https://doi.org/10.26219/heal.aueb.6623

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Employing Bayes Factors in Clinical Trials

By

Anastasios K. Tasoulas

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
February 2021

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΧΡΗΣΗ ΠΑΡΑΓΟΝΤΩΝ BAYES ΣΕ ΚΛΙΝΙΚΕΣ ΔΟΚΙΜΕΣ

Αναστάσιος Κ. Τασούλας

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Φεβρουάριος 2021

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

ACKNOWLEDGEMENTS

My family, friends, and professors who encouraged me and recognized my efforts. Special thanks to my supervisor Prof. Ioannis Ntzoufras, and co-supervisor Dr. Stavros Nikolakopoulos, for their invaluable assistance.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

VITA

I was born in 1993. Graduated from the National and Kapodistrian University of Athens with a BSc in Mathematics in 2018. Completed military service in 2019.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

ABSTRACT

Anastasios Tasoulas

Employing Bayes Factors in Clinical Trials

February 2021

The goal of this dissertation is to examine how Bayes factors can be utilized for hypothesis testing in clinical trials; namely, in tests of superiority, equivalence, and non-inferiority. Bayes factors are not as popular in medical research as frequentist (classical) hypothesis testing methods, despite their many claimed advantages. A major reason for this is that frequentist inference is focused on the long-run assessment of type I and II error rates, a feature of high importance in clinical trials. We demonstrate that Bayes factors can be adapted to satisfy such a requirement as well. The sensitivity of the Bayes factor to the choice of prior distribution is also of interest. In particular, we examine the use of local versus non-local priors, as the latter have been demonstrated to provide a means of "fairer" testing by balancing the rate of evidence accumulation towards hypotheses when each are true. After analyzing the theoretical dimensions of how to implement Bayes factors in clinical trials, we perform a simulation study aimed at comparing their performance in tests of superiority, equivalence, and non-inferiority, under different settings and specifications.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

ΠΕΡΙΛΗΨΗ

Αναστάσιος Τασούλας

ΧΡΗΣΗ ΠΑΡΑΓΟΝΤΩΝ BAYES ΣΕ ΚΛΙΝΙΚΕΣ ΔΟΚΙΜΕΣ

Φεβρουάριος 2021

Στόχος της παρούσας διατριβής είναι να εξετασθεί το πώς οι παράγοντες Bayes μπορούν να εφαρμοστούν στους ελέγχους υποθέσεων κλινικών δοκιμών˙ συγκεκριμένα, στους ελέγχους ανωτερότητας, ισοδυναμίας, και μη-κατωτερότητας. Οι παράγοντες Bayes δεν είναι τόσο διαδεδομένοι στην κλινική έρευνα όσο οι κλασικές (frequentist) μέθοδοι ελέγχων υποθέσεων, παρά τα πολλά προτερήματα που τους αποδίδονται. Ένας μείζων λόγος για αυτό, είναι ότι η κλασική συμπερασματολογία είναι επικεντρωμένη στον μακροπρόθεσμο καθορισμό της πιθανότητας σφαλμάτων τύπου I και II, ένα χαρακτηριστικό ύψιστης σημασίας στις κλινικές δοκιμές. Δείχνουμε ότι οι παράγοντες Bayes μπορούν να προσαρμοστούν ώστε να ικανοποιήσουν και αυτήν την ανάγκη. Η ευαισθησία των παραγόντων Bayes στην επιλογή της εκ-των-προτέρων (prior) κατανομής είναι επίσης σημείο ενδιαφέροντος. Συγκεκριμένα, εξετάζουμε τη χρήση local και non-local prior κατανομών, καθώς έχει αποδειχθεί ότι οι τελευταίες μπορούν να προσφέρουν έναν τρόπο «δικαιότερου» ελέγχου, εξισορροπώντας την ταχύτητα συσσώρευσης των αποδεικτικών στοιχείων προς υποστήριξη των υποθέσεων, όταν κάθε μία από αυτές είναι αληθής. Αφού αναλύσουμε τη θεωρητική διάσταση της χρήσης παραγόντων Bayes στις κλινικές δοκιμές, εκτελούμε μια άσκηση προσομοίωσης με στόχο να συγκρίνουμε την απόδοσή τους σε ελέγχους ανωτερότητας, ισοδυναμίας, και μη-κατωτερότητας, υπό διαφορετικές συνθήκες και ρυθμίσεις.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

TABLE OF CONTENTS

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

LIST OF TABLES

https://doi.org/10.26219/heal.aueb.6623

 https://doi.org/10.26219/heal.aueb.6623

LIST OF FIGURES

https://doi.org/10.26219/heal.aueb.6623

 https://doi.org/10.26219/heal.aueb.6623

CHAPTER 1

INTRODUCTION

1.1 Frequentist Hypothesis Testing and its Limitations

1.1.1 Historical Background

The term 'frequentist' or 'classical hypothesis testing' is generally understood as the method of determining the credibility of a proposed hypothesis, using a probability called the p-value as a criterion. A typical example goes as follows:

> A sample $X = (x_1, x_2, \ldots, x_n)$ of data is collected, presumably originating from a distribution $f(X|\theta)$, where θ is a parameter of interest. We seek to test the hypothesis H_0: $\theta = \theta_0$ against the alternative H_1: $\theta \neq \theta_0$. Under the assumption that H_0 is true, a test statistic $T(X|H_0)$ is constructed that follows a probability distribution. Using this result, we derive the p-value, which is the probability p of $T(X|H_0)$ taking on a value equal or more extreme than what was observed, given the sample X. If p falls beneath a threshold, usually $\alpha = 0.05$ by convention, hypothesis H_0 is rejected, and we claim that θ differs from θ_0 at the $\alpha \cdot 100\%$ significance level. Otherwise, H_0 is not rejected, and we proceed to act as if it is true.

This method is a largely unacknowledged amalgamation of two distinct statistical techniques that were originally established to address different problems. The first one is the p-value, originally introduced by Karl Pearson

https://doi.org/10.26219/heal.aueb.6623

(Pearson, 1900), but formally outlined by Sir R. A. Fisher (Fisher, 1992). The second one is the technique of hypothesis testing, as formulated by mathematicians Jerzy Neyman and Egon Pearson (Neyman & Pearson, 1928).

Fisher defined the *p*-value, also called 'significance probability', in the same way as it is properly understood today: the probability of observing an experimental result equal or more extreme than what was already observed, given a hypothesis. However, it was merely intended to quantify the credibility of a given hypothesis subject to nullification (hence the term 'null' hypothesis) in a single experiment, within what was then called a 'significance test'. Background knowledge relevant to the experiment at hand could also be taken into consideration before drawing a conclusion (Fisher, 1973), but generally Fisher viewed small *p*-values as evidence pointing towards rejection of the hypothesis. In his own words:

> *"If p is between 0.1 and 0.9, there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05, and consider that [lower values of p] indicate a real discrepancy"* (Fisher, 1992).

Soon after, Neyman and Pearson introduced the hypothesis test. Unlike the significance test, this one included an additional 'alternative' hypothesis H_1, usually taken to be the complement of the null H_0. The existence of two hypotheses introduced two types of errors, seen in the following table[1]:

1 In Neyman and Pearson's view, it is possible to accept the null hypothesis. The notion that the null can only either be rejected, or fail to be rejected, is a legacy of Fisher's significance test, where no alternative hypothesis exists (Goodman, 1993).

https://doi.org/10.26219/heal.aueb.6623

	H_0 true (H_1 false)	**H_0 false (H_1 true)**
Reject H_0 (Accept H_1)	Type I error	OK
Accept H_0 (Reject H_1)	OK	Type II error

Table 1: Error types in hypothesis tests.

Under each hypothesis, the probabilities of type I and II errors (α and β respectively) could be mathematically calculated, and thus minimized based on the test's design. Critical regions corresponding to those probabilities would be constructed for a particular test statistic *a priori*. Subsequent observations translating to values of the statistic within those critical regions would result in rejection of the null hypothesis. Even though this method provided no way to evaluate the credibility of a given hypothesis in a single experiment, the idea was that after many similar experiments, the proportions of type I and II errors would converge towards these probabilities α and β, and thus not greatly exceed them. In the researchers' own words:

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

It becomes clear that unlike Fisher's significance test, the hypothesis test of Neyman and Pearson was *not* a tool of inference, but a decision rule designed to yield desirable long-term results. A source of confusion has been the fact that the value of Neyman and Pearson's test statistic could be expressed in terms of a probability, thus facilitating a connection to Fisher's p-value (Biau et al., 2010). The two methods ended up being used in tandem so frequently by researchers that they eventually merged, despite Fisher's objections:

 https://doi.org/10.26219/heal.aueb.6623

> *"On the whole the ideas (a) that a test of significance must be regarded as one of a series of similar tests applied to a succession of similar bodies of data, and (b) that the purpose of the test is to discriminate or "decide" between two or more hypotheses, have greatly obscured their understanding, when taken not as contingent possibilities but as elements essential to their logic"* (Fisher, 1973).

1.1.2 Misconceptions and Drawbacks

The philosophical differences at the roots of these two methods have led to researchers characterizing them 'contradictory', or 'incompatible' (Hubbard & Bayarri, 2003, and Goodman, 1999a), but go largely unrecognized in the academic literature, paving the way for erroneous interpretations and misunderstandings of the fundamental concepts, even by statisticians (Kalbfleisch & Sprott, 1976). Common mistakes include the reading of the p-value as a data-specific version of the type I error rate α (Goodman, 1993, Hubbard & Bayarri, 2003, and Kalbfleisch & Sprott, 1976), as the probability of the null hypothesis H_0 being true, and many others (Goodman, 2008).

These issues however, do not exhaust the list of reasons why frequentist hypothesis testing is rendered problematic. Many practical limitations arise that are endemic to frequentist inference in general (Wagenmakers et al., 2008), but as far as hypothesis tests go, some troubling ones include:

- The inherent inability of p-values to quantify statistical evidence *for* a hypothesis. Furthermore, many researchers claim that their mechanism of quantifying evidence *against* hypotheses is flawed as well: If p-values reflected the totality of the evidence, then the same p-values would have equal implications for experiments of varying sample sizes, yet the opposite has been argued (Peto et al., 1976).

https://doi.org/10.26219/heal.aueb.6623

- As the p-value is calculated over a portion of the sample space, changes in that space can affect it. This means that arbitrary factors outside of the observed data, such as the researcher's sampling plan, can influence a test's verdict. In a notorious example, the same data under a Binomial formulation ($x \in \{0, 1, \ldots, N\}$, $N \in \mathbb{N} \setminus \{0\}$) offer an opposing conclusion than under a Negative Binomial one ($x \in \mathbb{N}$) (Berger & Wolpert, 1988).
- In order for the p-value to be calculated, the probability distribution of the test statistic must be specified, which depends on the value of the parameter of interest under the null hypothesis. This introduces obstacles when seeking to test *interval* null hypotheses, i.e.: hypotheses which define a parameter of interest to be within a range, rather than having a single value (point null hypotheses).
- Difficulty of interpreting results when it comes to testing non-nested hypotheses. Consider H_1: $\mu_1 < \mu_2 < \mu_3 < \mu_4$ against H_2: $\mu_1 < \mu_3 < \mu_2 < \mu_4$. It is not obvious which hypothesis to take as the null, and trying both introduces the risk of rejecting both of them (or not rejecting either one). Additionally, consider having a p-value of 0.049 when H_1 is the null, but one of 0.051 when H_2 is, for significance level $\alpha = 0.05$. One is rejected, the other is not, but the p-values' 'weights of the evidence' are effectively equal (Wagenmakers et al., 2008).

Nevertheless, frequentist hypothesis testing delivers results adequately enough to maintain its dominant status in fields such as sociology, psychology, and medicine. A proposed alternative is to test hypotheses using the method of Bayes factors, as they possess many desirable inferential properties that are absent from the frequentist toolbox (Wagenmakers et al., 2008).

https://doi.org/10.26219/heal.aueb.6623

1.2 The Bayesian Framework

In a broad sense, the objective of the Bayesian school of statistical inference is to quantify, in probabilistic terms, how much one's existing, or 'prior' beliefs regarding a state of reality change after the observation of data. Its name derives from Thomas Bayes, an 18th century English Protestant cleric, whose posthumously published letter to a colleague (Bayes, 1763) set the foundations for what later came to be defined as the Bayes theorem. The simplest form of this theorem (Gill, 2014) states that for two events A and B, where B has a non-zero probability of occurrence $P(B) \neq 0$, the conditional probability of A given B is equal to:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.2.1}$$

The Bayes theorem can also be expressed in terms of probability distribution functions. Suppose a sample x is obtained from a sampling distribution $f(x|\theta)$, where θ is the parameter of interest. If θ is assumed to follow a 'prior' distribution $p(\theta)$ *before* the data are observed, then its 'posterior' distribution *after* the observed data are taken into consideration is given by:

$$\underbrace{p(\theta|x)}_{posterior} = \frac{f(x|\theta)p(\theta)}{f(x)} \propto \underbrace{f(x|\theta)}_{likelihood}\underbrace{p(\theta)}_{prior}, \tag{1.2.2}$$

where $f(x) = \int f(x|\theta)p(\theta)d\theta$ denotes the marginal distribution (or marginal likelihood) of the sample x, and plays the role of a normalizing constant that does not depend on θ. Since the focus is on parameter θ, and not the sample x, it is entirely appropriate to write $f(x|\theta)$ as $L(\theta|x)$, explicitly denoting the likelihood. Subsequently, the posterior distribution $p(\theta|x)$ can be used to make inferential statements about the parameter θ, like estimate its mean, variance, or construct a probability interval, formally called a 'credibility interval'.

https://doi.org/10.26219/heal.aueb.6623

This type of interval has a convenient interpretation[2] usually misattributed to its frequentist counterpart, the confidence interval.

The Bayesian attitude towards statistical inference essentially turns the classical (or frequentist) paradigm on its head: instead of treating observations as random quantities generated from a sequential sampling mechanism dependent on a fixed parameter θ, the Bayesian view conditions on the observations, and treats the parameter itself as a random quantity. Even though Bayesian methods have been known at least since the early 19th century[3] (Stigler, 2005), researchers were slow to adopt them in practice. An important reason for this delay was the difficulty involved in extracting closed-form expressions of posterior distributions, since the integrals used to calculate the marginal likelihood $f(x)$ were often intractable, especially in multi-dimensional problems (Robert & Casella, 2013, and Robert, 2007). As a consequence, Bayesian enthusiasts were for the better part of the 20th century restricted to the study and use of *conjugate* priors: distributions which lead to posteriors of the same family, thus making closed-form expressions readily available.

The eventual solution to this problem arrived in the late 1980s, when technological advances in processing power, and the increasing availability of computers, sped up the development and popularity of Markov Chain Monte Carlo (MCMC) techniques. Although known since the 1950s for their applications in physics (Metropolis et al., 1953), it wasn't until the very end of the 20th century that these techniques became popular among statisticians (Geman & Geman, 1984, and Gelfand & Smith, 1990). They were particularly attractive to Bayesian analysts, as they provided them with a means to simulate samples from posterior distributions (and thus make inferential statements), even when closed-form expressions for them were unavailable.

The resulting rise in the popularity of Bayesian inference didn't discourage its critics, one of their most popular objections being that Bayesian methods introduce subjective opinion through the prior distribution,

2 It represents an interval within which a parameter lies with some probability.

3 According to some researchers, it was Pierre-Simon, marquis de Laplace who actually discovered the Bayes theorem and expressed it in its general form (Stigler, 1986).

https://doi.org/10.26219/heal.aueb.6623

resulting in (partly) subjective posteriors, while distorting a scientific process that is supposed to be objective (Gelman, 2008). One response to such criticisms has been the use of 'uninformative' (also called 'vague', or 'diffuse') priors, as these do not favor (not significantly, at least) any parameter values over others, and thus allow inference to be maximally data-based (Gill, 2014). Another approach towards objective Bayesian analysis has been the use of 'empirical Bayes' methods, in which researchers allow the data to influence their choice of prior (Carlin & Louis, 2000).

