha\)-level interval has at least \(\alpha\)-level coverage. In addition, the construction is much more general than the one of Welch. Again, I feel that his argument would not have been shot down so quickly if these simulation tools were available in his time.

As a final take, consider the Bayesian solutions to the problem at hand. The likelihood can be written as

\[
L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \propto \frac{1}{\sigma_1^n \sigma_2^m} \exp \left\{ -\frac{1}{2} \left[ \frac{S_1^2 + n(\bar{x} - \mu_1)^2}{\sigma_1^2} + \frac{S_2^2 + m(\bar{y} - \mu_2)^2}{\sigma_2^2} \right] \right\},
\]
where \( S_i^2 \) denotes the sum-of-squares in each sample. Making the reparametrization to \( \psi = \mu_2 - \mu_1 \) and \( \lambda = \mu_1 + \mu_2 \), the Fisher information can be written as

\[
I(\psi, \lambda, \sigma_1^2, \sigma_2^2) = \frac{1}{2} \begin{bmatrix}
\frac{n}{2\sigma_1^2} & 0 & 0 & 0 \\
0 & \frac{m}{2\sigma_2^2} & 0 & 0 \\
0 & 0 & \frac{n}{\sigma_1^4} & 0 \\
0 & 0 & 0 & \frac{m}{\sigma_2^4}
\end{bmatrix}.
\]

Now then, Jeffreys’ overall prior can be seen to be \( \pi_J(\psi, \lambda, \sigma_1^2, \sigma_2^2) \propto (\sigma_1\sigma_2)^{-3} \). Using theorem 3.2, we have the reference prior, under any other grouping than that which yields Jeffreys’, that \( \pi_R(\psi, \lambda, \sigma_1^2, \sigma_2^2) \propto (\sigma_1\sigma_2)^{-2} \). As an interesting side note, Jeffreys (1961) proved that the prior \( \pi^*_J(\psi, \lambda, \sigma_1^2, \sigma_2^2) \propto (\sigma_1\sigma_2)^{-1} \), which one might consider to be a quite natural prior since both are scale-parameters, will actually lead to the fiducial solution.

In figure 4.13, results from two samples are displayed as credible-/confidence curves based on Fisher’s fiducial solution, \( C_f \), Welch’s approximate solution, \( C_w \), and marginal posterior distributions based on \( \pi_J \) and \( \pi_R \). In the left panel, 5 samples are drawn from \( X \) and \( Y \), with \( (\mu_1, \mu_2, \sigma_1, \sigma_2) = (2, 4, 2, 4) \), while the right panel has 10 samples from each distribution with the same means, but more unbalanced standard deviations \( (\sigma_1, \sigma_2) = (1, 9) \).

For the left panel, Jeffreys’ prior leads to the shortest intervals, followed by the Reference prior, then Welch’s approximation, and finally Fisher’s fiducial solution. This corroborates the results by Robinson that the fiducial solution is conservative. That being said, the curves are quite close close together, and one would need a high degree of confidence to obtain large discrepancies between the lengths of the intervals.

In the right panel, the fiducial solution and Welch’s approximation is virtually indistinguishable from each other, indicating that the fiducial solution works well in obtaining close to the correct coverage. The two posterior distributions also seem closer together in this plot, but are both narrower than the frequentist solutions, with maybe Jeffreys’ posterior being a little narrower than the one based on the reference prior – indicating approximately the correct coverage.
The goal of this example was simply to highlight how the frequentist and fiducial approaches performed, and compare the solutions with some objective Bayesian solutions. From a Bayesian point of view, it has been studied quite detailed by Ghosh & Kim (2001), who also explore the use of probability matching priors for this problem, and find excellent coverage even for low $n$.

While Fisher may not have cared to much about the coverage probabilities of his solution, considering the resulting distribution simply as one of epistemic probability, it could be interesting to pursue the idea of the Behrens-Fisher solution being an approximate confidence distribution further, as it is a rather open question on the literature. Simulations pointed to earlier seem to indicate that the Behrens-Fisher intervals are conservative, which should be deemed an admissible form of inference. I mean, as long as a 90% interval has coverage probability larger than, or equal to .9, it seems safe to use in practice. It would be much worse if the interval in question only had a 30% chance of covering the true parameter.

### 4.4 The Fieller-Creasy problem

The Fieller-Creasy problem is a famous problem in statistical inference, and it is included here as an example where frequentist and Bayesian inference give two distinctly different answers, and comparisons are hard, or impossible.

The problem can be stated in the following manner. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, where all $X_i$ and $Y_j$ are independent, and the variance unknown. The problem at hand then, in this relatively simple setup, is to estimate the ratio $\psi = \mu_2 / \mu_1$.

This problem has a long history in statistical inference, starting from the original solution by Fieller (1940), and further analysis by Creasy (1954).

Given the data as above, the statistics $\bar{X}$ and $\bar{Y}$ are distributed as

$$
\begin{pmatrix}
\bar{X} \\
\bar{Y}
\end{pmatrix} \sim N_2 \left( \begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}, \frac{\sigma^2}{n} I_2 \right)
$$

where $I_2$ denotes the $2 \times 2$ identity matrix. There is a nice geometry to this problem; the pair $(\bar{X}, \bar{Y})$ can be considered a point in 2-dimensional euclidean space, and we can make the transformation into polar coordinates by

$$
\bar{X} = R \cos \theta \quad \bar{Y} = R \sin \theta
$$

for random variables $R \in \mathbb{R}$ and $\theta \in [-\pi/2, \pi/2)$. The expectations $(\mu_1, \mu_2)$ transforms to $(r_0 \cos \theta_0, r_0 \sin \theta_0)$ for underlying true parameters $r_0$ and $\theta_0$. Then since

$$
\psi = \frac{\mu_2}{\mu_1} = \frac{r_0 \sin \theta_0}{r_0 \cos \theta_0} = \tan \theta_0,
$$

questions concerning $\psi$ amounts to questions concerning the angle $\theta_0$.

Next, make the following transformation of the statistics

$$
\begin{pmatrix}
Z_1 \\
Z_2
\end{pmatrix} = \begin{pmatrix}
\cos \theta_0 & \sin \theta_0 \\
-\sin \theta_0 & \cos \theta_0
\end{pmatrix} \begin{pmatrix}
\bar{X} \\
\bar{Y}
\end{pmatrix}.
$$
4.4. THE FIELLER-CREASY PROBLEM

This transformation amounts to a clockwise rotation in 2-dimensional space, and it is easy to check that

\[
\begin{pmatrix}
Z_1 \\
Z_2
\end{pmatrix} \sim N_2 \left( \begin{pmatrix}
r_0 \\
0
\end{pmatrix}, \frac{\sigma^2}{n} I_2 \right),
\]

and that the corresponding polar coordinates for \( Z_1 \) and \( Z_2 \) are by \( R \cos (\theta - \theta_0) \) and \( R \sin (\theta - \theta_0) \) respectively. Letting \( S_p^2 \) denote the pooled variance as in (4.36), and putting \( \tilde{R} = \sqrt{nR/S_p} \), the following quantity is a familiar pivot;

\[
\tilde{R}^2 \sin^2 (\theta - \theta_0) = \left( \frac{Z_2}{S_p/\sqrt{n}} \right)^2 = \left( \frac{\sqrt{n}Z_2/\sigma}{S_p/\sigma} \right)^2 \sim F_{1,n-1},
\]

namely \( F \)-distributed with \((1, n-1)\) degrees of freedom.