The debate between frequentist and Bayesian proponents still goes strong, covering a lot more points of contention than just the prior distribution (Bayarri & Berger, 2004, and Albers et al., 2018). However, despite their downsides (real, or imagined), Bayesian procedures provide a flexible and increasingly attractive framework with many claimed advantages over its frequentist counterpart (Wagenmakers et al., 2008), and a wide variety of applications in areas such as classification, model building, and hypothesis testing.

1.3 Bayes Factors

1.3.1 Definition and Interpretation

Suppose the goal is to compare two competing hypotheses[4] H_0 and H_1, given a sample x. Let $f(x)$ be the marginal likelihood of the observed sample, $P(H_i|x)$ be the posterior probability of a hypothesis being true, $P(H_i)$ be the prior probability, and $f(x|H_i)$ be the marginal likelihood of the sample under

4 In contrast to the frequentist framework, the null hypothesis has no special status in Bayesian testing. In fact, as a hypothesis can be accepted as well as rejected, the 'null' hypothesis, as in 'subject to nullification', does not technically exist, so the term tends to be avoided. Many researchers go as far as to abandon the phrase 'testing hypotheses' entirely, in favor of 'comparing models' (Gill, 2014, Carlin & Louis, 2000, and Bernardo & Smith, 2009). The terms will be used interchangeably throughout this dissertation.

https://doi.org/10.26219/heal.aueb.6623

the assumption that hypothesis H_i is true, for $i \in \{0, 1\}$. We assume that $P(H_0) + P(H_1) = 1$. By the Bayes theorem, it follows that:

$$P(H_0|x) = \frac{f(x|H_0)P(H_0)}{f(x)}, \tag{1.3.1}$$

$$P(H_1|x) = \frac{f(x|H_1)P(H_1)}{f(x)}, \tag{1.3.2}$$

where $f(x) = f(x|H_0)P(H_0) + f(x|H_1)P(H_1)$. Taking the ratio of the above equations eliminates the unconditional marginal likelihood and returns:

$$\underbrace{\frac{P(H_0|x)}{P(H_1|x)}}_{\textit{posterior odds}} = \underbrace{\frac{f(x|H_0)}{f(x|H_1)}}_{\textit{Bayes factor}} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{\textit{prior odds}}. \tag{1.3.3}$$

The Bayes factor of H_0 versus H_1 is denoted BF_{01}, and is defined as the ratio of the marginal likelihoods of the observed data x under the two different hypotheses. It can alternatively be expressed as the ratio of posterior to prior hypotheses odds:

$$BF_{01} = \frac{\dfrac{P(H_0|x)}{P(H_1|x)}}{\dfrac{P(H_0)}{P(H_1)}} = \frac{P(H_0|x)P(H_1)}{P(H_1|x)P(H_0)}. \tag{1.3.4}$$

If hypothesis H_i assumes the existence of a parameter vector $\theta_i \in \Theta_i$ with a prior $p_i(\theta_i)$, then the marginal likelihood $f(x|H_i)$ is obtained by integration over the parameter space. This amounts to calculating a weighted average of the likelihood for different values of a parameter, which is in turn distributed according to the prior distribution. The marginal likelihood is given by:

$$f(x|H_i) = \int_{\Theta_i} f(x|\theta_i, H_i) p_i(\theta_i) d\theta_i, \quad \text{for} \quad i \in \{0,1\}. \tag{1.3.5}$$

 https://doi.org/10.26219/heal.aueb.6623

Note that in the special case when the comparison is between simple point hypotheses H_0: $\theta = \theta_0$ versus H_1: $\theta = \theta_1$, the Bayes factor reduces to the likelihood ratio between them:

$$BF_{01} = \frac{f(x|\theta_0)}{f(x|\theta_1)}. \tag{1.3.6}$$

Bayes factors measure how strongly the evidence provided by the data supports a hypothesis (or model) over another. They were first introduced by English mathematician Sir Harold Jeffreys in his 1939 book 'Theory of Probability' (Jeffreys, 1939), but the term itself did not appear in his writings. The quantity BF_{01} that we identify today as the Bayes factor, is denoted by Jeffreys as K, and is given by:

$$K = \frac{\dfrac{P(q|\theta H)}{P(q'|\theta H)}}{\dfrac{P(q|H)}{P(q'|H)}} = \frac{P(q|\theta H)P(q'|H)}{P(q'|\theta H)P(q|H)}, \tag{1.3.7}$$

where q and q' are the null and alternative hypotheses respectively, θ are the observed data, and H is the prior information. The new quantity K is simply referred to as a new parameter through which one can perform significance tests. This may create the impression that Jeffreys' methodology is conceptually associated with Fisher's significance test, but that is not the case. Jeffreys' perspective was different, as he outlined a method of quantifying evidence both *in favor* and *against* a hypothesis compared to another, making it a *relative* measure of evidential strength, not an *absolute* one like Fisher's. In contrast to the p-value, which is a probability bounded from 0 to 1, the Bayes factor can stretch from zero to infinity, making its assessment of the evidence somewhat unclear. For this purpose, Jeffreys constructed a rough scale of how the strength of the evidence is to be measured for different values of the Bayes factor (Table 2), and his comments on how to interpret it are illuminating:

https://doi.org/10.26219/heal.aueb.6623

"We do not need [BF_{01}] with much accuracy. [...] It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are 10^4 or 10^{10} to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice" (Jeffreys, 1939).

Grade level	Bayes Factor	Interpretation
0	$BF_{01} > 1$	H_0 supported over H_1.
1	$1 > BF_{01} > 10^{-0.5}$	Evidence against H_0 not worth more than a bare mention.
2	$10^{-0.5} > BF_{01} > 10^{-1}$	Evidence against H_0 substantial.
3	$10^{-1} > BF_{01} > 10^{-1.5}$	Evidence against H_0 strong.
4	$10^{-1.5} > BF_{01} > 10^{-2}$	Evidence against H_0 very strong.
5	$10^{-2} > BF_{01}$	Evidence against H_0 decisive.

Table 2: Jeffreys' grading of evidential strength provided by Bayes factors.

Even though Jeffreys' scale is regarded by researchers as somewhat arbitrary, and no explicit thresholds for acceptance or rejection exist, as is the case with frequentist hypothesis testing (Gill, 2014), it does provide a useful benchmark for evaluating evidence. In a landmark paper on Bayes factors, authors Kass and Raftery (Kass & Raftery, 1995) expanded on it and presented a modified version (Table 3), with the Bayes factor undergoing a $2\times log_e$ transformation. This was justified on the grounds that is provides a Bayes factor on the same scale as the familiar likelihood ratio test statistic.

$2\times\log_e(BF_{10})$	BF_{10}	Interpretation
0 - 2	1 - 3	Evidence against H_0 not worth more than a bare mention.
2 - 6	3 - 20	Evidence against H_0 positive.
6 - 10	20 - 150	Evidence against H_0 strong.
> 10	> 150	Evidence against H_0 very strong.

Table 3: Kass and Raftery's modification of Jeffreys' scale.

Beyond their ability to quantify evidence for hypotheses[5], Bayes factors have intuitive and convenient interpretations. A Bayes factor of $BF_{01} = 5$, for example, can be interpreted as any of the following statements (Goodman, 1999b):

- *The observed data are five times as likely to occur under hypothesis H_0 than they are under H_1.*
- *The data support hypothesis H_0 five times as strongly than they do H_1.*
- *The odds of hypothesis H_0 versus H_1 after the experiment are five times what they were before it.*

Some researchers (Lavine & Schervish, 1999) insist on a more careful and conservative interpretation of Bayes factors as merely the change in the odds in favor of the null hypothesis (the one in the enumerator), before and after observing the data. This interpretation is obvious when looking at formula (1.3.4), and it emphasizes the fact that equal Bayes factors can have different implications depending on one's prior knowledge.

5 Bayes factors or their logarithms are often referred to as the "weight of the evidence" (Good, 1985, and Good, 1992).

Consider three different scenarios where two hypotheses H_0 and H_1 are compared, and our prior convictions regarding the null H_0 range between suspicion, ambivalence, and support, as expressed by the prior odds in the following table:

Prior belief of H_0	**Prior odds**	**Bayes factor BF_{01}**	**Posterior odds**
Suspicion	0.1 (0.09)	5	0.5 (0.33)
		0.2	0.02 (0.01)
Ambivalence	1 (0.5)	5	5 (0.83)
		0.2	0.2 (0.16)
Support	10 (0.9)	5	50 (0.98)
		0.2	2 (0.66)

Table 4: Posterior odds for different prior beliefs and Bayes factor values. Prior and posterior probabilities of hypotheses in parentheses.

A Bayes factor of $BF_{01} = 5$ provides moderate evidence for hypothesis H_0 on its own, but if we have prior reasons to suspect it (represented by prior odds = 0.1), then the posterior odds lean in favor of the alternative H_1 (posterior odds = 0.5) and stronger evidence is required for us to change our opinion. Conversely, if we have a Bayes factor $BF_{01} = 0.2$, but have strong prior reasons to support H_0 (represented by prior odds = 10), the Bayes factor may point against the hypothesis on its own, but the posterior odds still end up leaning in its favor (posterior odds = 2). If we are uncertain either way (prior odds = 1), then the posterior odds are equal to the Bayes factor, which determines our final verdict completely.

It is important to keep in mind that a Bayes factor in favor of a hypothesis H_0, does not exclude the possibility of that hypothesis being dubious or generally insufficient in describing the reality underlying an experiment. It only implies that H_0 is better *compared* to its alternative H_1. It is entirely possible that in the grand scheme of things both H_0 and H_1 are pretty much equally implausible when compared to other alternatives.

https://doi.org/10.26219/heal.aueb.6623

1.3.2 Advantages over Frequentist Testing

The use of Bayes factors for comparing hypotheses gives researchers access to a wide variety of desirable mathematical and inferential properties that constitute an important advantage over classical methods. A detailed list of the most important ones follows, along with explanations of frequentist shortcomings.

1. *Bayes factors can provide evidence in favor of a hypothesis*.

This has been mentioned earlier and is one of the most important differences between Bayesian and frequentist testing, given that the *p*-value can only provide evidence *against* a hypothesis. This property does more than render obsolete the awkward and counter-intuitive clarification that *"one can never accept the alternative hypothesis, but only fail to reject the null"* (Wilkinson, 1999). Consider the example of an experimental setting where the objective is to seek evidence for the absence of an effect (e.g.: H_0: $d = 0$) under certain conditions. Classical tests are not equipped to provide a satisfactory answer to such a problem, but the best they can do is *suspend disbelief* of the null hypothesis (Wagenmakers, 2007). A researcher claiming the absence of an effect after discovering an insignificant *p*-value would be treating absence of evidence as evidence of absence, a serious inferential error (Altman & Bland, 1995). There exist many settings where the objective is to prove the null hypothesis, and Bayes factors are perfectly capable of providing such an affirmative answer (Gallistel, 2009, Rouder et al., 2009, and Wetzels et al., 2009).

2. *Bayes factors have an inherent mechanism of automatic parsimony*.

For clarification, the law of parsimony, also known as 'Ockham's razor', is a rule of scientific inference stating that when competing theories are equally good at describing reality, the simplest one (in the case of statistical models that would the one with the fewest parameters) should be preferred (Domingos, 1999). Consider a comparison between hypotheses H_0 and H_1. If

https://doi.org/10.26219/heal.aueb.6623

the observed data x are independent, and $x^{(k)} = \{x_i\colon i \in \{1, 2, \ldots, k\}\}$ for integers $k \in \{1, 2, \ldots, n\}$ then their likelihoods can be expressed as:

$$f(x|H_i) = f(x_n|x^{(n-1)}, H_i)\, f(x_{n-1}|x^{(n-2)}, H_i) \cdots f(x_2|x_1, H_i)\, f(x_1|H_i). \qquad (1.3.8)$$

Equation (1.3.8) demonstrates that the marginal likelihood is the product of sequential 'one step-ahead' out-of-sample predictions of the observed data: *"Given all previous observations* x_1, x_2, …,x_k, *what is the likelihood of this specific* x_{k+1} *being next?"* This justifies the interpretation of the Bayes factor as a measure of what model predicts the observed data better. A result of this property is that when it comes to comparing models, Bayes factors automatically favoring the one with the best predictive ability, also tend to favor the one that is the most parsimonious as a consequence. This happens because overly complex models with more parameters than needed tend to overfit the data, but their fit is based on a failure to distinguish between random variation (noise), and genuine information (signal). Thus they perform poorly when making new out-of-sample predictions, resulting in lower marginal likelihoods, and Bayes factors designating them as unfavorable (Wagenmakers et al., 2008).

3. *Bayes factors allow comparisons between non-nested hypotheses.*

In contrast to frequentist tests, whether the null hypothesis is nested within the alternative or not does not affect Bayes factors, as it is not a requirement for extracting any of the marginal likelihoods. This allows researchers to make easy and elegant comparisons in settings where classical methods would be hard to implement (Klugkist et al., 2005, and Huntjens et al., 2006).

4. *Bayes factors allow the null hypothesis to be expressed as an interval.*

Frequentist tests usually need the explicit specification of a point value for the parameter of interest under the null, in order to define the distribution of

https://doi.org/10.26219/heal.aueb.6623

the test statistic, and thus derive the p-value. Bayes factors face no such limitation; they can accommodate interval null (and alternative) hypotheses by averaging the likelihood over all possible values for the parameter of interest. All needed is a prior density specifying how that parameter is distributed within the interval in question.

5. *Bayes factors allow for easy monitoring of the accumulation of evidence*.

Although frequentist methods do not forbid researchers from pausing and observing the evidence before a study is complete, such an action introduces complications, because running multiple tests increases the risk of committing type I errors, and special adjustments of the error rate α are needed to remedy this. Bayes factors, unlike frequentist tests, are not premised on the long-run control of error rates. Thus, frequency of testing does not require adjusting the threshold of what constitutes significant evidence. This makes them a more sensible choice in settings where interim analysis is common, as in psychology or clinical trials. This property is not only practical, but also ethical, since Bayesian testing allows the collection of data to stop when *"a point has been proven or disproven, or [when] the data collector runs out of time, money, or patience"* (Edwards et al., 1963), thus ensuring that no resources will be wasted or patients inconvenienced more than is truly necessary. Bayesian inference has this advantage because it acts in accordance with the likelihood principle[6] (Kass & Raftery, 1995). This makes it immune to a whole host of drawbacks and paradoxes that plague frequentist methods, which violate the principle instead (Berger & Wolpert, 1988).

6. *Bayes factors allow researchers to take model uncertainty into account*.

Assume that we seek to compare m potential models (hypotheses) H_1, H_2, …, H_m for explaining the data x at hand. Using prior model odds and the

6 The likelihood principle is a conjecture in statistics and probability theory which states that observations with equal or proportional likelihoods contain the same information about a parameter of interest θ (Casella & Berger, 2002). The implication is that external factors irrelevant to the data, such as sampling intentions or stopping rules, should not affect inference.

https://doi.org/10.26219/heal.aueb.6623

Bayes factor property that $BF_{ik} = BF_{ij} \times BF_{jk}$, we can calculate all of their posterior probabilities using the formula:

$$\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}} \Leftrightarrow \text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}. \tag{1.3.9}$$

When the number of candidate models m is reasonably small as to be manageable, then instead of being forced to choose only one of them, we can use an 'average' model that is the weighted mean of all individual ones, with their posterior probabilities as weights, although many different methods of combining them exist (Raftery et al., 1993, and Madigan & Raftery, 1994). These types of models have demonstrated enhanced predictive performance (Raftery et al., 1996), but even when this is not our goal, this property highlights the ability of Bayesian hypothesis testing to provide a quantification of model uncertainty in the form of a probability. Such a feature is not available in the frequentist framework, and it makes results easier to interpret and understand, even by non-experts[7].

7. *Bayes factors are not biased against the null hypothesis*.

It has been observed that as the sample size gets very large, classical tests systematically reject the null hypothesis (Lin et al., 2013). This causes problems in settings where samples are large by the experiment's very nature, like demographic or social studies. Bayes factors do not have this problem (Kass & Raftery, 1995). This behavior of classical tests may not be present in experiments of smaller size, but the frequentist bias against the null hypothesis is still quietly present, as it is inherent and structural. By conditioning the test statistic on the null hypothesis, p-values only express the 'unusualness' of the observed data *under the null*, hence failing to account for the possibility that the data may be even more unusual under the alternative

7 The correct interpretation of posterior model probability, is the probability of a model being the true one *among those being compared* (assuming that the true model is one of them).

https://doi.org/10.26219/heal.aueb.6623

H_1 (Wagenmakers et al., 2018). The comparative nature of Bayes factors eliminates this problem.

8. *Bayes factors provide a coherent quantification of the evidence.*

The Bayes factor's interpretation as a measure of evidence is intuitive and easy to understand, unlike the *p*-value's, which has spawned a wide variety of previously mentioned misconceptions. A Bayes factor reflects only the evidence provided by the observed data; a *p*-value does not, as it is a tail-end probability dependent on the density of data that could have been observed, but weren't. As Sir Harold Jeffreys put it: *"What the use of [p-values] implies, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure"* (Jeffreys, 1939). Additionally, one can use the formula $BF_{ik} = BF_{ij} \times BF_{jk}$ to measure evidence from multiple comparisons. If, for example, the data support H_1 twice as strongly as they do H_2, and H_2 three times as strongly as they do H_3, then it follows that they support H_1 six times as strongly as they do H_3. No equivalent result exists in the frequentist toolbox (Wagenmakers et al., 2010).

1.3.3 Methods of Computation

The way a Bayes factor is calculated is by either deriving exact solutions or approximations of the marginal likelihoods in its enumerator and denominator, and then computing their ratio. Assuming the existence of a parameter vector $\theta \in \Theta$, following a prior distribution $p(\theta)$, the integral[8] one seeks to calculate is:

$$f(x) = \int_{\Theta} f(x|\theta)\, p(\theta)\, d\theta \,. \tag{1.3.10}$$

8 The notation of the marginal likelihood in this section omits any indication of hypotheses H for simplicity, although what is implied is the same integral as the one in equation (1.3.5).

https://doi.org/10.26219/heal.aueb.6623

Except for a limited number of cases where closed-form expressions of the marginal likelihoods exist, the computation of such integrals is not a straightforward task. In problems where parameter θ is of high dimension, they can quickly become complicated or intractable. Many methods have been put forward by researchers attempting to overcome this problem, ranging from simulation-based to asymptotic ones. An outline of the most important ones follows, describing their methodology along with some advantages and disadvantages.

1. Laplace Approximation

This asymptotic method is based on the assumption that the posterior distribution $f(\theta|x) \propto f(x|\theta)p(\theta)$ is approximately Normal (DiCiccio et al., 1997), and highly peaked around its mode, which is usually the case for large samples (Kass & Raftery, 1995). According to this method, the marginal likelihood is approximated by:

$$\hat{f}_{\mathscr{L}}(x) = (2\pi)^{\frac{d}{2}} |\tilde{\Sigma}|^{\frac{1}{2}} f(x|\tilde{\theta})\, p(\tilde{\theta}), \tag{1.3.11}$$

where d is the dimension of parameter vector θ, $\tilde{\theta}$ is the posterior mode of θ, and $\tilde{\Sigma}$ is the negative inverse of the Hessian matrix of the logarithm of the posterior density, at $\tilde{\theta}$. Having an error of order $O(n^{-1})$, the Laplace approximation has been determined to provide satisfactory results even for problems of moderate sample size, although researchers Kass and Raftery point to problematic behavior when the size is smaller than $5d$ (Kass & Raftery, 1995). Many variants of this method exist. One of them is

$$\hat{f}_{MLE}(x) = (2\pi)^{\frac{d}{2}} |\hat{\Sigma}|^{\frac{1}{2}} f(x|\hat{\theta})\, p(\hat{\theta}), \tag{1.3.12}$$

which is different from the original in that it uses a maximum-likelihood estimator $\hat{\theta}$ instead of the posterior mode. This version offers no improvement in accuracy, but its advantage lies in the fact that many

 https://doi.org/10.26219/heal.aueb.6623

statistical programs have built-in functions to calculate maximum-likelihood parameter estimates, allowing for faster and more efficient computation (Kass & Raftery, 1995).

Another variant is the Laplace-Metropolis estimator (Raftery, 1996, and Lewis & Raftery, 1997), which uses simulated posterior sampling to approximate the quantities $\widetilde{\theta}$ and $\widetilde{\Sigma}$. For MCMC output $\theta^{(1)}$, $\theta^{(2)}$, …, $\theta^{(T)}$, the estimator is given by:

$$\hat{f}_{MH}(x)=(2\pi)^{\frac{d}{2}}|\bar{\Sigma}|^{\frac{1}{2}} f(x|\bar{\theta})\, p(\bar{\theta}), \tag{1.3.13}$$

where $\bar{\theta}=\frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$, and $\bar{\Sigma}=\frac{1}{T-1}\sum_{t=1}^{T}\left(\theta^{(t)}-\bar{\theta}\right)\left(\theta^{(t)}-\bar{\theta}\right)^{T}$.

2. Prior Monte Carlo Approximation

The marginal likelihood (1.3.5) is the likelihood of the observed sample averaged over the prior distribution. Thus it can be approximated by simulating a sample $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}$ from the prior $p(\theta)$ and calculating the mean

$$\hat{f}_{Prior}(x)=\frac{1}{m}\sum_{i=1}^{m} f(x|\theta^{(i)}). \tag{1.3.14}$$

This method is relatively easy and straightforward, but inefficient. The data density $f(x|\theta)$ is frequently peaked compared to the prior $p(\theta)$, resulting in likelihood summands close to zero for most sampled values of θ (Carlin & Louis, 2000).

3. Harmonic Mean Monte Carlo Approximation

An alternative approach to sampling from the prior is sampling from the posterior (Ntzoufras, 2011). Using equation

$$\frac{1}{f(x)}=\int\frac{1}{f(x|\theta)}\frac{f(x|\theta)\,p(\theta)}{f(x)}d\theta=\int\frac{1}{f(x|\theta)}p(\theta|x)\,d\theta, \tag{1.3.15}$$

https://doi.org/10.26219/heal.aueb.6623

it follows (Newton & Raftery, 1994) that the marginal likelihood can be approximated using a posterior sample $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}$ through the estimator

$$\hat{f}_{HM}(x)=\left(\frac{1}{m}\sum_{i=1}^{m} f(x|\theta^{(i)})^{-1}\right)^{-1}. \tag{1.3.16}$$

This method is an improvement compared to the inefficient prior Monte Carlo estimator, but has the drawback of being unstable (Carlin & Louis, 2000).

4. Importance Sampling Approximation

A different option is to sample values of parameter θ from a so-called ‘importance density’, and then to approximate the marginal likelihood integral using a weighted likelihood average where the ratio of prior to importance densities acts as a weight. For a sample $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}$ from importance distribution $\tilde{p}(\theta)$, the marginal likelihood is estimated by

$$\hat{f}_{IS}(x)=\frac{\sum_{i=1}^{m} w_i f(x|\theta^{(i)})}{\sum_{i=1}^{m} w_i}, \quad \text{where} \quad w_i=\frac{p(\theta^{(i)})}{\tilde{p}(\theta^{(i)})}. \tag{1.3.17}$$

The importance sampling approach is an improvement over $\hat{f}_{Prior}(x)$ (Bos, 2002). Note that $\hat{f}_{IS}(x)=\hat{f}_{Prior}(x)$ when $\tilde{p}(\theta)=p(\theta)$ (Perrakis et al., 2014).

5. Schwarz Criterion Approximation

This is not a method of approximating the marginal likelihood, but the Bayes factor itself. It has been proven (Schwarz, 1978) that for sample size n,

$$\lim_{n\to+\infty}\left(\frac{S-\log(BF_{01})}{\log(BF_{01})}\right)=0, \tag{1.3.18}$$

where the amount $S = \log(f(x|\hat{\theta}_0, H_0)) - \log(f(x|\hat{\theta}_1, H_1)) - \frac{1}{2}(d_0 - d_1)\log(n)$ is called the Schwarz criterion, d_i are the dimensions of the parameter vectors under each model, and $\hat{\theta}_i$ are maximum-likelihood estimates. It follows that the quantity *exp*(S) can be used as a rough approximation of the Bayes factor. With a relative error of order $O(1)$, this approach is not very accurate (Kass & Raftery, 1995), but its advantage lies in that it provides a 'shortcut' to the Bayes factor without considerations for any prior distributions (Carlin & Louis, 2000), as they can be difficult to specify properly. This will be discussed in further detail in the next chapter. Lastly, the quantity $-2S$ is called the Bayesian Information Criterion (BIC) and is a popular tool in model selection.

6. Savage-Dickey Density Ratio

This method is similarly used for calculating the Bayes factor, and not the marginal likelihoods that produce it. Suppose we have a parameter vector $\theta=(\varphi,\psi)\in\Theta$, where φ are the parameters of interest, and ψ are nuisance parameters. When nested models H_0: $\varphi = \varphi_0$ versus H_1: $\varphi \neq \varphi_0$ are being compared, the Bayes factor is given by

$$BF_{01} = \frac{f(x|H_0)}{f(x|H_1)} = \frac{p(\varphi=\varphi_0|x, H_1)}{p(\varphi=\varphi_0|H_1)}. \qquad (1.3.19)$$

While the method applies only to nested models, it holds under relatively general conditions, and has the advantage of being easy to use (Wagenmakers et al., 2010).

Many other techniques for computing Bayes factors exist, such as marginal likelihood approximation through Gibbs sampling (Chib, 1995), Metropolis-Hastings sampling (Chib & Jeliazkov, 2001), bridge sampling (Meng & Wong, 1996), and path sampling (Gelman & Meng, 1994). A review of MCMC-based methods is provided by Han and Carlin (2001).

https://doi.org/10.26219/heal.aueb.6623

1.4 Closing Remarks

In this chapter, we sought to explore the foundations of classical (frequentist) hypothesis testing, point out its shortcomings, and introduce the alternative hypothesis testing method of Bayes factors. The next chapter will focus on principles of 'objective' Bayesian hypothesis testing, and how they relate to the prior distribution. Another area of interest is the use of local and non-local priors, and how they affect inference. To our knowledge, such priors have not yet been examined in clinical trial related hypothesis tests, which are the focus of this dissertation.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

CHAPTER 2

PRINCIPLES OF OBJECTIVE BAYESIAN MODEL SELECTION

2.1 Role of the Prior Distribution

Within the Bayesian framework, a general rule for parameter estimation is that as the sample size increases, the amount of information induced to the posterior by the data outweighs and eventually dominates any information conveyed by the prior distribution. Flat or diffuse priors need smaller numbers of observations compared to more informative ones in order to "wash out" any prior effect. However, all priors become asymptotically irrelevant after adding more and more observations, and produce similar posterior distributions determined almost exclusively by the data. This phenomenon does not manifest when dealing with Bayes factors, which are highly sensitive to the choice of parameter priors, even for large samples (Kass, 1993). In fact, this characteristic is one of the main criticisms against Bayesian hypothesis testing (Gill, 2014, Kass & Raftery, 1995, and Wagenmakers et al., 2010), as it not only increases the need for sensitivity analysis in research, but also introduces a variety of complications whereby Bayesian and frequentist tests produce opposing conclusions under certain circumstances. The latter is commonly known as the 'Lindley paradox'[9].

The paradox lies in the fact that, under specific conditions, an increase in the sample size leads to Bayes factors *accepting* the null hypothesis[10] rather

9 Alternatively named the 'Lindley-Bartlett', the 'Jeffreys-Lindley', or the 'Jeffreys paradox'. First formulated in its present form by Dennis V. Lindley (Lindley, 1957), with additional comments by Maurice S. Bartlett (Bartlett, 1957).

10 The null H_0 is assumed to be nested within the alternative H_1.

https://doi.org/10.26219/heal.aueb.6623

than rejecting it, as is the case with frequentist tests. A typical example to illustrate this behavior is the following (Sprenger, 2013):

Consider a normally distributed sample $x \sim N(\theta,\sigma^2)$, with variance σ^2 known and fixed, and suppose one seeks to test the hypothesis H_0: $\theta = \theta_0$ against the alternative H_1: $\theta \neq \theta_0$. Further, consider the prior $\theta \sim N(\theta_0, \sigma_\theta^2)$ under H_1, the function

$$z(x)=\frac{\bar{x}-\theta_0}{\sigma/\sqrt{n}}, \tag{2.1.1}$$

and assume that the sample mean $\bar{x}$ is different from θ_0 at the limit of significance for some level α, for an arbitrary sample size n. In other words, for any size n, a sample $x = x(n)$ can be found such that $z(x) = \pm Z_{\alpha/2}$, since $z(x)$ follows a standard normal distribution under the null hypothesis. Under these circumstances, the Bayes factor of H_0 over H_1 is given by

$$BF_{01}=\sqrt{1+\frac{n\sigma_\theta^2}{\sigma^2}}\exp\left(-\frac{1}{2}\frac{n\sigma_\theta^2}{n\sigma_\theta^2+\sigma^2}z^2(x)\right). \tag{2.1.2}$$

The first implication of equation (2.1.2) is that as the sample size grows, the Bayes factor tends to infinity in support of H_0, even if the data correspond to a p-value that implies rejection. This result comes in opposition to frequentist tests' rejection of null hypotheses for large sample sizes, hence the apparent paradox. A second important implication was noted by Bartlett, who pointed out that the Bayes factor also supports the null H_0 when the prior variance σ_θ^2 of parameter θ becomes very large. In fact, it has been demonstrated that making the prior variance arbitrarily high will systematically lead to preference for the simplest hypothesis regardless of the data (Wagenmakers et al., 2010).

The consequence of the latter result is that diffuse or flat priors are risky choices in Bayesian hypothesis testing. Diffuse priors are used liberally in problems of parameter estimation because they are a convenient way to represent ignorance, and thus lead to "objective" posterior results. Due to the

https://doi.org/10.26219/heal.aueb.6623

Lindley paradox (from Bartlett's side), this approach does not lead to valid results in hypothesis testing when the prior is specified carelessly. The use of relatively non-informative priors will lead to activation of the Lindley paradox and to eventual undermining of the inferential process. The use of improper priors is similarly proscribed, as they lead to divergent marginal likelihood integrals that require unknown multiplicative constants to be fully determined (Ntzoufras, 2011), making the computation of Bayes factors infeasible (Carlin & Louis, 2000).

One solution is to avoid diffuse priors in Bayesian hypothesis testing (Shafer, 1982). Other researchers refer to the absence of the Lindley paradox in the case of interval parameter hypotheses as an indication that the problem is due to the philosophy of point null hypotheses, and not due to the prior distributions (Aitkin, 1991). Another group of researchers, including Jeffreys, suggest not to be bothered by the use of improper priors on nuisance parameters appearing under both hypotheses, since evidence suggests that this approach does not severely affect results (Kass & Raftery, 1995). Another seemingly obvious solution could be the use of the Schwarz criterion, as it has the property of being asymptotically independent of any prior distribution. However, its inaccuracy and ambiguity concerning which sample size is sufficient for the validity of its asymptotic results, make it a subpar choice (Kass & Raftery, 1995).

While some researchers attempt to reconcile the apparent contradictions introduced by the Lindley paradox on a philosophical level (Spanos, 2013, and Robert, 2014), others have followed the more practical approach of proposing modified versions of Bayes factors that allow the use of improper priors, and are more robust to changes in prior variance.

2.2 Bayes Factor Variants

This section presents some alternative versions of the Bayes factor, most of them adapted to accommodate the use of improper parameter priors. A

https://doi.org/10.26219/heal.aueb.6623

summary definition is provided along with a description of their most important properties, advantages, and possible drawbacks.