This invites the construction of \( \alpha \)-level confidence sets by

\[
C_\alpha = \left\{ \theta_0 : \tilde{R}^2 \sin^2 (\theta - \theta_0) \leq F_{1,n-1}^\alpha \right\},
\]

where \( F_{1,n-1}^\alpha \) is \( \alpha \) quantile of the \( F \) distribution. A confidence distribution is not available though, since this pivot is not monotone in \( \theta_0 \), and it has some strange properties. Noticeably, if \( \tilde{R}^2 \leq F_{1,n-1}^\alpha \) then \( C_\alpha = \mathbb{R}! \).

Consider, after an experiment has been done, that the statistics \((\bar{x}, \bar{y}, s_p)^\top = (1.6014, 2.7514, 1.721)\) are obtained from 4 observations. The corresponding statistics \( R \) and \( \theta \) is by

\[
R = \sqrt{\bar{x}^2 + \bar{y}^2} = 3.1835 \quad \theta = \arctan \frac{\bar{y}}{\bar{x}} = 1.0437,
\]

which makes \( \tilde{R}^2 = 13.69 \). This implies that all confidence intervals at level \( \alpha = F_{1,n-1}^{-1}(13.69) = 0.9657 \) or higher will be the entire real line. Other possibilities are that the confidence set is an interval, or the complement of an interval, the situation is illustrated in figure 4.14.

In the figure, the pivot (4.43) is plotted from \( \psi \in (-10, 10) \) corresponding to \( \theta_0 \in (-1.47, 1.47) \), the confidence sets consists of the parts of the parameter space where the pivot dips under the required \( F \)-quantiles. We see that the 50% confidence set is in fact an interval, while the 90% is the complement of an interval. All sets with larger confidence than 96.57% consists of the entire real line, as stated. Also added to the plot, at its lowest point, is a tick mark at the natural point estimator for \( \psi \), namely \( \bar{y}/\bar{x} \).

This example is sometimes cited as a setup where frequentist inference is nonsensical. What are we to make of the (exact) confidence intervals obtained from this setup? Surely a confidence interval consisting of the entire real line must have 100% confidence? Are there then levels of confidence, between 1 and for example 0.9657 for which we really can’t say anything, where the data doesn’t allow confidence statements? This is the view of Schweder & Hjort (2016) and Cox (2006), and is also my own. There are simply situations, where the data obtained from an experiment does not carry enough information about \( \psi \) to make precise confidence statements at all levels of confidence, there is nothing inherently wrong with this, although it might seem strange at first glance.
From the Bayesian point of view, a reference prior is available, with a proper marginal posterior distribution. Yang & Berger (1998) give the reference prior for this problem as

$$\pi(\psi, \mu_1, \sigma^2) \propto \sigma^{-2}(1 + \psi^2)^{-1/2},$$

and the marginal posterior by

$$\pi(\psi|\mathbf{x}, \mathbf{y}) \propto (1 + \psi^2)^{-1} \left( S^2 + \frac{n(\bar{y} - \psi \bar{x})^2}{1 + \psi^2} \right)^{-(2n-1)/2},$$

where $S^2 = \sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2$.

Plugging in our data from previous, we obtain the marginal posterior distribution in figure 4.15, where the dotted vertical line represents the estimated $\psi = \bar{y}/\bar{x}$ as before. The distribution in this case is unimodal, centered a little to the right of the frequentist point estimate.

It is difficult to compare the two plots in figures 4.14 and 4.15, as they differ both in form and in logical content, the pivot simply giving rise to confidence regions in the parameter space, while the marginal posterior distribution a bona fide probability distribution – albeit based on an ‘objective prior’. The reference posterior distribution gives rise to highest posterior credible intervals that are nicely defined, for all $\alpha \in (0, 1)$, in sharp contrast to the frequentist situation. Liseo (1990) studied the coverage probabilities of posterior quantiles based on this distribution, and found that they often are pretty close to the frequentist coverage probability.

Both methods depend heavily on the location of $(\mu_1, \mu_2)$ in the parameter space. A $\mu_1$ close to zero is likely to make the ratio $\bar{y}/\bar{x}$ unstable, resulting in poor performance of both the Bayesian and frequentist solutions. There seems to be no quick fix from the Bayesian point of view either, with inferences being crucially dependent on the samples obtained.
Figure 4.15: The reference marginal posterior distribution for $\psi$ in the Fieller-Creasy problem.
Chapter 5
Concluding remarks

In this thesis, I have looked at two vastly different approaches to statistical inference; Fisher’s fiducial argument, specifically focusing on the modern version in terms of confidence distributions, and the theory of objective Bayesianism. Though the theories are different in approaches and underlying logic, they are trying to achieve the same goal, and they often give quite similar results. In the examples looked at, sometimes we were able to achieve an exact match between the marginal Bayesian posteriors, and an optimal confidence distribution, i.e. one that is uniformly most powerful while still admitting the correct coverage. In all cases considered, exact matching was achieved with Jeffreys’ prior, and typically the reference prior would either equal Jeffreys’, or they would yield the same posterior distributions. Outside the context of exactly matching an optimal CD, the reference prior approach seems to outperform Jeffreys’, when they give different posteriors. Looking at linear-combinations of normal means, the reference prior approach is superior to Jeffreys’ when comparing with the CD based on either the t-distributed pivot, or the Behrens-Fisher distribution, this is most likely due to a better approach to handling nuisance parameters.

The goal of the thesis was not to solve any new problems of fiducial inference, but to highlight connections between fiducial inference and the ‘objective’ Bayesian inference. Fisher intended his fiducial distribution to be an objective epistemic probability distribution on the parameter space, allowing definite probability statements about the parameter – without appealing to Bayes’ theorem. Comparing the distributions from fiducial inference, represented by the confidence distribution approach, to the most compelling Bayesian solution for objective inference, this thesis provides a few answers, but leaves many left to solve.

There are themes in the thesis that one might want to pursue further, either building directly upon the work done here, or completely new directions. In the following I will indicate some of these.

5.1 Exact matching and uniformly optimal CDs

We’ve seen that exact matching between Bayesian posterior distributions and joint fiducial distributions, is only possible in a few limited cases. Within the exponential family, this problem seems to be relatively closed due to the findings by (Veronese & Melilli [2017], reviewed in section 4.2.1). While their results establishes conditions for exact matching of joint distributions, there is no guarantee that the marginal
distribution for a parameter of interest $\psi$, matches the uniformly optimal CD. It would be of interest to close this problem as well, perhaps by putting some further conditions on the prior distribution $\pi(\cdot)$.

Veronese & Melilli (2017) worked within the cr-NEF, i.e. members of the natural exponential family whose density can be written on the form

$$p_\theta(x) = p_\theta(T_1)p_\theta(T_2|T_1)p_\theta(T_3|T_1, T_2) \cdots p_\theta(T_d|T_1, \ldots, T_{d-1}),$$

where each conditional density is also a member of the NEF. Constructing a joint fiducial distribution from this, by converting each term into a fiducial distribution, and taking the product, this joint distribution will coincide with the Bayesian posterior distribution based on the reference prior. Now then, since each fiducial distribution in the product above, is also a confidence distribution, the fiducial distribution constructed from $p_\theta(T_d|T_1, \ldots, T_{d-1})$ will also be uniformly optimal by the theory in chapter 2.

Let $\{\theta_1, \ldots, \theta_d\}$ denote the natural parameters corresponding to the jointly sufficient statistics $\{T_1, \ldots, T_d\}$ from above, so that a CD constructed from $p_\theta(T_d|T_1, \ldots, T_{d-1})$ is uniformly optimal for $\theta_d$. The marginal posterior distribution for $\theta_d$ is

$$p(\theta_d|x) \propto \int_{\theta_{d-1}} \cdots \int_{\theta_1} \pi_R(\theta_1, \ldots, \theta_d)p_\theta(x)d\theta_1 \cdots d\theta_{d-1}$$

$$= p_\theta(T_d|T_1, \ldots, T_{d-1})M(\theta_d, T_1, \ldots, T_{d-1}),$$

where the function $M$ is meant to capture what’s left over from integrating away the parameters $\theta_1, \ldots, \theta_{d-1}$. The point now is that this marginal posterior distribution will only correspond to the uniformly optimal confidence distribution, if the distribution of $T_d|T_1, \ldots, T_{d-1}$ is a member of the correct families and the function $M$ is in fact proportional to the prior ensuring exact coverage. For a natural exponential family with density proportional to

$$\exp \{\theta T(x) - K(\theta)\},$$

the only distributions who admit such a prior distribution is again; Normal with known variance, Gamma with known shape, binomial, Poisson and negative-binomial. In addition, the prior $\pi$ that ensures exact coverage is the flat prior for the normal location parameter, and $\pi \propto K'(\theta)$ for the others, where $K'$ denotes the derivative of $K$, see Veronese & Melilli (2015) for a proof of this statement.

To summarize then, we would need $p_\theta(T_d|T_1, \ldots, T_{d-1})$ and the function $M(\theta_d, T_1, \ldots, T_{d-1})$ to satisfy some rather strong conditions in order to have exact matching between uniformly optimal confidence distributions and marginal posteriors.

It would be of interest to investigate in which cases this happens, and what restrictions must be in place to ensure it.

5.2 Approximate matching and PMPs

In section 3.6 I briefly covered some of the basic results on posterior matching. That is, to what order does the quantiles of a Bayesian posterior distribution match
the frequentist coverage probability. The classical result is by Welch & Peers (1963), who proved that all continuous priors will yield first-order matching posteriors, while Jeffreys’ prior will yield second-order matching posteriors. In some special cases, with further restrictions on the likelihood, Jeffreys’ prior will result in a third-order matching posterior, but that is pretty much it besides the exact cases mentioned in the preceding section.

One major topic in this direction, that there simply wasn’t room to explore, is probability matching priors. Prior distributions that are specifically tailored to achieve posterior matching up to some intended order. I briefly mentioned them in section 3.6.1 but it would have been interesting to explore this topic as well. Some of the examples considered already have a great deal of theory in this direction laid out, such as the Behrens-Fisher problem, and the more general linear combinations of normal means. Quite often, the reference prior approach will actually yield first- or second-order PMPs. Notably, when the parameters are orthogonal, as when utilizing theorem [3.2] reference priors often are PMPs.

If \( \psi \) is some scalar, parameter of interest, and \( (\lambda_1, \ldots, \lambda_p) \) is a vector of nuisance parameters, and \( i_{ij} \) is the \((i, j)\)-th element of a diagonal information matrix \( I(\psi, \lambda_1, \ldots, \lambda_p)^{-1} \). Then, any prior distribution satisfying

\[
\pi(\psi, \lambda_1, \ldots, \lambda_p) \propto i_{11}^{-1/2} h(\lambda_1, \ldots, \lambda_p),
\]

where \( h(\cdot) \) is an arbitrary function, is a second-order PMP for \( \psi \) (Datta & Ghosh 1995a). With a third-order PMP sometimes within reach by how the function \( h(\cdot) \) is chosen.

It could be interesting to compare PMPs to the uniformly optimal CDs, as they aim more directly for optimal frequentist properties.

## 5.3 Paired Exponentials

In section 4.1 I considered the problem of estimating \( \delta \) in a setup where we observe exponential pairs \((X_i, Y_i)\) with \( X_i \sim \text{Expo}(\theta) \) and \( Y_i \sim \text{Expo}(\theta + \delta) \). The optimal confidence distribution for this parameter is readily found without much trouble, but the Bayesian approach is difficult. I derived two different reference priors

\[
\pi_R(\delta, \theta) \propto \delta^{-4+\frac{2}{\sqrt{3}}} \left( \frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2} \right)^{1/2},
\]

\[
\pi_R^*(\delta, \theta) \propto \delta^{-1} \left( \frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2} \right)^{1/2},
\]

based on how the sequence of compact sets were chosen. \( \pi_R \) is based on compact rectangles in the original parametrization, using the algorithm sketched out by Berger & Bernardo (1989). The reference prior \( \pi_R^* \) is based on compact rectangles in the orthogonal parametrization of \((\psi, \lambda)\), and used the theorem of Datta & Ghosh (1995b).

The performance of the posterior distribution based on these two different priors is puzzling. \( \pi_R \) will yield posterior distributions that seemingly always are proper, with performance comparable, or slightly below, that of Jeffreys’ prior in this problem. \( \pi_R^* \) seems to sometimes yield posterior distributions on par with Jeffreys’ prior, but other times the posterior distributions do not appear to be proper, even for
moderately large \( n \). Consider the statistics \( A = \sum (x_i + y_i) \) and \( S = \sum y_i \), from \( n \) observations. This makes the posterior distribution

\[
p(\delta, \theta | A, S) \propto \delta^a \left( \frac{1}{\theta^2} + \frac{1}{(\theta + \delta)^2} \right)^{1/2} \theta^n (\theta + \delta)^n \exp \{-\delta A - \theta S\},
\]

with \( a = -1 \) for \( \pi_R \) and \( a = -4 + 5/\sqrt{2} \) for \( \pi_R^* \).

Depending on the sampled values of \((A, S)\), \( \pi_R^* \) can yield results on par with Jeffreys’ prior, or results that are non-sensical, perhaps improper. In figure 5.1, the marginal posterior distributions are displayed, computed for the two reference priors, and Jeffreys’ prior – at 4 different values of the sampled statistics. The value of \( A \) is kept fixed at 4, while \( S \) changes from 2 in the top left, 1 in the top right, 0.5 in bottom left and 0.25 in the bottom right, all at \( n = 10 \).

As \( S \) becomes smaller, the posterior distribution based on \( \pi_R^* \) creeps closer and closer to Jeffreys’, which we’ve seen has good properties in this example. The same happens for \( \pi_R \), but it seems to be shifted to the right compared to the two others. Comfortingly, the posterior based on \( \pi_R \) doesn’t keep traveling right for even lower values of \( S \), and at larger values of \( n \), all three posterior distributions look the same.