Pseudo-Bayes Factor

This variant (Geisser & Eddy, 1979, and Gelfand & Dey, 1994) is based on a cross-validation criterion of leave-one-out predictive densities. Let $x = \{x_1, x_2, \ldots, x_n\}$ be the sample, and let $x_{(k)}$ denote the set $x \setminus x_k, k \in \{1,2,\ldots,n\}$. If $x_{(k)}$ is used to turn a prior $p(\theta)$ into a proper posterior $p(\theta|x_{(k)})$, then the pseudo-Bayes factor is given by

$$BF_{01}^{Ps} = \frac{\prod_{k=1}^{n} \left(\int_{\Theta_0} f_0(x_k|\theta_0)\, p_0(\theta_0|x_{(k)})\, d\theta_0 \right)}{\prod_{k=1}^{n} \left(\int_{\Theta_1} f_1(x_k|\theta_1)\, p_1(\theta_1|x_{(k)})\, d\theta_1 \right)}. \tag{2.2.1}$$

Local Bayes Factor

This Bayes factor variant was developed by Smith and Spiegelhalter (1980), and Spiegelhalter and Smith (1982) with the aim of allowing the use of improper priors when comparing nested linear models. To illustrate the concept, assume we compare hypotheses H_0 and H_1, with improper priors p_0 and p_1 respectively for parameter θ, such that

$$\begin{aligned} p_0(\theta) \propto g_0(\theta) \;\Leftrightarrow\; p_0(\theta) = c_0\, g_0(\theta), \\ p_1(\theta) \propto g_1(\theta) \;\Leftrightarrow\; p_1(\theta) = c_1\, g_1(\theta), \end{aligned} \tag{2.2.2}$$

where g_0 and g_1 are functions with finite integrals, and c_0, c_1 are some unknown constants. While these improper priors do not affect posterior densities, their impact on Bayes factors is critical. In this problem, the Bayes factor of the compared hypotheses equals

$$BF_{01} = \frac{c_0}{c_1} \times \frac{\int_{\Theta_0} f(x|\theta_0)\, g_0(\theta_0)\, d\theta_0}{\int_{\Theta_1} f(x|\theta_1)\, g_1(\theta_1)\, d\theta_1}, \tag{2.2.3}$$

https://doi.org/10.26219/heal.aueb.6623

so when the ratio of multiplicative constants is unknown, it is impossible to make any inferences. The solution of the local Bayes factor entails using an imaginary sample x_0 called a 'training set' to estimate the ratio c_0 / c_1. That estimate is given by

$$\widehat{\left[\frac{c_0}{c_1}\right]} = BF_{01}^{(0)} \times \frac{\int_{\Theta_1} f(x_0|\theta_1) g_1(\theta_1) d\theta_1}{\int_{\Theta_0} f(x_0|\theta_0) g_0(\theta_0) d\theta_0}, \tag{2.2.4}$$

where $BF_{01}^{(0)}$ represents the Bayes factor calculated using the training set alone, but is arbitrarily set to 1. Finally, the desired Bayes factor is approximated using the original data:

$$BF_{01} = \widehat{\left[\frac{c_0}{c_1}\right]} \times \frac{\int_{\Theta_0} f(x|\theta_0) g_0(\theta_0) d\theta_0}{\int_{\Theta_1} f(x|\theta_1) g_1(\theta_1) d\theta_1}. \tag{2.2.5}$$

While this approach is imaginative, it raises two important objections. First, no definitive rule on how to construct the training sample exists. Smith and Spiegelhalter state that it has to be the smallest possible sample that results in proper posterior distributions[11] while maximizing support for the simplest hypothesis, but these criteria are somewhat vague. Second, the decision to set the training sample Bayes factor equal to 1 is dubious and lacks rigorous justification (Gill, 2014).

Posterior Bayes Factor

Another variant that allows the use of improper priors is the posterior Bayes factor, proposed by Aitkin (1991), which replaces the usual marginal likelihood ratio with the ratio of posterior means:

11 Such a sample is referred to as a 'minimal sample size'. Specifically, it is the smallest possible sample size that results in finite marginal likelihoods using objective priors, and is typically the number of parameters in a model.

https://doi.org/10.26219/heal.aueb.6623

$$BF_{01}^{P}=\frac{\int_{\Theta_0} f_0(x|\theta_0)\, p_0(\theta_0|x)\, d\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)\, p_1(\theta_1|x)\, d\theta_1}. \tag{2.2.6}$$

Although the use of the posterior is an attempt to make it robust to the prior (Gill, 2014) the double use of the data has drawn criticism, as it results in violation of the likelihood principle, and a tendency to support more complicated models (Ntzoufras, 2011).

Partial Bayes Factor

This variant, proposed by O' Hagan (1995), relies on subsamples of the real data, rather than imaginary ones. Specifically, it subsets the data x into a training set y, uses it to turn improper priors into proper posteriors, and uses those with the rest of the data z to define the Bayes factor $BF_{01}(z|y)$. The undetermined ratio of multiplicative improper prior constants is eliminated by the property that $BF_{01}(z|y)=BF_{01}(x)/BF_{01}(y)$. The partial Bayes factor has the asymptotic property of consistently picking the correct model when the size ratio $|x|$ / $|y|$ tends to infinity (Gill, 2014).

Fractional Bayes Factor

In the same paper where he presented the partial Bayes factor, O' Hagan also presented another variant that aimed to minimize the importance of training set selection. If x and y are the full data and the training set respectively, the author defines their size ratio η = $|y|$ / $|x|$, and uses the formula $f(y|\theta) \approx f(x|\theta)^{\eta}$ to approximate the training sample likelihood through the size-adjusted full sample likelihood. This approximation leads to the fractional Bayes factor, given by

$$BF_{01}^{\eta}(x)=\frac{\int_{\Theta_0} f_0(x|\theta_0)\, p_0(\theta_0)\, d\theta_0}{\int_{\Theta_1} f_1(x|\theta_1)\, p_1(\theta_1)\, d\theta_1}\times\frac{\int_{\Theta_1} f_1(x|\theta_1)^{\eta}\, p_1(\theta_1)\, d\theta_1}{\int_{\Theta_0} f_0(x|\theta_0)^{\eta}\, p_0(\theta_0)\, d\theta_0}. \tag{2.2.7}$$

https://doi.org/10.26219/heal.aueb.6623

This approach replaces the choice of training sample with the less restrictive choice of its size, but creates the need for constant adjustments when Bayes factors have to be updated for new observations (Gill, 2014).

Intrinsic Bayes Factor

This variant was developed by Berger and Pericchi (1996a, 1996b). Similar to the partial Bayes factor, it uses subsamples that can turn improper priors $p_i(\theta_i)$ into proper posteriors, and then uses said posteriors to construct Bayes factors. For subsample $x(l)$, authors define

$$BF_{01}(x(l)) = \frac{\int_{\Theta_0} f_0(x(-l)|\theta_0, x(l))\, p_0(\theta_0|x(l))\, d\theta_0}{\int_{\Theta_1} f_1(x(-l)|\theta_1, x(l))\, p_1(\theta_1|x(l))\, d\theta_1} = BF_{01}^{N} \times BF_{10}^{N}(x(l)), \qquad (2.2.8)$$

where $x(-l)$ denotes all data points except those in subset $x(l)$, $p_i(\theta_i|x(l))$ is a proper posterior of θ, BF_{01}^{N} is the common Bayes factor, and $BF_{10}^{N}(x(l))$ is a Bayes factor using only the data from the subset $x(l)$. The superscript N denotes the use of improper (non-informative) parameter priors, and the final product in equation (2.2.8) demonstrates that any undetermined multiplicative constants from these improper priors necessarily cancel out.

The authors bypass the problem of how to select $x(l)$ by averaging $BF_{01}(l)$ over all possible sample subsets that fulfil the necessary criteria. If S such subsamples exist in total, then the (arithmetic) intrinsic Bayes factor can be calculated by

$$BF_{01}^{A} = \frac{1}{S} \sum_{l=1}^{S} BF_{01}(x(l)), \qquad (2.2.9)$$

but other summary statistics can be used, like a geometric mean, a harmonic mean, or a median. While averaging over all potential training sets fixes the problem of selection, it is at the same time a drawback. In settings where sizeable samples are common (e.g.: social sciences) the number of potential training sets becomes prohibitively large for the method to be used as intended, and random sampling has to be employed instead (Gill, 2014).

https://doi.org/10.26219/heal.aueb.6623

Additionally, computing a summary statistic of the individual Bayes factors leads to multiple use of the data, making it a non-fully Bayesian model selection procedure.

To the extent that the aforementioned Bayes factor variants enable the use of improper priors, they can be considered steps towards a more objective approach to model selection. However, the fact that they still rely on entirely subjective choices (specification of training set or its size) seems to defeat that very purpose, and makes them less than ideal. The question of what constitutes truly objective analysis has been of great interest to statistical researchers, and many have attempted to provide a coherent definition of the concept in the service of Bayesian inference.

2.3 Objective Bayesian Analysis

The need to define 'objective' versus 'subjective' Bayesian analysis has sparked discussions that transcend statistical inference altogether, and take on philosophical dimensions. One such recent suggestion by Gelman and Hennig (2017) has been to abandon this inadequate binary paradigm, and to replace 'objectivity' with concepts such as 'transparency' and 'impartiality', and 'subjectivity' with 'context dependence' and 'awareness of multiple perspectives'. Despite such efforts, no universally agreed upon definition of objective Bayesianism exists, and many Bayesian researchers even go as far as to claim that no form of statistical analysis can be truly objective (Berger, 2006).

Since objective Bayesian model comparison is a subset of objective Bayesian analysis, it naturally follows that the lack of strict definition for the latter applies to the former as well. Despite that, a conventional understanding of objective Bayesian model comparison does exist, and it mainly revolves around prior distributions and how they can be optimized to be as uninformative as possible. Research in this context traditionally involves the

 https://doi.org/10.26219/heal.aueb.6623

proposal of new methods and priors, followed by *post hoc* evaluations of their objectivity, as well as other properties which ensure a well-behaved model comparison methodology.

2.3.1 Criteria for Objective Bayesian Model Comparison

Researchers Bayarri et al. (2012) presented a list of seven strictly defined criteria that future objective priors should fulfil. They are categorized into four different classes:

- One basic criterion
- Three consistency criteria
- One predictive matching criterion
- Two invariance criteria

The notation that follows is:

$$M_0: f_0(x|\alpha) \text{ for the full model,}$$
$$M_i: f_i(x|\alpha, \beta_i) \text{ for other models,}$$

where β_i are vectors of model-specific parameters, α is the vector of common ones, and $p_i(\alpha,\beta)=p_i(\beta_i|\alpha)p_i(\alpha)$ is the joint parameter prior, for $i\in\{1,2,\ldots,N\}$.

Criterion 1: Basic

This criterion states that prior distributions $p_i(\beta_i|\alpha)$ for model-specific parameters should be proper, to prevent the emergence of undetermined multiplicative constants that would render the Bayes factor arbitrary. Furthermore, they should not be excessively vague, in the sense of attributing almost all of their mass outside of reasonable parameter values.

Criterion 2: Model Selection Consistency

Assuming that one of the models being compared is the true one, this criterion states that priors should be chosen in such a way so that the true

https://doi.org/10.26219/heal.aueb.6623

model's posterior probability converges to 1 as the sample size diverges to infinity. If several priors satisfy this requirement, it seems reasonable to prioritize them according to their speed of convergence (Consonni et al., 2018).

Criterion 3: Information Consistency

If for any model M_i a sequence x_n of n-sized datasets exists such that the likelihood ratio

$$\Lambda_{i0}(y_n) = \frac{\sup_{\alpha,\beta_i} f_i(x_n|\alpha,\beta_i)}{\sup_{\alpha} f_0(x_n|\alpha)} \to \infty \quad \text{as} \quad n \to \infty, \quad \text{then} \quad BF_{i0}(x_n) \to \infty. \tag{2.3.1}$$

In other words, asymptotically and for the same sets of data, the ratio of likelihoods averaged with a prior as a weight (Bayes factors), should lead to the same conclusion as the ratio of maximized likelihoods.

Criterion 4: Intrinsic Prior Consistency

It is sometimes possible for model features such as sample size n or even the data themselves to affect the form of a parameter prior. This criterion states that when this is the case, these effects should disappear as n becomes asymptotically large, and that the prior $p_i(\beta_i|\alpha,n)$ should converge to a proper one $p_i(\beta_i|\alpha)$, called an *intrinsic* prior when it exists.

Criterion 5: Predictive Matching

Before this criterion is described, the concept of predictive matching should be defined. Two models M_i and M_j with priors p_i and p_j respectively, are 'predictively matched' for a sample size m when their marginal likelihoods based on an m-sized sample are close (in terms of some distance measure), and 'exactly predictively matched' when they are equal.

The idea behind this criterion is that when using a minimal sample size it should not be possible to draw support for any hypothesis, since all degrees of freedom are expended on setting parameters and none are left for model discrimination. Proceeding from this intuition, the criterion states that for an

https://doi.org/10.26219/heal.aueb.6623

appropriately defined minimal sample size used in comparing two models, parameter priors should be chosen in such a way so as to at least result in predictive matching (Bayes factor close to one), when the ideal outcome of exact predictive matching (Bayes factor equal to one) cannot be obtained.

Criterion 6: Measurement Invariance

This is the shortest of the criteria in terms of description. It states that units of measurement used for observations or parameters should not affect Bayesian answers.

Criterion 7: Group Invariance

If the models being compared have density structures that are invariant to a group transformation G_0, then prior distributions $p_i(\beta_i|\alpha)$ should be chosen so that the marginal distributions $f_i(x|\alpha)=\int f_i(x|\alpha,\beta_i)\,p_i(\beta_i|\alpha)\,d\beta_i$ are also invariant under that transformation.

2.3.2 Frequentist Properties of Bayesian Tests

As previously mentioned, frequentist methods are dominant in science, while Bayesian ones find researchers hesitant to adopt them, in large part due to the stigma of subjectivity. One way to bridge this gap could be to demonstrate that the two approaches are actually complementary, and that seemingly conflicting results are not necessarily irreconcilable. Objective Bayesian methods are the most promising route towards that goal (Bayarri & Berger, 2004), unifying Bayesian and frequentist conclusions in a wide variety of statistical problems, or imbuing Bayesian answers with often needed frequentist properties (Berger, 2006).

Regarding model comparisons, one major consolidation of Bayesian and frequentist results was achieved by connecting Bayes factors to a powerful but underappreciated modification of classical testing, called ‘conditional frequentist testing’. The idea is to allow type I and II error rates to vary by

 https://doi.org/10.26219/heal.aueb.6623

conditioning them on a statistic that measures the strength of the evidence, rather than keeping them constant throughout tests. Building upon existing ideas, researchers Berger et al. (1994) and Berger (2008), outlined this method when comparing simple hypotheses H_0 and H_1 (although it has since generalized to accommodate comparisons of composite hypotheses as well):

- Calculate the usual p-values p_0 and p_1, when H_0 and H_1 are the null hypotheses respectively.
- Define statistic $S = S(x) = \max\{p_0, \ p_1\}$.
- Reject H_0 if $p_0 < p_1$, reject H_1 otherwise.
- Report conditional type I and II error probabilities $\alpha(S)=P(Reject\, H_0|S, H_0)$, and $\beta(S)=P(Accept\, H_0|S, H_1)$.

The connection of conditional frequentist testing to Bayesian testing was achieved when the authors proved that

$$\alpha(S)=\frac{BF_{01}}{1+BF_{01}}, \text{ and } \beta(S)=\frac{1}{1+BF_{01}}, \tag{2.3.2}$$

when assuming equal prior probabilities for the hypotheses. The result benefits both sides, as Bayesians can use objective methods to give frequentist interpretations to their results, and frequentists can calculate conditional error probabilities through the Bayesian route without having to perform complex computations.

Another interesting method of connecting frequentist and Bayesian conclusions in hypothesis testing, is the minimum Bayes Factor (Goodman, 2001, and Page & Satake, 2017), which allows the construction of an "exchange rate" between Bayes factors and p-values, enabling direct comparison of results. The minimum Bayes factor is designed to reflect the greatest amount of evidence against the null hypothesis for a given p-value. In doing so it provides an objective lower bound on Bayes factors, making it an example of an objective Bayesian procedure (Held & Ott, 2018).

 https://doi.org/10.26219/heal.aueb.6623

For tests based on the assumption of a normally distributed (or approximately so) test statistic, the minimum Bayes factor of null H_0 vs H_1 is given by

$$MBF_{01} = \exp\left(-\frac{Z^2}{2}\right), \qquad (2.3.3)$$

where Z is the number of standard errors away from the null value. For example, a p-value of 0.05 (Z = 1.96) corresponds to a minimum Bayes factor of 0.15 (Page & Satake, 2017), meaning that the evidence supports the null 15% as strongly as it does the most plausible alternative.