This behaviour is quite strange, and I have not been able to find a good explanation. The reparametrization I did in order to find the reference prior using the algorithm of [Berger & Bernardo] (1989), leaves a gap in the parameter space at \( \delta = 0 \) where the parametrization is not one-to-one – this is most likely the source of the discrepancy, and it might make the theorem by [Datta & Ghosh] (1995b) invalid, though I haven’t found anything in the paper to indicate that this is so. The 1989 paper by Berger and Bernardo finds the reference prior for the product of two normal means, under the restriction that both means are strictly larger than zero – similar to the restriction I enforced. And my proofs and techniques are following
their approach. A conjecture might be that the value \(-4 + 5/\sqrt{2} \approx -0.46\) is as close to \(-1\) as one can get, while still maintaining a proper distribution on average. It becomes a question of choosing between a prior that on average performs well, \(\pi_R\), or a prior that sometimes performs well, and sometimes quite horribly.

The problem was included in the thesis, as an example of how difficult the theory of 'objective' Bayes can be, this is particularly interesting when contrasted to the CD solution, which is so simple. It would be of interest to derive the reference prior in this problem without restriction in place, hopefully we should see similar results compared to the other problems, where the reference prior seemed to be slightly better than Jeffreys’ prior on average over the sample space.

5.4 Behrens-Fisher

In section 4.3.2, the Behrens-Fisher problem was touched upon briefly. This is a problem of immense historical, and practical importance, and could probably warrant an entire thesis of its own; exploring suggested solutions from different angles and generalizations. The main question I tried to indicate is whether or not Fisher’s fiducial solution is a confidence distribution, exactly, approximately, or conservatively – i.e. \(\alpha\)-level quantiles have at least \(\alpha\)-level coverage. Some initial simulations seem to indicate that Fisher’s solution is an approximate confidence distribution, close to uniformity even for moderately small sample sizes, such as \(n = 10\). To answer this question more fully would require either a large simulation study, with some measure of closeness to pure uniformity, perhaps the KL-divergence, or maybe some asymptotic expansions for a more analytic proof. The studies that have been done seem to suggest that Fisher’s solution is at least a conservative confidence distribution, though it remains an open question.

5.5 Epistemic probability

The fiducial argument rests on transferring aleatory probability from the sample space, into the parameter space. Once we are in the parameter space, the good old rules for probability calculus no longer seem to apply, at least not under the desideratum that exact coverage should be conserved. Clearly, once the randomness has been transferred, it no longer adheres to the axioms laid down by Kolmogorov – which is troublesome. First, and foremost, it means we need to be clear in our thought process when working fiducially, paying attention to the underlying logic, and checking resulting distributions as we go along.

Fisher regarded his fiducial distribution as a proper epistemic posterior distribution on the parameter space, regarding the previously fixed parameter of interest, now as a random variable. I will agree with most of the critics that this is in fact incorrect, and continue to regard my parameters as fixed unknown quantities. If the fiducial distribution is also a confidence distribution, I am more willing to call the resulting distribution epistemic in a sense, but not go so far as to consider the parameter as a random variable. I do feel that the confidence distribution in a sense captures something more akin to this interpretation of probability. This seems also to be the view of Schweder (2017), who argues that confidence is in fact epistemic probability!
Hampel (2003) regards the confidence distributions as the only proper fiducial argument, as it was laid out in Fisher’s early papers – before he started regarding the parameters as random. Hampel thinks that “... one of the reasons why we still have problems with fiducial probabilities, is that we lack an adequate formalism for frequentist epistemic probabilities.” (Hampel 2003, p. 8, my emphasis) Perhaps Fisher’s fiducial argument is a new type of probability, with it’s own set of rules – we simply need to find out what they are. Perhaps we need a new set of axioms, tailored to this specific system of logic. In this regard, see also Schweder & Hjort (2016, ch. 15.4) and Schweder & Hjort (2017).
Bibliography


Liseo, B. (1990), Elimination of nuisance parameters with reference non informative priors, Technical Report 90-58C, Purdue University.


Appendix A

Proofs

A.1 Proof of Lemma 4.1

I will restate the Lemma to make it clear where we are heading:

**Lemma 4.1.** Let

\[
I_i^1(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} \log \left( \frac{\lambda - 1}{\lambda} \right) \, d\lambda
\]  

(4.12)

and,

\[
K_i^{-1}(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{1 + \lambda^2}}{\lambda^2} \, d\lambda.
\]  

(4.13)

These integrals equals,

\[
I_i^1(\psi) = \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) + \frac{\psi}{k_i} - \log \left( \frac{\psi}{k_i} \right) \right) + B_7,
\]  

(4.14)

\[
K_i^{-1}(\psi) = \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} - 2 \log \left( \frac{\psi}{k_i} \right) + \sqrt{2} \log (1 + \psi/l_i) \right) + B_{10},
\]  

(4.15)

where $B_7$ and $B_{10}$ are bounded functions of $k_i$ and $l_i$. □

**Proof.** We start by noting that due to the chosen parametrization, the parameter of interest $\psi$ only crops up in the expressions at the edges of the integration area. Then, the integral taken over the interior of the integration area, is simply a constant independent of $\psi$. The main idea then is to approximate the integrand by simpler functions at the edges of the integration area, and take the integral over these instead.

I will first prove that:

\[
I_i^1(\psi) = \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) + \frac{\psi}{k_i} - \log \left( \frac{\psi}{k_i} \right) \right) + B_7
\]  

(A.1.1)

where $B_7$ is a bounded function of $k_i$ and $l_i$.

We begin by fixing a suitably small $\epsilon > 0$, and splitting the area of integration into

\[
(1 + \psi/k_i, 1 + \epsilon), \ (1 + \epsilon, 1 + \epsilon^{-1}) \text{ and } (1 + \epsilon^{-1}, 1 + \psi/l_i),
\]  

(A.1.2)
we can always arrange the sequences \( \{k_i\} \) and \( \{l_i\} \) such that these intervals exist.

Then rewrite the integrand in \( I_1(\psi) \) as \( I_1(\psi) = I_2(\psi) - I_3(\psi) \) where,

\[
I_2(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \sqrt{1 + \frac{1}{\lambda^2}} \frac{\log (\lambda - 1) \, d\lambda}{\lambda - 1} \tag{A.1.3}
\]

\[
I_3(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \sqrt{1 + \frac{1}{\lambda^2}} \frac{\log (\lambda) \, d\lambda}{\lambda - 1}. \tag{A.1.4}
\]

Let \( I_{21}(\psi) \), \( M_1 \), \( I_{22}(\psi) \) denote the parts of \( I_2(\psi) \) taken over the respective intervals \([A.1.2]\), and note that \( M_1 \) is simply a constant independent of \( l_i \) and \( k_i \). In the following \( M_i \) denotes constant terms while \( B_i \) denotes bounded terms.