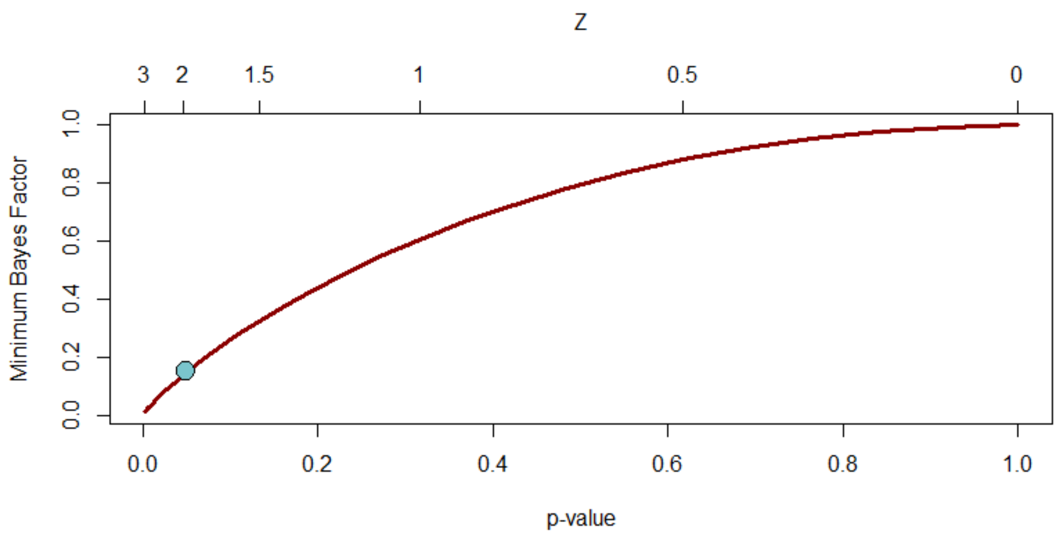

Figure 1: Comparison of p-values and minimum Bayes factors for a two-sided test. Blue dot is the example of p = 0.05.

Minimum Bayes factors have been claimed to be at least as objective as p-values, and have demonstrated that in many cases p-values tend to overstate the evidence against a null hypothesis (Goodman, 1999b). Other techniques that attempt to bridge the gap between Bayesian and frequentist testing are the *posterior* and *partial posterior predictive p-values*, both of which use objective priors with the data to derive posterior densities, and proceed to calculate the p-values through them (Bayarri & Berger, 2004).

 https://doi.org/10.26219/heal.aueb.6623

2.4 Popular Priors

This section provides a short description for a selection of popular priors used in objective Bayesian analysis, their properties, advantages, and disadvantages.

g-Prior

Proposed by statistician Arnold Zellner (1986), *g*-priors are used to model the behavior of covariate coefficients in problems of multiple linear regression. Let m_l be one such model, where

$$Y|X_l, \beta_l, \sigma^2, m_l \sim N_n(X_l \beta_l, \sigma^2 I_n), \tag{2.4.1}$$

$Y \in \mathbb{R}^{n\times 1}$ is the observations vector, $X_l \in \mathbb{R}^{n\times(p_l+1)}$ is the design matrix, p_l is the number of non-intercept covariates, β_l is the vector of covariate coefficients, σ^2 is the variance, and I_n is an identity matrix. The *g*-prior on β_l is given by

$$\beta_l|\sigma^2, m_l \sim N_{p_l+1}\left(\mu_{\beta_l}, g(X_l^T X_l)^{-1}\sigma^2\right), \text{ with } p(\sigma^2|m_l) \propto \sigma^{-2}. \tag{2.4.2}$$

The prior mean μ_{β_l} can be set equal to zero, or specified using imaginary data, preferably generated from the null model. These priors allow for closed-form expressions of posterior distributions, are computationally efficient, and relatively easy to use since the only unspecified hyperparameter is g. For $g = n$, *g*-priors asymptotically lead to Bayes factors that behave like a BIC-based model selection procedure. While consistent in terms of model selection, *g*-priors have the drawback of information inconsistency (Liang et al., 2008, and Consonni et al., 2018).

Hyper-*g* and Hyper-*g*/*n* Prior

Building upon Zellner's suggestion to treat hyperparameter g in *g*-priors as a random variable that follows some prior distribution $p(g)$ (Zellner, 1986),

https://doi.org/10.26219/heal.aueb.6623

researchers Liang et al. (2008) introduced the family of hyper-g-priors, where g has a probability density

$$p(g)=\frac{\alpha-2}{2}(1+g)^{-\frac{\alpha}{2}},\ \ g>0, \tag{2.4.3}$$

which is proper for any $\alpha > 2$. This distribution corresponds to placing a $Beta(1, \alpha/2 - 1)$ prior on parameter $w = g/(1 + g)$, which is a factor controlling the posterior shrinkage of the non-intercept, centered covariate coefficients. One disadvantage of hyper-g priors is their tendency to inflate posterior inclusion probabilities of unimportant covariates towards 0.5 (Consonni et al., 2018). Like g-priors, hyper-g-priors are computationally efficient, and have the added advantage of being information consistent. They are also model selection consistent, except in the case where the true model is the intercept-only model (Liang et al., 2008). This latter limitation is fixed by a slight modification called the hyper-g/n prior, where g has density

$$p(g)=\frac{\alpha-2}{2n}\left(1+\frac{g}{n}\right)^{-\frac{\alpha}{2}},\ \ g>0. \tag{2.4.4}$$

Robust Prior

This prior was proposed by Bayarri et al. in the same paper where the seven criteria of objective Bayesian model selection were presented (Bayarri et al., 2012), and it satisfies all of them in the context of multiple linear regression. For a comparison between models m_0 and m_l, the robust prior under m_l is given by

$$p_l^R(\beta_0,\beta_{l\backslash 0},\sigma)\propto\sigma^{-1}\int_0^{\infty} N_{p_l-p_0}(\beta_{l\backslash 0}|0,g\Sigma_{l\backslash 0})\pi^R(g)dg, \tag{2.4.5}$$

where β_0 is the coefficient vector of the covariates that are common to both models[12], while $\beta_{l\backslash 0}$ is the coefficient vector of covariates exclusive to

12 This *does not* imply that β_0 is the complete covariate coefficient vector of m_0, or that m_0 is necessarily nested within m_l.

 https://doi.org/10.26219/heal.aueb.6623

model m_l. Also $\Sigma_{l\setminus 0}=\sigma^2(V_{l\setminus 0}^T V_{l\setminus 0})^{-1}$, $V_{l\setminus 0}=(I_n-X_0(X_0^T X_0)^{-1}X_0^T)X_{l\setminus 0}$, X_0 and $X_{l\setminus 0}$ are design matrices, p_l and p_0 are numbers of non-intercept covariates under m_l and m_0 respectively, σ^2 is the variance, and

$$\pi^R(g)=\alpha[\rho_l(b+n)]^\alpha (g+b)^{-(\alpha+1)} 1_{\{g>\rho_l(b+n)-b\}}, \tag{2.4.6}$$

with $\alpha > 0$, $b > 0$, and $\rho_l \geq b/(b+n)$. Robust priors have the advantage of providing closed-form expressions for Bayes factors. For a default parameter value of $\alpha = 3$, the hyper-g prior can be derived from a robust prior with parameters $\alpha = 1/2$, $b = 1$, and $\rho_l^{-1} = n + 1$, while the hyper-g/n prior from a robust prior with parameters $\alpha = 1/2$, $b = n$, and $\rho_l^{-1} = 2$ (Consonni et al., 2018).

Shrinkage Priors

When dealing with regression problems of high dimension, it is crucial to identify unimportant regression covariates, and discard them. This way, we can reduce the model to a (computationally) manageable size in terms of parameters. That can be achieved with various regularization methods that introduce a penalty in their estimation procedure, which eliminates components with trivial effects. In the Bayesian framework, such regularization methods can be implemented using special ‘shrinkage priors’ (van Erp et al., 2019). One of the older regularization methods, called Ridge regression (Hoerl & Kennard, 1970) can be carried out in the Bayesian context by having covariate coefficients β_j follow the conditional prior

$$\beta_j|\lambda,\sigma^2 \sim N\left(0,\frac{\sigma^2}{\lambda}\right),\ j\in\{1,2,\ldots,p\},\ \text{ where }\ \lambda\sim Cauchy_{>0}(0,1) \tag{2.4.7}$$

is the penalty parameter that controls the amount of shrinkage (Hsiang, 1975). Another popular regularization method is the LASSO (Tibshirani, 1996), which can be implemented under the Bayesian paradigm by having

 https://doi.org/10.26219/heal.aueb.6623

$$\begin{aligned}&\beta_j|\tau_j^2,\sigma^2 \sim N(0,\sigma^2\tau_j^2),\\&\tau_j^2|\lambda^2 \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right),\ \ j\in\{1,2,\ldots,p\},\\&\lambda \sim Cauchy_{>0}(0,1).\end{aligned} \tag{2.4.8}$$

Many other shrinkage priors have been proposed as default choices for high-dimensional problems, and more are being developed and assessed based on their computational efficiency, frequentist properties, and ease of interpretation. More on the subject can be found in Consonni et al. (2018), and van Erp et al. (2019).

2.5 Point versus Interval Hypotheses in Bayesian Testing

The majority of statistical tests involve a formulation of the null hypothesis as a point, or H_0: $\theta = \theta_0$, referred to as a point null hypothesis. This approach has long served statistical analysts in many fields of science, and researchers have claimed that it is a good approximation to the actual problem in a large number of cases (Berger & Sellke, 1987). Many others however, have expressed criticism against it and have argued that in practice, a point null hypothesis can rarely, if ever, hold exactly (Held & Ott, 2018, Morey & Rouder, 2011, and Kim & Robinson, 2019). In the words of statistician Jacob Cohen: *"[The point null hypothesis] can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false)"* (Cohen, 1992). A consequence which appears in both frequentist and Bayesian testing, is that even when the true value differs from the point null by a margin of no practical importance, the hypothesis may still be rejected given an adequate (yet indeterminate) sample size (Morey & Rouder, 2011).

A suggested solution is to abandon the point null hypothesis formulation in favor of an interval-based one (Fig. 2) (Hodges & Lehmann, 1954).

https://doi.org/10.26219/heal.aueb.6623

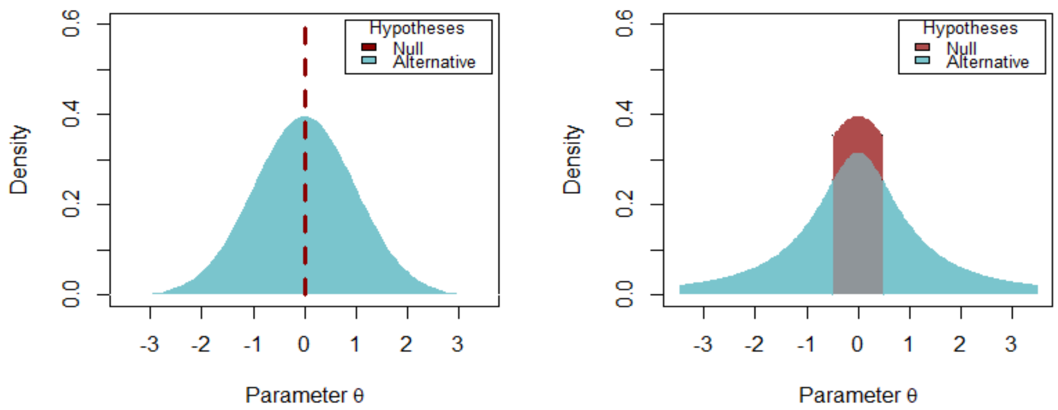

Figure 2: An example of a point null (left) compared to an interval null hypothesis (right).

An example of such a test is the minimum-effect test, comparing hypothesis H_0: $|\theta - \theta_0| \leq \delta$ versus H_1: $|\theta - \theta_0| > \delta$, where δ is determined by the researcher, and signifies the maximum difference between the null and the true value that can be considered unimportant. Many other types of interval null tests exist, and Bayes factors provide an easy way to implement them; all needed are two prior distributions for the parameter in question, with supports on the parameter value regions defined by each hypothesis.

2.5.1 Local and Non-Local Prior Approaches

When both null and alternative hypotheses are intervals, it is not sufficient to only define how the parameter of interest is distributed under the alternative; a similar distribution is required for the null hypothesis as well. Suppose null H_0 is compared to alternative H_1 for parameter $\theta \in \Theta$, such that

$$\begin{aligned} H_0 &: \theta \sim p_0(\theta),\ \ \theta \in \Theta_0, \\ H_1 &: \theta \sim p_1(\theta),\ \ \theta \in \Theta_1, \end{aligned} \tag{2.5.1}$$

https://doi.org/10.26219/heal.aueb.6623

where $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$, and $p_0(\theta) = 0 \ \forall \theta \in \Theta_1$. When prior $p_1(\theta)$ is positive on Θ_0, it is called a 'local alternative prior', and H_1 a 'local alternative hypothesis'.

Historically, most Bayesian hypothesis tests have relied on local priors, but practical and philosophical arguments against their use exist. A general objection is that the existence of a parameter region which is consistent with both hypotheses eliminates any notion of meaningful distinction between them. In the case of interval null tests, it has been observed that hypotheses with common support can make Bayes factors converge to finite non-zero values for large sample sizes (Morey and Rouder, 2011). This behavior violates the model selection consistency criterion of objective analysis. In the case of point null tests, an undesirable feature is a severe imbalance in the rate of accumulation of evidence. When the alternative hypothesis is true, evidence in its favor accumulates at an exponential rate, but when the null is true, evidence in its favor accumulates at a sublinear rate (Johnson & Rossell, 2010).

A solution to these problems is the use of alternative priors that have no common support regions with the null. Morey and Rouder (2011) proposed the use of 'non-overlapping' priors for interval null tests, while researchers Johnson and Rossell (2010) proposed 'non-local' priors for point null tests. Such priors improve the balance between rates of evidence accumulation, and vastly enhance the asymptotic ability of Bayes factors to detect true hypotheses (Morey & Rouder, 2011, and Johnson & Rossell, 2010). Non-overlapping priors are straightforward (Fig. 30, Appendix), but as non-local ones are more complicated, a strict definition is necessary. Here, we adopt the definition of Johnson and Rossell (2010). Under the conditions described in (2.5.1), prior $p_1(\theta)$ is called a 'local alternative prior density' when

$$\exists \varepsilon > 0 : p_1(\theta) > \varepsilon \ \ \forall \theta \in \Theta_0, \tag{2.5.2}$$

and a 'non-local alternative prior density' when $\forall \ \varepsilon > 0 \ \exists \ \zeta = \zeta(\varepsilon)$ such that

https://doi.org/10.26219/heal.aueb.6623

$$p_1(\theta) < \varepsilon \;\; \forall\, \theta \in \Theta : \; \inf_{\theta_0 \in \Theta_0}\{|\theta - \theta_0|\} < \zeta \,. \tag{2.5.3}$$

Such a definition imposes a smooth transition to zero on the alternative prior when moving from Θ_1 to Θ_0, visualized in Fig. 3 and 4.