By a simple Taylor expansion at \( \lambda = 1 \) we have

\[
\sqrt{1 + \frac{1}{\lambda^2}} = \sqrt{2} - \frac{\lambda - 1}{\sqrt{2}} + O((\lambda - 1)^2), \tag{A.1.5}
\]

in an interval near \( \lambda = 1 \). Here, the notation \( O(\cdot) \) is used to denote functions that are bounded on an interval near \( \lambda = 1 \). Specifically, if \( f(x) \) is \( O(x) \) near \( x = a \), there exists \( \delta \) and \( M \) such that

\[
|f(x)| \leq M|x| \text{ whenever } 0 < |x - a| < \delta.
\]

Multiplying by \( \log (\lambda - 1) \) and dividing by \( (\lambda - 1) \) will introduce a singularity at \( \lambda = 1 \), but the approximation still holds on an interval \([1, 1 + \delta]\). We have

\[
\sqrt{1 + \frac{1}{\lambda^2}} \frac{\log (\lambda - 1)}{\lambda - 1} = \frac{\sqrt{2} \log (\lambda - 1)}{\lambda - 1} - \frac{\log (\lambda - 1)}{\sqrt{2}} + \log (\lambda - 1)O(\lambda - 1)
\]

Now then

\[
I_{21}(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{1 + \lambda^2} \log (\lambda - 1)}{\lambda(\lambda - 1)} \, d\lambda = \int_{1+\psi/k_i}^{1+\psi/l_i} \frac{\sqrt{2} \log (\lambda - 1)}{\lambda - 1} - \frac{\log (\lambda - 1)}{\sqrt{2}} \, d\lambda + \int_{1+\psi/k_i}^{1+\psi/l_i} H_1(\lambda) \, d\lambda
\]

\[
= \frac{1}{\sqrt{2}} \left[ M_3 + (\psi/k_i) \log(\psi/k_i) - \psi/k_i - \log^2(\psi/k_i) \right] + B_1, \tag{A.1.6}
\]

where \( H_1 \) denotes the remainder, which is of the form \( \log (\lambda - 1)O(\lambda - 1) \). The integral of the remainder term \( (H_1) \) is a bounded function of \( k_i \), because, by the definition of \( O(\cdot) \) there exists a constant \( M_4 \) such that:

\[
\int_{1+\psi/k_i}^{1+\psi/l_i} |H_1(\lambda)| \, d\lambda \leq M_4 \int_{1+\psi/l_i}^{1+\psi/k_i} \left| \log (\lambda - 1) (\lambda - 1) \right| d\lambda = M_5 + \frac{M_4}{2} \log(\psi/k_i)(\psi/k_i)^2 + \frac{M_4}{4} (\psi/k_i)^2.
\]

As \( k_i \to \infty \) this expression simply becomes \( M_5 \) since the expressions involving \( k_i \) goes to zero.
Next we need to expand $\text{A.1.3}$ at $\lambda = \infty$. By putting $z = \lambda^{-1}$ and rewriting this expression as
\[
\left( \sqrt{1 + z^2} \frac{1}{\lambda} \left( \frac{1}{1 - z} \right) \right) \left[ \log(\lambda) + (\log(1 - z)) \right],
\]
we can use well known series expansions around $z = 0$ for the terms containing $z$ and obtain
\[
\sqrt{1 + \frac{1}{\lambda^2}} \log(\lambda - 1) = \left( 1 + O \left( \frac{1}{\lambda^2} \right) \right) \left( 1 + \frac{1}{\lambda} \right) \left( \log(\lambda) + O \left( \frac{1}{\lambda} \right) \right)
\]
\[
= \frac{\log(\lambda)}{\lambda} + O \left( \frac{1}{\lambda^2} \right),
\]
for large $\lambda$.

The integral becomes
\[
I_{22}(\psi) = \int_{1+\epsilon^{-1}}^{1+\psi/l_i} \frac{\log(\lambda)}{\lambda} d\lambda + \int_{1+\epsilon^{-1}}^{1+\psi/l_i} H_2(\lambda) d\lambda
\]
\[
= \frac{1}{2} \left[ M_6 + \log^2 (1 + \psi/l_i) \right] + B_2. \tag{A.1.7}
\]
Again, the integral of the remainder term is bounded because there exists $M_7$ such that:
\[
\int_{1+\epsilon^{-1}}^{1+\psi/l_i} |H_2(\lambda)| d\lambda \leq M_7 \int_{1+\epsilon^{-1}}^{1+\psi/l_i} \left| \frac{1}{\lambda^2} \right| d\lambda
\]
\[
= -M_7 \left[ \frac{1}{1 + \psi/l_i} - M_8 \right]
\]
\[
\rightarrow M_9 \text{ as } l_i \rightarrow 0.
\]

Putting $\text{A.1.6}$ and $\text{A.1.7}$ together we obtain
\[
I_2^0(\psi) = \int_{1+\psi/k_i}^{1+\psi/l_i} \sqrt{\frac{1}{\lambda^2} \log (\lambda - 1)} d\lambda
\]
\[
= \frac{1}{\sqrt{2}} \left( \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) - \frac{\psi}{k_i} - \log^2 \left( \frac{\psi}{k_i} \right) \right) + \frac{\log^2 (1 + \psi/l_i)}{2} + B_3. \tag{A.1.8}
\]

Next we will prove that :
\[
I_3^0(\psi) = -\sqrt{2} \frac{\psi}{k_i} + \frac{\log^2 (1 + \psi/l_i)}{2} + B_6. \tag{A.1.9}
\]
Again, we split the area integration like in $\text{A.1.2}$ and denote the integrals as $I_{31}^0(\psi)$, $M_{10}$ and $I_{32}^0(\psi)$. We again note that the integral over $(1 + \epsilon, 1 + \epsilon^{-1})$ is simply a constant, and use series expansions at $\lambda = 1$ and $1/\lambda = 0$ to obtain to approximate the integrand.

Since we no longer have a singularity in the log-term, we have the following Taylor series around $\lambda = 1$
\[
\sqrt{1 + \frac{1}{\lambda^2}} \log (\lambda) = \sqrt{2}(\lambda - 1) + O \left( (\lambda - 1)^2 \right),
\]
which yields

\[
\frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} \log [\lambda] = \sqrt{2} + O(x - 1) \text{ near } \lambda = 1
\]

For large values of \( \lambda \), we have as before

\[
\frac{\sqrt{1 + \lambda^2}}{\lambda(\lambda - 1)} \log [\lambda] = \frac{\log (\lambda)}{\lambda} + O \left( \frac{1}{\lambda^2} \right) \text{ for large } \lambda.
\]

This makes

\[
I_{31}(\psi) = -\sqrt{2} \frac{\psi}{k_i} + B_4
\]

\[
I_{32}(\psi) = \frac{\log^2 (1 + \psi/l_i)}{2} + B_5,
\]

since the integrals of the remainders are bounded. This proves the expression in A.1.9 and 4.14 now follows.

Lastly, we need to prove that

\[
K_i^{-1}(\psi) = \frac{1}{\sqrt{2} k_i^0} - \sqrt{2} \log \left( \frac{\psi}{k_i} \right) + \log (1 + \psi/l_i) + B_{10}, \quad (A.1.10)
\]

which follows easily from the work already done by splitting the area of integration like before, denoting the integrals \( K_i^{-1}(\psi) \), \( M_{10} \), \( K_{i2}^{-1}(\psi) \), and noting that

\[
\frac{\sqrt{1 + \lambda^2}}{\lambda - 1} = \frac{\sqrt{2}}{\lambda - 1} - \frac{1}{\sqrt{2}} + O(\lambda - 1) \text{ near } \lambda = 1 \quad (A.1.11)
\]

\[
= \frac{1}{\lambda} + O \left( \frac{1}{\lambda^2} \right) \text{ for large } \lambda. \quad (A.1.12)
\]

Then we have

\[
K_1^{-1}(\psi) = \frac{1}{\sqrt{2} k_i} - \sqrt{2} \log (\psi/k_i) + B_8
\]

\[
K_2^{-1}(\psi) = \log (1 + \psi/l_i) + B_9,
\]

and 4.15 follows, and the lemma is proved.

\[\Box\]

**A.2 Proof of Lemma 4.2**

Again, I restate the lemma

**Lemma 4.2.** Let \( K_i(\psi) \) and \( \pi_i(\psi) \) be as defined above. Let \( \{k_i\}_{i=1}^\infty \) be a sequence diverging to infinity and put \( l_i = k_i^{-1} \). Then, putting \( \psi_0 = 1 \) we have that

\[
\lim_{i \to \infty} K_i(\psi) \pi_i(\psi) = \psi^{-4 + \frac{2}{\sqrt{2}}} \quad (4.17)
\]

\[\Box\]
A.2. PROOF OF LEMMA 4.2

Proof. We start by rewriting the expression as,

\[ \frac{K_i(\psi) \pi_i(\psi)}{K_i(1) \pi_i(1)} = \frac{K_i(\psi)}{K_i(1)} \frac{1}{\psi} \exp \left\{ K_i(\psi) I_1^i(\psi) - K_i(1) I_1^i(1) \right\}. \] \hfill (A.2.1)

Now, if the limits exists and aren’t zero we can do

\[ \lim_{i \to \infty} \frac{K_i(\psi) \pi_i(\psi)}{K_i(1) \pi_i(1)} = \frac{1}{\psi} \left( \lim_{i \to \infty} K_i(\psi) \right) \exp \left\{ \lim_{i \to \infty} \left[ K_i(\psi) I_1^i(\psi) - K_i(1) I_1^i(1) \right] \right\}. \]

I will first prove that the limit labeled I equals one. Start by writing

\[ K_i(\psi) = \frac{1/k_i - 2 \log(1/k_i) + \sqrt{2}\log(1+k_i) + B_{12}}{\psi/k_i - 2 \log(\psi/k_i) + \sqrt{2}\log(1+\psi k_i) + B_{12}} \]

\[ = \frac{1/k_i - 2 \log(1/k_i) + \sqrt{2}\log(k_i) + \sqrt{2}\log(1+1/k_i) + B_{12}}{\psi/k_i - 2 \log(\psi/k_i) + \sqrt{2}\log(\psi k_i) + \sqrt{2}\log(1+1/(\psi k_i)) + B_{12}} \]

and note that the terms $1/k_i$, $\psi/k_i$, $\log(1+1/k_i)$ and $\log(1+1/(\psi k_i))$ will go to zero as $k_i \to \infty$. In addition, the bounded function $B_{12}$ will become small compared to the log terms, so we may write

\[ \lim_{i \to \infty} \frac{K_i(\psi)}{K_i(1)} = \lim_{i \to \infty} \frac{-2 \log(1/k_i) + \sqrt{2}\log(k_i)}{-2 \log(\psi/k_i) + \sqrt{2}\log(\psi k_i)} \]

\[ = \lim_{i \to \infty} \frac{2 \log(k_i) + \sqrt{2}\log(k_i)}{2 \log(\psi) + 2 \log(\psi k_i) + \sqrt{2}\log(k_i)} \]

\[ \text{L'Hopital's Rule} \]
\[ = \lim_{i \to \infty} \frac{\frac{2}{k_i} + \sqrt{2}}{\frac{2}{k_i} + \sqrt{2}} = 1, \]

and the limit I is proved.

Next we need to tackle the limit labeled II. Start by writing $K_i(\psi) I_1^i(\psi)$ as

\[ K_i(\psi) I_1^i(\psi) = \frac{1}{\sqrt{2}} \left( \psi k_i \log \left( \frac{\psi}{k_i} \right) + \frac{\psi}{k_i} - \log^2 \left( \frac{\psi}{k_i} \right) + B_{7} \right) \]

\[ = \frac{1}{\sqrt{2}} \left( \psi k_i \log \left( \frac{\psi}{k_i} \right) - 2 \log \left( \frac{\psi}{k_i} \right) + \sqrt{2}\log(1+\psi k_i) \right) + B_{10} \]

\[ = \frac{\psi}{k_i} \log \left( \frac{\psi}{k_i} \right) + \psi/k_i - \log^2 \left( \frac{\psi}{k_i} \right) + B_{13} \]

\[ \psi/k_i + \sqrt{2}\log \left( \psi^{-\sqrt{2}} \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) \right) + B_{14}, \] \hfill (A.2.2)

and make some simplifications about this expression’s behavior in the limit.

For any value of $\psi$, the terms $\psi/k_i \log \left( \frac{\psi}{k_i} \right)$, $\psi/k_i$ and $B_{13}$ in the numerator are bounded as $k_i \to \infty$, so they can be removed when taking the limit. In addition, the $\log^2 \left( \frac{\psi}{k_i} \right)$ term can be written $\log^2 \left( \psi \right) - 2 \log \left( \psi \right) \log \left( k_i \right) + \log^2 \left( k_i \right)$, where we note that $\log^2 \left( \psi \right)$ is also bounded.

Therefore, the expression essentially behaves like

\[ \frac{2 \log \left( \psi \right) \log \left( k_i \right) - \log^2 \left( k_i \right)}{\psi/k_i + \sqrt{2}\log \left( \psi^{-\sqrt{2}} \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) \right) + B_{14}} \]
Keeping the limiting behavior in mind, the limit in II essentially becomes

\[
\lim_{i \to \infty} \left[ \frac{2 \log (\psi) \log (k_i) - \log^2 (k_i)}{\psi/k_i + \sqrt{2} \log \left( \psi^{-\sqrt{2}} \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) \right) + B_{14}} \right] + \frac{\log^2 (k_i)}{1/k_i + \sqrt{2} \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{15}} \right] \tag{A.2.3}
\]

Which can be split into two separate pieces

\[
\lim_{i \to \infty} \left[ \frac{2 \log (\psi) \log (k_i)}{\psi/k_i + \sqrt{2} \log \left( \psi^{-\sqrt{2}} \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) \right) + B_{14}} \right] \tag{A.2.4}
\]

\[
+ \lim_{i \to \infty} \left[ \frac{\log^2 (k_i)}{1/k_i + \sqrt{2} \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{15}} \right] \tag{A.2.5}
\]

Let’s start with evaluating the limit in (A.2.4). Pulling out the \( \log (\psi) \) term in the numerator, and combining all the bounded terms in the denominator we get

\[
\log (\psi) \lim_{i \to \infty} \frac{2 \log (k_i)}{\sqrt{2} \log (k_i^{1+\sqrt{2}}) + B_{16}} = \log (\psi) \left( 2 - \sqrt{2} \right) \tag{A.2.6}
\]