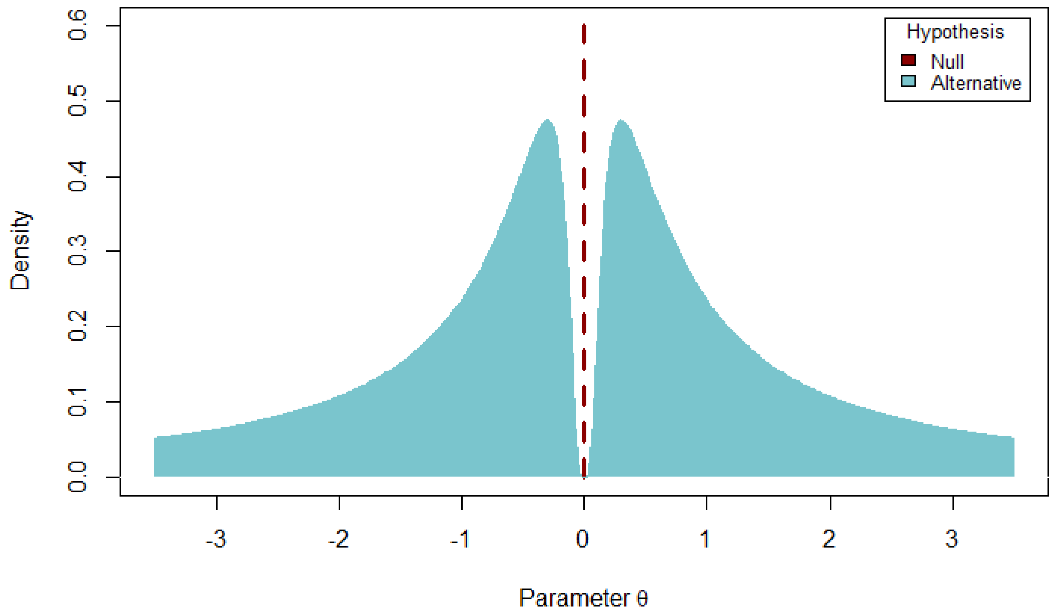

Figure 3: Point null hypothesis with non-local alternative prior.

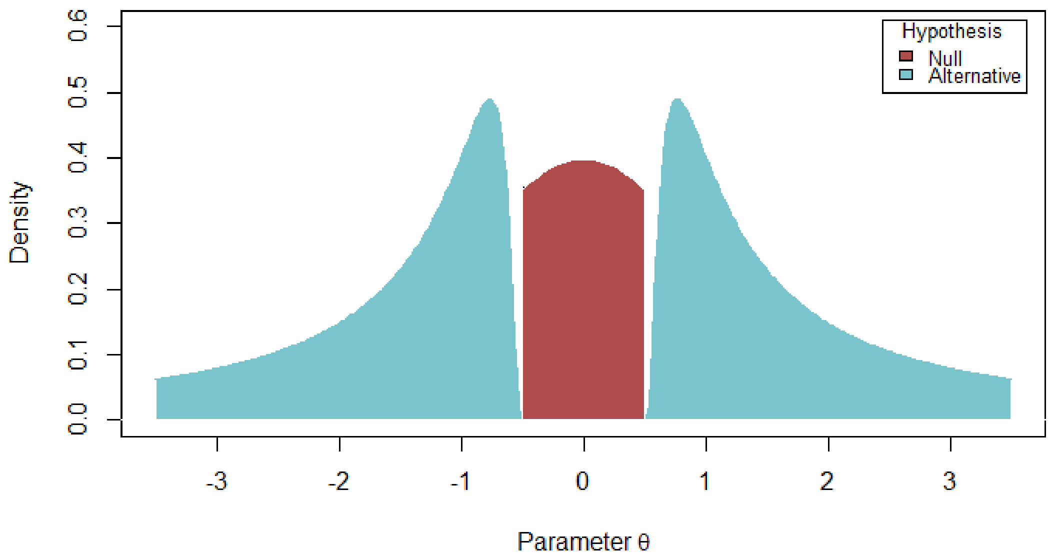

Figure 4: Interval null hypothesis with non-local alternative prior.

https://doi.org/10.26219/heal.aueb.6623

2.6 Closing Remarks

While the various schools of objective Bayesianism succeed in resolving many deficiencies of broader Bayesian inference, they are not entirely free of shortcomings themselves. In particular, the extensive use of improper priors in objective Bayesian analysis is criticized as an erroneous utilization of probability theory, since improper priors are not true probability distributions.

Another criticism is the myopic focus on the prior distribution when it comes to subjectivity. The form of the likelihood, following from the choice of model, has the capacity to affect results, just like the prior. Therefore, the degree to which choice of model is arbitrary, is the degree to which approaches that seek to eliminate subjectivity should take it into consideration. However, it is often the case that objective priors cannot be chosen independently of model, with the latter being a prerequisite for the former. Objective Bayesians claim that a statistical model provides the general context, without which it is not possible to define the notion of objectivity. Hence, any drawbacks resulting from the model-dependent priors are the price that we have to pay in order to achieve a certain degree of objectivity (Berger, 2006).

Finally, objective Bayesian methods lack the solid axiomatic foundations that subjective Bayesian methods have, resulting in a variety of incoherences like the marginalization paradox (Dawid et al., 1973). The topic of objective Bayesian analysis is still under intense research. New objective priors are being proposed and added to the existing arsenal, which is already large enough to overwhelm any casual analyst. Having such a great variety of objective Bayesian approaches at hand can distract from the fact that subjective Bayesian methods work sufficiently well in the majority of problems (Fienberg, 2006).

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

CHAPTER 3

HYPOTHESIS TESTING IN CLINICAL TRIALS

3.1 Purpose and Design of Clinical Trials

Generally speaking, clinical trials are experiments that seek to measure the safety or efficacy of different medical interventions in human subjects. The focus of this chapter is on a particular subset of broader clinical trials, known as 'randomized controlled trials' (RCTs), which draw conclusions after a period of observing patients who have been randomly assigned to two or more groups. One group, called the 'control group', receives the standard treatment (or no treatment at all if it does not exist), and the other group(s) receive(s) the new or competing treatments. The aim of such a trial is to determine the efficacy of the new treatment in comparison with the standard or other competing treatments[13]. RCTs have gained significant ground in medical research during the past decades. When properly conducted, they are the most reliable technique for evaluating medical interventions (Friedman et al., 2015, and Pocock, 2013). In the following of this dissertation, 'clinical trials' are to be understood as referring to RCTs, although the two terms are not generally interchangeable in the academic literature.

3.1.1 Clinical Trial Phases

Before a new drug can be mass produced and released to the public, it has to undergo a thorough, formal examination in order to establish its safety and

13 Concerns of whether individual patients are given optimal medical care are outside the scope of clinical trials.

https://doi.org/10.26219/heal.aueb.6623

efficacy. This process is typically organized into four phases, each with a different structure and goal.

Phase I

Generally performed on small groups of healthy volunteers, phase I trials are conducted to establish drug safety. Their aim is to determine the drug dose that can be administered without resulting in unacceptable toxicity levels. Initial estimates are commonly extrapolated from tests on animals. After multiple dosage-escalating experiments, the maximum tolerated dose (MTD) is determined, which corresponds to the maximum accepted level of toxicity. For example, it could be set as the dosage level at which a specified percentage of subjects start to develop serious side-effects. The size of a phase I drug trial usually ranges from 20 to 100 subjects (Mahan, 2014). These sample sizes are flexible rules of thumb that are not strictly adhered to in any trial phase.

Phase II

After the MTD is established, drug trials move to phase II, where the objective is to determine the effectiveness of a drug. They are conducted on patients suffering from the disease that the drug is targeting (Armitage et al., 2008), and monitor their response. While the typical size of a phase II trial is greater than of a phase I, ranging from 100 to 300 subjects (Mahan, 2014), it is still too small to provide adequate power for confirming treatment effects (Friedman et al., 2015). Phase II trials can be thought of as screening processes that select the most promising of drugs that can proceed to the next phase of more extensive testing (Armitage et al., 2008). Optimal dose finding is also a usual goal for a phase II trial.

Phase III

In this stage, trials are usually comparative and *confirmatory*. They seek to establish treatment efficacy according to a prospectively defined level of evidence. Specifically, the drug under examination is compared to the

https://doi.org/10.26219/heal.aueb.6623

standard form(s) of therapy, in order to evaluate its effectiveness. These types of comparative trials are the most scientifically rigorous. They are the most extensive clinical investigations of new treatments, and may require large groups of subjects, even up to 3,000 (Mahan, 2014), in order to achieve adequate power. Some researchers reserve the term 'clinical trial' only for phase III experiments (Pocock, 2013), while referring to phases I and II as 'clinical studies' (Friedman et al., 2015). Statistical testing regarding equivalence, superiority or non-inferiority of a new treatment is conducted during this phase.

Phase IV

After completing phase III trials and being approved for public use, drugs continue being monitored to identify potential efficacy and safety problems that failed to be detected previously. Phase IV trials can be thought of as 'post-marketing' surveillance studies that examine how the drug works in the real world. Although the scientific value of their conclusions is limited (Pocock, 2013), they have the potential to detect adverse effects, and even lead to drugs being recalled from the market (Mahan, 2014).

3.1.2 Principles of Clinical Trial Design

Clinical trials are essentially statistical experiments. Thus, they are subject to basic principles of scientific experimentation, such as replication, control of variability, and minimization of bias (Pocock, 2013, and Armitage et al., 2008). The execution of a scientifically rigorous and valid clinical trial largely depends on its design, which has to be planned carefully in advance, if the desired result is to be achieved. This subsection lists the most important aspects of a clinical trial's design, and the considerations that have to be made by the researchers with regards to every one of them.

https://doi.org/10.26219/heal.aueb.6623

Randomization

As stated previously, clinical trials are studies of comparative nature, because they categorize subjects into different 'intervention' and 'control' groups, in order to monitor their responses to new therapies. The assignment of a subject to either the intervention or the control group, is not determined by the researchers, but by a process of randomization. This leads to a more balanced allocation of subjects across groups with regards to their prognostic characteristics (e.g.: age, sex, smoker or non-smoker, etc.), ensuring that differences in treatment outcomes will not be due to systematic reasons. For example, an unbalanced allocation of smokers into the study groups of a trial measuring lung health, would lead to biased results. Such a phenomenon is referred to as 'confounding'. Randomization of subjects is the standard method for suppressing the effects of confounding, thus minimizing bias. More importantly, random allocation guarantees the validity of statistical tests, by making it possible to ascribe a probability distribution to the differences in group outcomes (Friedman et al., 2015). Failure to perform randomization properly can result in a trial that is just as biased as a non-randomized one (Pocock, 2013).

Blinding

Even though proper randomization avoids bias from one source, it does not entirely eliminate it. Patients' knowledge of what treatment they receive can induce psychological bias that influences their responses, and thus distorts the trial's measurements. A way to eliminate this is to mask the treatments from the patients, or to use inactive duplicates of their medicine, known as placebos. For example, if an old treatment A (control group) is compared to a new treatment B (intervention group), they can be formulated to look similar to each other, and clinical staff may administer them to patients according to their group, without revealing each treatment's identity. In the case where a new treatment is being compared to the absence of treatment, the control group receives a placebo version of the therapy administered to the intervention group. This kind of masking where only the

https://doi.org/10.26219/heal.aueb.6623

patients are unaware of their treatment is called 'single-blinding'. However, knowledge of treatment by the clinical staff can also lead to bias, as the staff is responsible for monitoring, recording, and evaluating the patients' responses. This can be avoided by concealing the identity of the treatment from both patients and physicians, in what is called 'double-blinding'. While double-blinding is not always possible for more invasive treatments such as surgery or chemotherapy, it is ideal for eliminating two potential sources of bias during treatment evaluation (Friedman et al., 2015).

Trial Size

Clinical trials require an adequately large sample size, so that statistical tests can have enough power to detect treatment effects. For any given significance level α, power $1 - \beta$, and expected treatment effect δ, the necessary sample size n can be estimated through power calculations. Choice of treatment effect δ is usually based on what has been discovered in earlier (i.e.: phase II) trials, or trials for treatments with similar mechanisms of action. As the choice of δ can be arbitrary, even in the presence of prior knowledge, power calculations offer a mere approximation of the required trial size (Armitage et al., 2008). For example, if the expected observed effect δ is overestimated, the sample size will turn out to be insufficient, resulting in an underpowered trial. Ultimately, the final choice is up to the researchers, who are tasked with defining a trial size large enough to detect important effects, yet small enough so as not to render the experiment infeasible or unreasonably costly.

Protocol

A crucial part of every clinical trial is the protocol. It is a formal document which describes in exhaustive detail every aspect of the study's design, conduct, and analysis. It is written before the implementation of the trial and it serves as a well planned guide. A protocol not only assists communication among the trial's research staff, but also provides information regarding the trial to third parties such as inspectors or regulatory agencies. It

https://doi.org/10.26219/heal.aueb.6623

describes the trial's motivation and goals, how patients are to be enrolled, randomized, examined, how data are to be collected, analyzed, and assessed, along with various administrative details and sources of funding. An important part of the protocol is the statistical analysis plan (SAP), in which every detail of the statistical analysis to be conducted is described (e.g.: tests to be performed, adjustments for multiple comparisons, handling of missing data, etc). Various guidelines have been published on how to write proper protocols that adhere to international standards, such as the SPIRIT 2013 Statement (Chan et al., 2013).

3.1.3 Ethical and Legal Considerations

Since clinical trials are experiments on human subjects, several ethical issues arise that have to be dealt with. At their core, these ethical issues are mainly about finding a balance between patient suffering, and medical progress through experimentation. One major ethical conundrum is the claim that administering placebos to patients amounts to deceiving them (Pocock, 2013). Another ethical problem arises when a study is not yet complete, but accumulating evidence starts to indicate that one treatment may be superior to another. Under these conditions, it could be considered unethical for a doctor to continue providing an inferior treatment to one group of patients. These kinds of problems have been studied by medical and legal experts, leading to the creation of international ethical guidelines that clinical research has to adhere to, such as the Declaration of Helsinki (World Medical Association, 2001). Ethical considerations may also be codified into national law, whereby each country sets its own standards.

https://doi.org/10.26219/heal.aueb.6623

3.1.4 Regulation of Statistical Aspects of Clinical Trials

Claims concerning the efficacy of proposed new drugs translate to statistical hypotheses to be tested in the light of data emerging from appropriate studies (e.g.: phase III randomized controlled trials). Claims can also refer to the efficacy of generic or biosimilar products, in which case the hypothesis of interest is the equivalence between the currently approved product, and its potential generic/biosimilar replacement.

The use of statistics in drug authorization is regulated (International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use, 1998) in order to ensure public health safety. There are several regulatory guidelines dictating the statistical methods to be employed for specific inferential problems commonly encountered in drug development (European Medicines Agency, 2014 and 2017, and U.S. Food and Drug Administration, 2017). Such guidelines are mostly driven by the desire for strictly *confirmatory* inference regarding drug efficacy. The strive for 'objectivity' in decisions concerning drug licensing, has given frequentist statistics an indisputably dominant position in the field of drug development. Frequentist hypothesis testing (i.e.: null hypothesis significance testing, or NHST) is founded on the concepts of type I (false positive) and II (false negative) errors and their long-run rates of occurrence, which in turn are established as the decision criteria in drug authorization.

Bayesian methods, however, are receiving increased attention and are being encouraged for use in specific areas of medical research, despite not being premised on the long-run control of type I and II errors. For example, the U.S. Food and Drug Administration (FDA) has released guidelines on the use of Bayesian statistics, where it urges a simulation-based evaluation of clinical trial operating characteristics (type I and II errors rates) (U.S. Food and Drug Administration, 2010). Apart from highlighting the critical importance of controlling type I and II error rates in clinical research (a traditionally frequentist achievement) this also indicates that Bayesian methods can be appropriately adjusted to fulfil similar requirements.

 https://doi.org/10.26219/heal.aueb.6623

3.2 Tests for Superiority, Equivalence, Non-Inferiority

The three most common types of tests conducted in clinical trials are the superiority, equivalence, and non-inferiority tests. This section presents the motivations behind them, and how they are implemented within the frequentist framework.

Superiority Testing

Traditionally, most clinical trials have been conducted in order to establish the superiority of a new treatment to a standard treatment (or to the absence of treatment) (Friedman et al., 2015). That can be determined by a one-sided test called a superiority test, formulated as

$$\begin{aligned} H_0&:\theta=0, \\ H_1&:\theta>0, \end{aligned} \tag{3.2.1}$$

where θ is the treatment effect difference ($\theta_{\text{Treatment}} - \theta_{\text{Control}}$, assuming higher is better). This is equivalent to the lower end of a $100(1-\alpha)\%$ confidence interval for θ being greater than zero (Fig. 5).

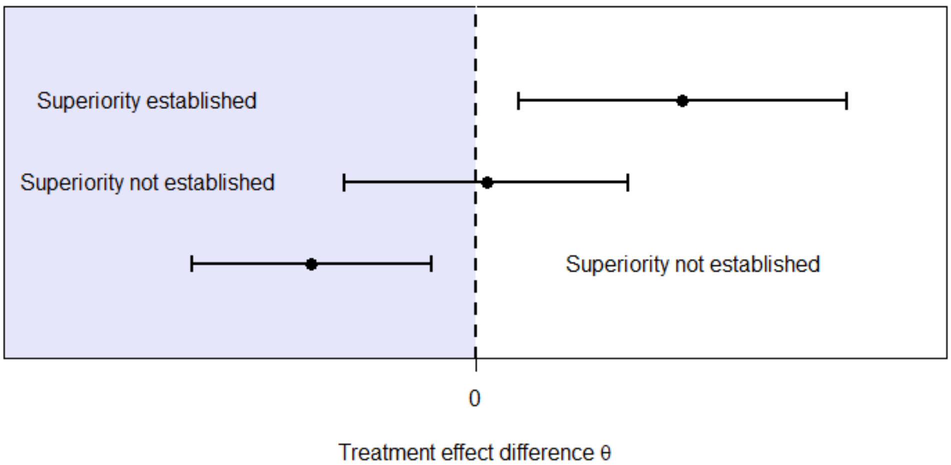

Figure 5: Visualized confidence interval conclusions for superiority testing.

https://doi.org/10.26219/heal.aueb.6623

Equivalence Testing

A common goal of clinical trials is not to determine whether a new treatment is superior to the standard one, but whether the two treatments are equivalent in their effects. If equivalence is established, no other better treatment is available, and the new treatment is known to have lower cost, fewer negative side-effects, or other advantages, then it makes sense to replace the standard treatment with the new one. Equivalence is not meant in the strict sense that the two treatments have *exactly* equal effects, but rather that they are close enough to be considered equivalent. This "closeness" is defined by an *equivalence margin* that is chosen by the researchers, and signifies the maximum acceptable difference between the treatment effects that can be considered unimportant. In terms of hypotheses, equivalence tests are expressed as

$$\begin{aligned} &H_0 : \theta \notin (\theta_L, \theta_U), \\ &H_1 : \theta \in (\theta_L, \theta_U), \end{aligned} \tag{3.2.2}$$

where θ is the treatment effect difference, and $\theta_L < 0$ and $\theta_U > 0$ are the lower and upper equivalence margins respectively. Equivalence can only be asserted when H_0 is rejected. This can be determined by conducting two one-sided tests (TOST), through the following steps (Armitage et al., 2008):

1. Test H_{10}: $\theta = \theta_L$ against H_{11}: $\theta > \theta_L$, and derive p-value p_1.
2. Test H_{20}: $\theta = \theta_U$ against H_{21}: $\theta < \theta_U$, and derive p-value p_2.
3. If $p = \max\{p_1, p_2\} < \alpha/2$, we reject H_0 and assert equivalence.

Visually, equivalence is demonstrated when a $100(1 - \alpha)\%$ confidence interval lies entirely within (θ_L, θ_U) (Fig. 6). In classical point null equivalence tests where H_0: $\theta = 0$ and H_1: $\theta \neq 0$, it is not possible for p-values to quantify evidence *for* H_0. The TOST procedure constitutes a creative workaround by moving the assumption of equivalence from the null to the alternative hypothesis. This enables quantification of evidence *against* non-equivalence, and hence *for* equivalence.

https://doi.org/10.26219/heal.aueb.6623

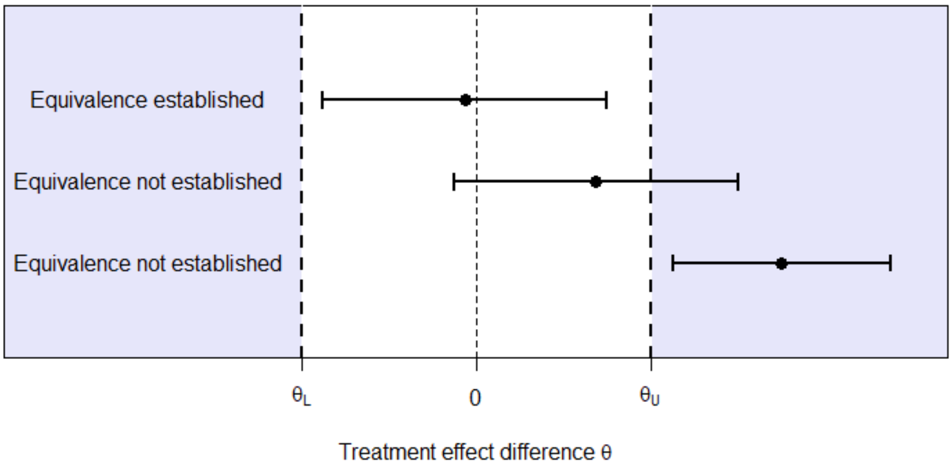

Figure 6: Visualized confidence interval conclusions for equivalence testing.

Non-Inferiority Testing

Following a motivation similar to the one behind equivalence tests, non-inferiority tests have the goal of demonstrating that a new treatment is not worse than the standard treatment. Just like in equivalence tests, the phrase "not worse than" is not literal, but defined by an effect difference between some negative margin and zero (assuming positive values imply improvement). A value within this interval technically implies an inferior treatment, but researchers tolerate it for the sake of the treatment's other potential benefits. The hypotheses being compared are

$$H_0 : \theta = \delta, \\ H_1 : \theta > \delta, \tag{3.2.3}$$

where $\delta < 0$ is the non-inferiority margin. Rejecting the null and establishing non-inferiority is equivalent to a $100(1-\alpha)\%$ confidence interval with its lower bound greater than δ (Fig. 7).

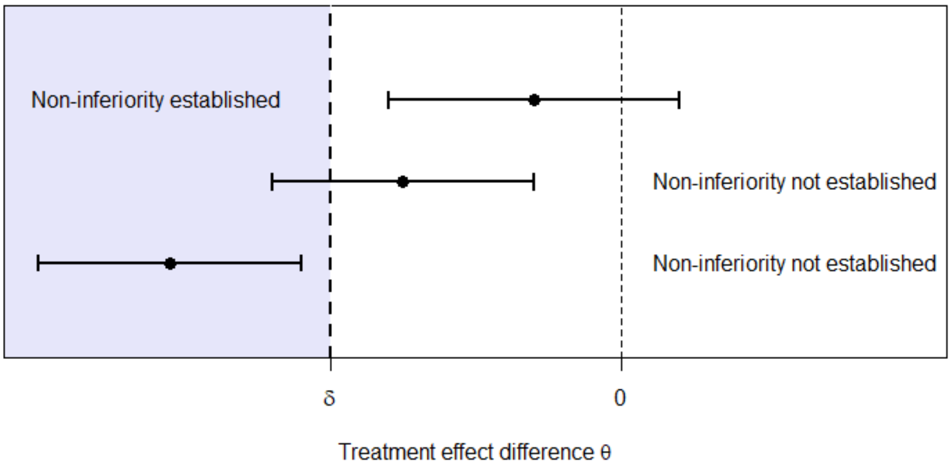

Figure 7: Visualized confidence interval conclusions for non-inferiority testing.

Superiority, equivalence, and non-inferiority tests are crucial in biomedical research, but their implementation from a frequentist perspective makes them vulnerable to the drawbacks inherent in frequentist inference, as described in Chapter 1. Bayesian methods in general, and Bayes factors in particular, resolve many of these deficiencies, and have inferential properties that make them increasingly popular in clinical trials (van Ravenzwaaij et al., 2019).

3.3 Bayesian Methods in Clinical Research

Despite the historical dominance of frequentist approaches, Bayesian ones have become increasingly prominent in broader biomedical research (Xuefeng & Yunling, 2011, and Ashby & Smith, 2000). Areas other than phase III clinical trials include diagnostic testing, epidemiology, survival analysis and meta-analysis (Ashby & Smith, 2000, and Ibrahim et al., 2012). Due to its flexibility, the Bayesian approach is better equipped to deal with complex models, such as longitudinal or hierarchical ones, which are common in health research (Xuefeng & Yunling, 2011, Ibrahim et al., 2012, and Berry, 2006). Furthermore, to the extent that the goals of confirmatory clinical trials are

https://doi.org/10.26219/heal.aueb.6623

decision-oriented rather than inferential, it has been argued that Bayesian decision rules are an optimal choice (Ashby & Smith, 2000).

Regardless of whether the goal is decision-making or inference however, the incorporation of evidence from past studies can be of great importance. Through the prior distribution, the Bayesian framework allows for elegant incorporation of historical data and expert opinion into any ongoing study. Spiegelhalter et al. (1993, 1994) have promoted the use of a wide range of such prior distributions, loosely categorizing them into:

- Vague or reference priors, used to represent ignorance by being minimally informative.
- Clinical priors, formalizing the opinions of well-informed medical experts.
- Sceptical priors, representing the beliefs of researchers unenthusiastic about the new treatment, and their doubts that it will have any large effect.
- Enthusiastic priors, representing beliefs opposite to those behind sceptical priors, and a reluctance to stop the experiment even when evidence supporting the null hypothesis (no treatment effect) has started to emerge.

Another proposed prior that utilizes historical data is the power prior (Ibrahim et al., 2012, and Campbell, 2017)

$$p(\theta|x_0, \alpha_0) \propto p_0(\theta) f(x_0|\theta)^{\alpha_0}, \tag{3.3.1}$$

where x_0 are the historical data, p_0 is the prior before x_0 were observed, and α_0 is a scalar from 0 to 1 that controls the influence of the historical data. For $\alpha_0 = 1$, the power prior is equal to the posterior from the old data, while $\alpha_0 = 0$ is equivalent to historical data having zero influence. Using a power prior in a problem with new data x and historical data x_0 returns a posterior

$$p(\theta|x, x_0, \alpha_0) \propto p_0(\theta) f(x_0|\theta)^{\alpha_0} f(x|\theta). \tag{3.3.2}$$

https://doi.org/10.26219/heal.aueb.6623

When it comes to phase III clinical trials, Bayesian methods not only offer a basis for inference, but also for the regulation of various aspects of the trial's design, such as sample size (Whitehead et al., 2008). They are thus particularly well-suited to adaptive design trials, where adjustments take place according to the observed data (Jack Lee & Chu, 2012, and Chang & Boral, 2008). In an example related to stopping rules, researchers may employ data monitoring and prior-to-posterior analysis to predict the probability of observing a certain effect $\delta \sim p(\delta|x)$, if the trial continues (Spiegelhalter et al., 1993). Based on that prediction, they can elect to either resume the trial, or to terminate it early as futile and save resources.

When pondering the advantages and disadvantages of the frequentist and Bayesian camps, one should not assume that they are necessarily incompatible. In fact, there is increasing evidence that the two outlooks converge into a cooperative future which capitalizes on the strengths of both (Jack Lee & Chu, 2012). For example, Chang and Boral (2008) have proposed a Bayesian-frequentist hybrid approach for phase III clinical trials, while Inoue et al. (2005) have discovered that some Bayesian and frequentist methods for determining sample size give identical answers.

The reconciliation of Bayesian and frequentist outlooks is not simply a potential advantage, but sometimes an obligation for researchers. Many regulatory agencies and clinical review boards still insist on an explicitly frequentist assessment of each study's operational characteristics, even when the analysis itself is purely Bayesian (U.S. Food & Drug Administration, 2010, Berry, 2006, and Chang & Boral, 2008). This has primarily to do with keeping the (overall) type I error rate α below conventional benchmarks (e.g.: 5%), thus guarding against false-positive results. An intuitive way of achieving this when testing with posterior distributions could be to set a small threshold α, and reject the null H_0: $\theta \leq \theta_0$ when $p(\theta \leq \theta_0 \mid x) \leq \alpha$, but other ways have been proposed (Quan et al., 2019).

A second important consideration relates to the power of a Bayesian test to detect differences when they exist. Consider a common approach to power calculations, assuming a normal test statistic $X_n \sim N(\delta, \sigma^2/n)$, where δ

https://doi.org/10.26219/heal.aueb.6623

represents the effect difference, and δ_α is a specific difference that we would like the test to detect. The sample size n to achieving power $1-\beta$ is selected so that it satisfies $P(X_n > 1.96\,\sigma/\sqrt{n} | \delta = \delta_\alpha) = 1-\beta$. A Bayesian variant of this (Spiegelhalter et al., 1993) could be to let δ be distributed according to a prior $p(\delta)$, and then to calculate the unconditional power

$$\int P(X_n > 1.96\,\sigma/\sqrt{n} | \delta)\, p(\delta)\, d\delta. \tag{3.3.3}$$

Another way to estimate an experiment's type I and II error rates for different sample sizes is by performing repeated Monte Carlo simulations (U.S. Food and Drug Administration, 2010). Evidently, even though Bayesian methods are not designed to control long-run error rates, they can be imbued with such frequentist properties without losing their distinctive advantages.

3.3.1 Present Bayes Factor Approaches

A common objective in medical research is the comparison of the means of two groups. A popular way to do this is through the two-sample t-test, assuming equal variances (Gönen et al., 2005). For independent observations x_{ij}, $j \in \{1, 2, \ldots, n_i\}$, $i \in \{1, 2\}$ following a $N(\mu_i, \sigma^2)$ distribution, hypothesis H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$ is tested using the statistic

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p/\sqrt{n_\delta}}, \tag{3.3.4}$$

which follows a t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. The 'effective sample size' is $n_\delta = (n_1^{-1} + n_2^{-1})^{-1}$, and the common (or 'pooled') variance s_p^2 is given by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{3.3.5}$$

https://doi.org/10.26219/heal.aueb.6623

Bayes Factor of Gönen et al. (2005)

Under the Bayesian paradigm, the two-sample comparison requires the specification of prior distributions for the difference $d = \mu_1 - \mu_2$, as well as for the variance σ^2. A straightforward choice could be to place a normal prior on $d \sim N(\lambda, \sigma_d^2)$, and an improper prior on σ^2. However, this parameterization can be problematic, as it does not take into consideration the scaling of d, and thus the choice of the variance hyperparameter σ_d^2 is unclear. A solution can be offered by switching the parameterization from $\{\mu_1, \mu_2, \sigma^2\}$ to $\{\mu=(\mu_1 + \mu_2)/2, d, \sigma^2\}$, and placing a prior on the standardized difference $\delta = d/\sigma$. The benefit of this approach is that it provides an intrinsic scaling for the effect sizes that is easier to model and applies broadly to different populations (Rouder et al., 2009, and Gönen et al., 2005). Under the new parameterization, the model describing the data is given by

$$\begin{aligned} &Y_i = \mu + \frac{\delta\sigma}{2} - \delta\sigma X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \\ &\delta \sim N(\lambda, \sigma^2_{\delta}), \\ &p(\mu, \sigma^2) \propto 1/\sigma^2 \end{aligned} \tag{3.3.6}$$

and X_i is a dummy variable indicating whether a subject belongs to the control group or not. Based on the above formulation, Gönen et al. (2005) obtained a closed-form Bayes factor for testing H_0: $\mu_1 = \mu_2 = \mu$ against H_1: $\mu_1 \neq \mu_2$. Under assumptions of normally distributed, homoskedastic, and independent data, and the priors in equation (3.3.6), the Bayes factor is given by

$$BF_{01} = \frac{T_v(t|0,1)}{T_v(t|n_\delta^{1/2}\lambda, 1+n_\delta\sigma_\delta^2)}, \tag{3.3.7}$$

where t is the observed value of the two-sample test statistic (with equal variances assumed), and $T_v(x \mid a, b)$ is the density function of a non-central t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom, location parameter a and scale parameter $b^{1/2}$. For prior mean of δ equal to $\lambda = 0$, which can be thought

 https://doi.org/10.26219/heal.aueb.6623

of as expressing prior ignorance concerning effect direction, the Bayes factor simplifies to

$$BF_{01} = \left[\frac{1 + t^2 / v}{1 + t^2 / \{ v (1 + n_\delta \sigma_\delta^2) \}} \right]^{-\frac{(v+1)}{2}} \times (1 + n_\delta \sigma_\delta^2)^{\frac{1}{2}}. \quad (3.3.8)$$

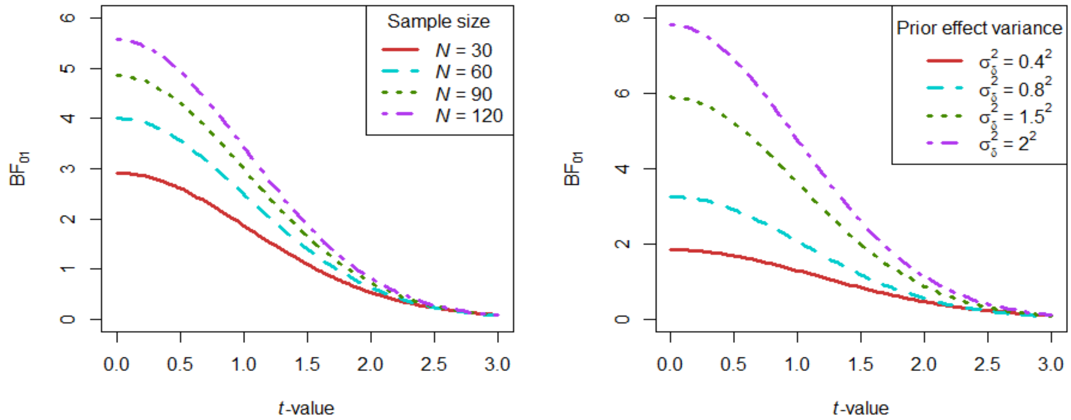

Figure 8: Simplified Gönen et al. (2005) Bayes factor, for different (full) sample sizes N, and prior effect variances σ_δ^2. The left graph assumes $\sigma_\delta^2 = 1$. The right graph assumes N = 60.