Evaluating the limit in equation (A.2.5) requires more work. Writing the two expressions on a common denominator yields a numerator that looks like this

\[
\log^2 (k_i) \left\{ \frac{\psi / k_i - 1 / k_i + \sqrt{2} \log \left( \psi^{-\sqrt{2}} \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) \right)}{\sqrt{2} \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{14} - B_{15}} \right\}. \tag{A.2.7}
\]

Now, \( B_{15} \) is essentially \( B_{14} \) evaluated at \( \psi = 1 \), while \( B_{14} \) was derived from \( B_{10} \) by canceling the \( 1/\sqrt{2} \) term in equation (A.2.2). From appendix A.3 we have that that content of \( B_{10} \) is \( M_{\epsilon} + O(\epsilon) + O(\psi/k_i) \) where \( M_{\epsilon} \) and \( O(\epsilon) \) are expressions whose size is determined by \( \epsilon \), and \( O(\psi/k_i) \) denote bounded terms of the form \( \psi/k_i \). Then we see that \( B_{14} - B_{15} \) is essentially just some bounded terms \( O(1/k_i) + O(\psi/k_i) \)

Before moving on to taking the limit, we can simplify the limit further by noting that

\[
\lim_{i \to \infty} \left\{ \frac{\psi / k_i \log^2 (k_i)}{\log \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) + B_{17}} \right\} \left\{ \frac{\log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{18}}{\log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{18}} \right\} = 0. \tag{A.2.8}
\]

This means that all the bounded terms in the numerator (A.2.7) can be removed, and we are left with evaluating the limit of

\[
L_1 = \frac{\log^2 (k_i) \log \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) - \log^2 (k_i) \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + \log^2 (k_i) \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right)}{\sqrt{2} \left\{ \log \left( k_i^{\sqrt{2}} + \psi k_i^{1+\sqrt{2}} \right) + B_{19} \right\} \left\{ \log \left( k_i^{\sqrt{2}} + k_i^{1+\sqrt{2}} \right) + B_{20} \right\}} \tag{A.2.9}
\]
which can be factorized as \( L_1 = \frac{1}{\sqrt{2}} (L_{11} L_{12} + L_{13} L_{14}) \) where,

\[
L_{11} = \frac{\log(k_i)}{\log\left(k_i^{\frac{1}{2}} + \psi k_i^{1+\sqrt{2}} \right) + B_{17}}
\]

\[
L_{12} = \frac{\log(k_i) \log\left(k_i^{\frac{1}{2}} + \psi k_i^{1+\sqrt{2}} \right) - \log(k_i) \log\left(k_i^{\frac{1}{2}} + k_i^{1+\sqrt{2}} \right)}{\log\left(k_i^{\frac{1}{2}} + k_i^{1+\sqrt{2}} \right) + B_{18}}
\]

\[
L_{13} = \frac{\log(k_i) \log \psi^{-\frac{1}{2}}}{\log\left(k_i^{\frac{1}{2}} + \psi k_i^{1+\sqrt{2}} \right) + B_{17}}
\]

\[
L_{14} = \frac{\log(k_i)}{\log\left(k_i^{\frac{1}{2}} + k_i^{1+\sqrt{2}} \right) + B_{18}}
\]

and the limit taken over each piece separately. By the same rewrite as in A.2.6, we have

\[
\lim_{i \to \infty} L_{11} = \lim_{i \to \infty} L_{14} = (1 + \sqrt{2})^{-1}.
\]

For \( L_{13} \) we get

\[
\lim_{i \to \infty} L_{13} = (1 + \sqrt{2})^{-1} \log(\psi^{-\frac{1}{2}})
\]

Lastly we need to tackle \( L_{12} \). Start by rewriting the expression as

\[
L_{12} = \left( \frac{\log(k_i)}{\log\left(k_i^{\frac{1}{2}} + \psi k_i^{1+\sqrt{2}} \right) + B_{18}} \right) \log\left( \frac{1 + \psi k_i}{1 + k_i} \right).
\]

Taking the limit of the first term, we get \((1 + \sqrt{2})^{-1}\), while the limit of the second term can be found by moving the limit inside the logarithm and applying L'Hôpital’s rule to obtain \(\log(\psi)\). In total this yields:

\[
\lim_{i \to \infty} L_{12} = \frac{\log(\psi)}{1 + \sqrt{2}}.
\]

Putting this all together we get

\[
\lim_{i \to \infty} L_1 = \frac{1}{\sqrt{2}} \left[ \frac{1}{1 + \sqrt{2}} \log(\psi) \left( \frac{1}{1 + \sqrt{2}} \right) \log\left( \frac{1 + \psi k_i}{1 + k_i} \right) \right]
\]

\[
= \log(\psi) \left( -5 + \frac{7}{\sqrt{2}} \right).
\]

(A.2.10)

and combining it with the limit from [A.2.6] we get

\[
\lim_{i \to \infty} \left[ K_i(\psi) I_1(\psi) - K_i(1) I_1(1) \right] = (2 - \sqrt{2}) \log(\psi) + \log(\psi) \left( -5 + \frac{7}{\sqrt{2}} \right)
\]

\[
= \left( \frac{5}{\sqrt{2}} - 3 \right) \log(\psi)
\]

(A.2.11)

and finally:

\[
\lim_{i \to \infty} \frac{K_i(\psi) \pi_i(\psi)}{K_i(1) \pi_i(1)} = \psi^{-4 + \frac{5}{\sqrt{2}}}.
\]

(A.2.12)
A.3 The error in the normalizing constant

For Lemma 4.2, we need to control the bounded error term, $B_{10}$, in equation A.1.10. In the following $M_\epsilon$ and $m_\epsilon$ will denote constants that are, respectively large or small, and is controlled by how $\epsilon$ is chosen. $h(k_i)$ will denote terms of the form $a\psi/(b+\psi k_i^\epsilon)$, which are bounded terms that tend to zero as $k_i \to \infty$.

From integrating A.1.11 and A.1.12 over the intervals $(1 + \psi/k_i, 1 + \epsilon)$ and $(1 + \epsilon^{-1}, 1 + \psi k_i)$ the remainder terms, $H_3(\lambda)$ and $H_4(\lambda)$ are of the form $O(\lambda - 1)$ and $O(\lambda^{-2})$, and thus by the definition of $O(\cdot)$:

\[
\int_{1+\epsilon}^{1+\epsilon} |H_3(\lambda)| \, d\lambda \leq M_1 \int_{1+\psi/k_i}^{1+\psi/k_i} |(\lambda - 1)| \, d\lambda \quad \text{for some constant } M_1
\]
\[
= M_1 (m_\epsilon - h(k_i)),
\]
\[
\int_{1+\epsilon^{-1}}^{1+\psi k_i} |H_3(\lambda)| \, d\lambda \leq M_2 \int_{1+\epsilon^{-1}}^{1+\psi k_i} \frac{1}{\lambda^2} \, d\lambda \quad \text{for some constant } M_2
\]
\[
= M_2 (m_\epsilon - h(k_i)).
\]

Combining this with the large constant $M_\epsilon$ one gets from integrating over $(1 + \epsilon, 1 + \epsilon^{-1})$ the error term $B_{10}$ is essentially

\[
B_{10} = M_\epsilon + O(\epsilon) + O(\psi/k_i) \quad \text{(A.3.1)}
\]

where $O(\epsilon)$ denotes terms bounded by $\epsilon$ (and thus can be made as small as we please), and $O(\psi/k_i)$ denotes bounded terms that tend to zero as $k_i \to \infty$. 
Appendix B

Selected R code

B.1 Algorithm from remark 4.2

```r
# Function for calculating CD for a single pair
cond.cdf = function(delta,S.obs,A.obs){
  k = (delta!=0)*(1-exp(-delta*S.obs))/(1-exp(-delta*A.obs))+(delta==0)*S.obs/A.obs
  return(k)
}

# Function for calculating CD-alpha for root finding
cond.cdf.root = function(delta,alpha,S.obs,A.obs){
  if (delta!=0){
    k = (1-exp(-delta*S.obs))/(1-exp(-delta*A.obs))
  } else{k = S.obs/A.obs}
  return(k-alpha)
}

n = 10
theta0 = 2
delta0 = 2

# Sampling data
x.obs = rexp(n,theta0)
y.obs = rexp(n,theta0+delta0)

# Initial root finding for individual CDs
lowerquants = rep(NA,n)
upperquants = rep(NA,n)
for (i in 1:n){
  s = y.obs[i]
  a = x.obs[i]+y.obs[i]
  lowerquants[i] = uniroot(cond.cdf.root,c(-10,10000),alpha=0.01,S.obs=s,A.obs=a,extendInt="yes")$root
  upperquants[i] = uniroot(cond.cdf.root,c(-10,10000),alpha=0.99,S.obs=s,A.obs=a,extendInt="yes")$root
}
lowerest = min(lowerquants)
upperest = max(upperquants)
delta.seq.est = seq(lowerest,upperest,length.out=10000)

# Creating the average CD
meanCD.matr = matrix(NA,nrow=n,ncol=length(delta.seq.est))
for (i in 1:n){
  meanCD.matr[i,] = cond.cdf(delta.seq.est,y.obs[i],y.obs[i]+x.obs[i])
}
meanCD = apply(meanCD.matr,2,median,na.rm=T)

# Finding the roots
lower = delta.seq.est[which(abs(meanCD-0.01)==min(abs(meanCD-0.01)))]
upper = delta.seq.est[which(abs(meanCD-0.99)==min(abs(meanCD-0.99)))]
#print(c(sum(x.obs+y.obs),sum(y.obs),lower,upper))

# Creating a new sequence in the parameter space
# This sequence needs to be constructed on the basis of the y-axis
```

conf = seq(0.01, 0.99, length.out=100)
delta = rep(NA, 100)
for (i in 1:100) {
delta[i] = delta.est[which(abs(meanCD-conf[i]) == min(abs(meanCD-
conf[i])))]
}
# The sequence delta should now contain all the interesting parts of the parameter space.

B.2 Marginal posterior from section 4.3.1

# Script for comparing confidence distributions and marginal posteriors in linear
combinations of normal means.

marg.posterior = function(data, B, psi, ci, a) {
ybar = sapply(data, mean)
n = sapply(data, length)
m = length(data)
Si = (n-1)*sapply(data, var)
post = rep(NA, length(psi))
for (i in 1:length(psi)) {
  A = (sum(Si+n*(ybar[psi[i]]*ci)/sum(ci^2/n))^2)
p = (sum(n)+2*a-m-1)/2)
  post[i] = (A-B)^(-p)
}
return(post)
}

getB = function(data, ci, n) {
ybar = sapply(data, mean)
n = sapply(data, length)
m = length(data)
Si = (n-1)*sapply(data, var)
fj = rep(0, m)
gj = rep(0, m)
# Loops to find fj and gj
for (i in 2:m) {
  if (i==2) {
  } else {
    gj[i] = 1/(n[i]+(n[1]-sum(gj[2:(i-1)]))*ci[i]^2)/sum(gj[2:(i-1)]))^2
  }
}
b = rep(0, m)
for (i in 2:m) {
  b[i] = 1/(n[i]+(n[1]-sum(gj[2:(i-1)]))*ci[i]^2)/sum(fj[2:(i-1)])*ci[i]/ci[1]^2
}
B = sum(b)
return(B)
}

# Setting parameters
m = 5
n = c(10, 10, 10, 3, 3)
mu = c(2, 2, 2, 2, 2)
sigma = 2
ci = c(1, 1, 1, 1, 1)

# Sampling data
y = list()
for (i in 1:m) {
  y[[i]] = rnorm(n[i], mu[i], sigma)
}

# Calculating sufficient statistics.
ybar = sapply(y,mean)
Si = sapply(y,var)
sp = sqrt(1/(sum(n)-m)*sum((n-1)*Si))

# Calculating the CD
psi.seq = seq(-30,30,by=.01)
CD = pt((psi.seq-as.numeric(ybar%*%ci))/(sp*sqrt(sum(ci^2/n))),df=(sum(n)-m))

## Adding the marginal posterior

# First calculating the B-constant
B = getB(y,ci,n)
# Getting the marginal posterior
m.post.J = marg.posterior(y,B,psi.seq,ci,(m+2)/2)
m.post.R1 = marg.posterior(y,B,psi.seq,ci,(m+1)/2)
m.post.R2 = marg.posterior(y,B,psi.seq,ci,1)

# Converting to CDFs
marg.cdf.J = rep(NA,length(psi.seq))
marg.cdf.R1 = rep(NA,length(psi.seq))
marg.cdf.R2 = rep(NA,length(psi.seq))
for (i in 1:length(psi.seq)){
marg.cdf.J[i] = sum(m.post.J[1:i])
marg.cdf.R1[i] = sum(m.post.R1[1:i])
marg.cdf.R2[i] = sum(m.post.R2[1:i])
}

# Normalizing
marg.cdf.R1 = marg.cdf.R1/max(marg.cdf.R1)
marg.cdf.R2 = marg.cdf.R2/max(marg.cdf.R2)

# The plotting
par(mfcol=c(2,2))
# CCs
plot(psi.seq,abs(1-2*CD),type="l",ylab="Confidence",xlab=expression(psi),xlim=c(2,10), ylim=c(0,1))
legend("bottomleft",inset=0.02,legend=c(expression(C(delta)),expression(paste(pi[J],sep="")), expression(paste(pi[R1],sep="")), expression(paste(pi[R2],sep=""))), lty=c(1,2,3,4),cex=0.9,box.lty=0,bty="n")
lines(psi.seq,abs(1-2*marg.cdf.J),lty=2)
lines(psi.seq,abs(1-2*marg.cdf.R1),lty=3)
lines(psi.seq,abs(1-2*marg.cdf.R2),lty=4)
# Log absolute difference
J.diff = log(abs(marg.cdf.J-CD))
R1.diff = log(abs(marg.cdf.R1-CD))
R2.diff = log(abs(marg.cdf.R2-CD))
plot(psi.seq,J.diff,type="l",ylab="Log absolute-difference",xlab=expression(psi),lty=2, xlim=c(2,10))
lines(psi.seq,R1.diff,lty=3)
lines(psi.seq,R2.diff,lty=4)
legend("bottomleft",inset=0.02,legend=c(expression(paste(pi[J],sep=""))), expression(paste(pi[R1],sep=""))), expression(paste(pi[R2],sep=""))), lty=c(2,3,4),cex=0.9,box.lty=0,bty="n")