Bayes Factor of Wang and Liu (2016)

Even though the Bayes factor of Gönen et al. (2005) is convenient to calculate, it is information inconsistent. Specifically, it converges to a non-zero constant when the t-statistic increases indefinitely (Wang & Liu, 2016, and Gronau et al., 2020). To eliminate this drawback, researchers Wang and Liu (2016) adjusted the model of Gönen et al. (2005) by placing a hyper-prior on σ_δ^2, given by

$$p(\sigma_\delta^2) = \frac{\kappa (\kappa \sigma_\delta^2)^b (1 + \kappa \sigma_\delta^2)^{-a-b-2}}{B(a+1, b+1)}, \quad (3.3.9)$$

which is a Pearson type VI distribution with shape parameters $a > -1$, $b > -1$, and a scale parameter $\kappa > 0$. With this adjustment, plus the setting of the prior mean at $\lambda = 0$, the resulting Bayes factor takes the form

https://doi.org/10.26219/heal.aueb.6623

$$BF_{01} = \left[\int_0^{+\infty} \left[\frac{1+t^2/v}{1+t^2/\{v(1+n_\delta \sigma_\delta^2)\}} \right]^{\frac{v+1}{2}} \times (1+n_\delta \sigma_\delta^2)^{-\frac{1}{2}} p(\sigma_\delta^2) d\sigma_\delta^2 \right]^{-1} . \quad (3.3.10)$$

With the particular choice of $\kappa = n_\delta$ and $b = (v + 1)/2 - a - 5/2$, it greatly simplifies to

$$BF_{01} = \frac{\Gamma((v+1)/2)\Gamma(a+1)}{\Gamma(v/2)\Gamma(a+3/2)} \left(1+\frac{t^2}{v}\right)^{-\frac{(v-2a-2)}{2}} . \quad (3.3.11)$$

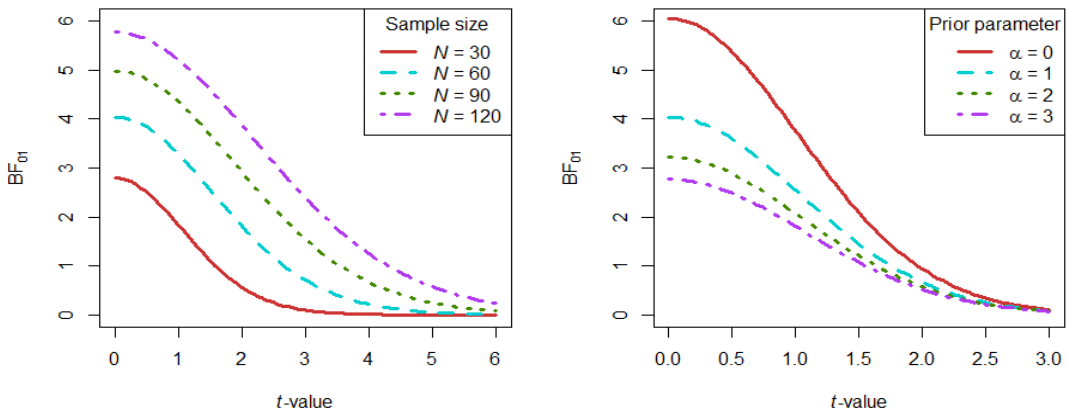

Figure 9: Simplified Wang and Liu (2016) Bayes factor, for different (full) sample sizes N, and prior parameters α. The left graph assumes α = 1. The right graph assumes N = 60.

JZS Bayes Factor of Rouder et al. (2009)

A different way to avoid the restrictive choice of placing a fixed value on the prior variance σ_δ^2, is to assume that it follows an Inverse-X^2 hyper-prior with one degree of freedom, an idea proposed by Zellner and Siow (1980) in the context of regression models. It was later proved (Liang et al., 2008) that integrating out σ_δ^2 using this prior results in a Cauchy-distributed standardized effect size $\delta = d/\sigma$. A Cauchy distribution is equivalent to a t-distribution with one degree of freedom, and its heavy tails make it a reasonable choice in problems where we should expect large effects. The combination of a Cauchy prior on standardized effect size, and of an improper Jeffreys prior $p(\sigma^2) \propto$

https://doi.org/10.26219/heal.aueb.6623

$1/\sigma^2$ on the variance of the data, is referred to as the Jeffreys-Zellner-Siow, or JZS prior, and is a model selection consistent, objective choice for both two-sample and one-sample Bayesian *t*-tests (Rouder et al., 2009, and Morey & Rouder, 2011). Researchers Rouder et al. (2009) maintained the model assumptions of Gönen et al. (2005), with the only adjustment of placing a Cauchy prior with scale parameter r on the standardized effect δ. The resulting JZS Bayes factor for testing H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$, takes the form

$$BF_{01} = \frac{\left(1+\frac{t^2}{v}\right)^{-\frac{(v+1)}{2}}}{\int_0^{+\infty} (1+n_\delta g r^2)^{-\frac{1}{2}} \left(1+\frac{t^2}{(1+n_\delta g r^2) v}\right)^{-\frac{v+1}{2}} (2\pi)^{-\frac{1}{2}} g^{-\frac{3}{2}} e^{-\frac{1}{2g}} dg}, \qquad (3.3.12)$$

where n_δ is the effective sample size, $v = n_1 + n_2 - 2$, and t is the value of the pooled-variance two-sample test statistic. As noted by the authors, and validated in Fig. 10, increasing the prior scale r results in more support for H_0.

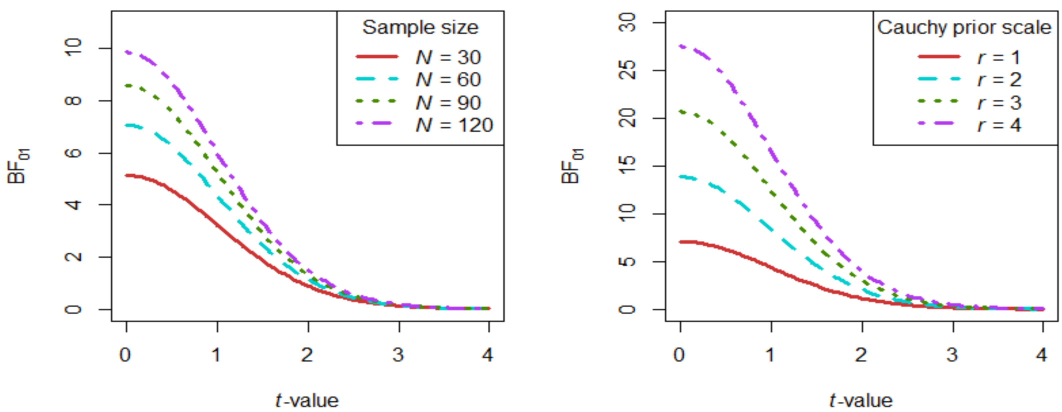

Figure 10: JZS Bayes factor, for different (full) sample sizes N, and Cauchy prior scales r. The left graph assumes r = 1. The right graph assumes N = 60.

Bayes Factor of Gronau et al. (2020)

Working within the framework of Gönen et al. (2005), researchers Gronau et al. (2020) proposed having the standardized effect size $\delta = d/\sigma$ follow a prior with density

$$p(\delta)=\frac{1}{\gamma}T_k\left(\frac{\delta-\mu_\delta}{\gamma}\right), \tag{3.3.13}$$

which is a generalized version of the t-distribution, with location hyperparameter μ_δ, scale hyperparameter γ, and k degrees of freedom. A motivation behind this formulation is that it could allow for easier incorporation of expert opinion into the design. The resulting Bayes factor takes the form

$$BF_{01}=T_v(t)\left[\int_{-\infty}^{+\infty}T_v(t|\sqrt{n_\delta}\delta)\gamma^{-1}T_k\left(\frac{\delta-\mu_\delta}{\gamma}\right)d\delta\right]^{-1}, \tag{3.3.14}$$

where $T_v(x|a)$ denotes the density of a non-central t-distribution, with noncentrality parameter a. It is worthy to note that for $\gamma = \sigma_\delta$ and $k \to +\infty$, this Bayes factor converges to the one of Gönen et al. (2005), while for the specific choice of $k = 1$ and $\mu_\delta = 0$, it is equivalent to the JZS Bayes factor.

3.4 Closing Remarks

This chapter aimed at familiarizing readers with the fundamental principles and motivations behind clinical trials. It also introduced the most common statistical tests conducted during clinical trials, from a frequentist perspective. Furthermore, it explored how broader Bayesian inference is utilized in biomedical research, and presented a list of popular Bayes factors for conducting Bayesian t-tests. Using Bayes factors to compare the means of two groups will be the exclusive focus of the simulation study in Chapter 4.

While formulated for comparisons between point null and composite hypotheses, the Bayes factors presented in this chapter can be extended to accommodate interval null hypotheses, by simply specifying a restricted range over which the prior-likelihood product is to be integrated. In a recent paper, researchers Ravenzwaaij et al. (2019) implemented a Bayes factor with Cauchy priors in phase III clinical trial hypothesis tests. For one-sided tests,

https://doi.org/10.26219/heal.aueb.6623

like those for superiority and non-inferiority, they derived Bayes factors by appropriately truncating the effect size priors to regions consistent with each hypothesis. Although the same can be applied to interval null equivalence tests, it would defeat their very purpose. The motive behind constructing arbitrary intervals around zero is to enable implementation of the TOST approach, as it circumvents the frequentist inability to provide evidence in favor of the null hypothesis. Since evidence for the null H_0: $\theta = 0$ can be quantified directly under the Bayesian paradigm, there is no imperative need for an interval null equivalence test. For those still inclined to implement one, Morey and Rouder (2011) have proposed a hybrid Bayes factor where the null is a controlled mixture of point and interval hypotheses.

One crucial detail is that all of the aforementioned Bayes factors involve a point null hypothesis H_0: $\theta = 0$ with a local alternative prior H_1: $\theta \sim p(\theta)$, meaning that $p(0) > 0$. As mentioned in the previous chapter, this can cause an imbalance in the rate of evidence accumulation. Specifically, when H_0 is true, evidence in its favor accumulates at a sublinear rate, but when H_1 is true, evidence in its favor accumulates at an exponential rate (Johnson & Rossell, 2010). For tests of superiority and non-inferiority, this problem may be evaded by replacing the point null hypothesis with an interval null, combined with a non-overlapping prior. As this is not possible for point null equivalence tests, the use of non-local priors seems to be a promising solution. Fixing the imbalance in the evidence accumulation rate is a small improvement that nevertheless contributes to the inferential advantages of Bayesian methods.

https://doi.org/10.26219/heal.aueb.6623

CHAPTER 4

BAYES FACTORS FOR CLINICAL TRIAL HYPOTHESIS TESTING

4.1 Non-Local Priors for Point Null Tests

Non-local alternative prior densities have the advantage of allowing us to control the rate of shrinkage towards zero when approaching regions consistent with the null hypothesis. This could be utilized, for example, in point null equivalence tests when standardized effects up to a certain magnitude δ are considered unimportant. To avoid rejection of the null for such trivial effects, non-local alternative priors can be modified so as to assign very low density to the region $(-\delta, \delta)$. Although the definition of non-local priors provided by Johnson and Rossell (2010) does not forbid their extension to problems of interval null testing, such settings would automatically eliminate the above advantage, as the obvious solution would be to extend the null region (θ_L, θ_U) to encompass $(-\delta, \delta)$ as well. As such, the non-local alternative priors examined in this chapter will be exclusively based on point null tests. In particular, focus will be given on the moment and inverse moment priors introduced in Johnson and Rossell (2010), and Johnson and Rossell (2012). The following paragraphs refer to testing hypothesis H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$.

Moment Prior (MOM)

Let $p_B(\theta)$ be a 'base' prior density with $2k$ finite integer moments, two bounded derivatives in a neighbourhood of θ_0, and $p_B(\theta_0) > 0$. For any positive integer k, the k-th moment prior density is given by

 https://doi.org/10.26219/heal.aueb.6623

$$p_M(\theta)=\frac{(\theta-\theta_0)^{2k}\,p_B(\theta)}{\int_{\Theta}(\theta-\theta_0)^{2k}\,p_B(\theta)\,d\theta}. \qquad (4.1.1)$$

Base priors $p_B(\theta)$ can be chosen to reflect tail beliefs about parameter θ; if large values seem likely, a heavy-tailed base prior like a *t*-distribution is a reasonable choice. Conversely, a thin-tailed Normal distribution should be preferred when large values are not expected. The former resulting prior would be called a *k*-th *t*-moment prior, and the latter a *k*-th normal moment prior.

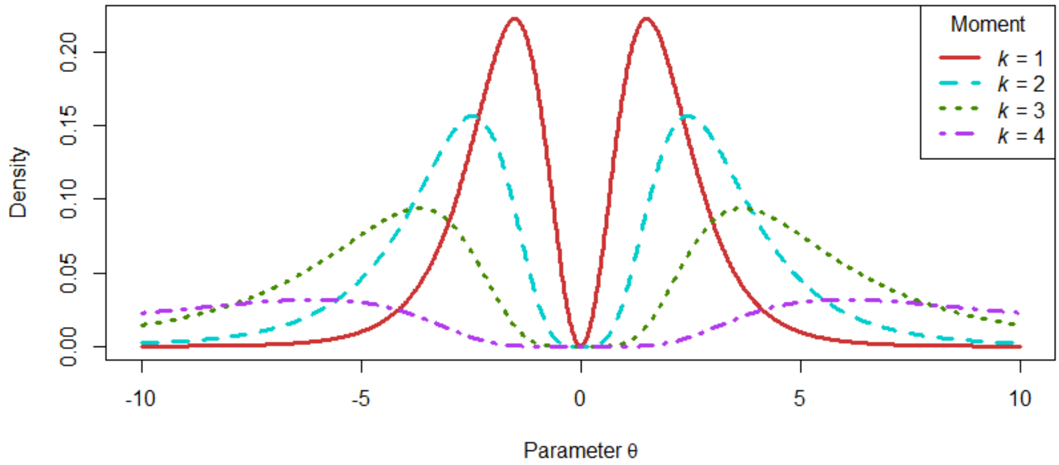

Figure 11: MOM priors with t-student base densities, for different order moments. Assuming degrees of freedom v = 9, and θ_0 = 0.

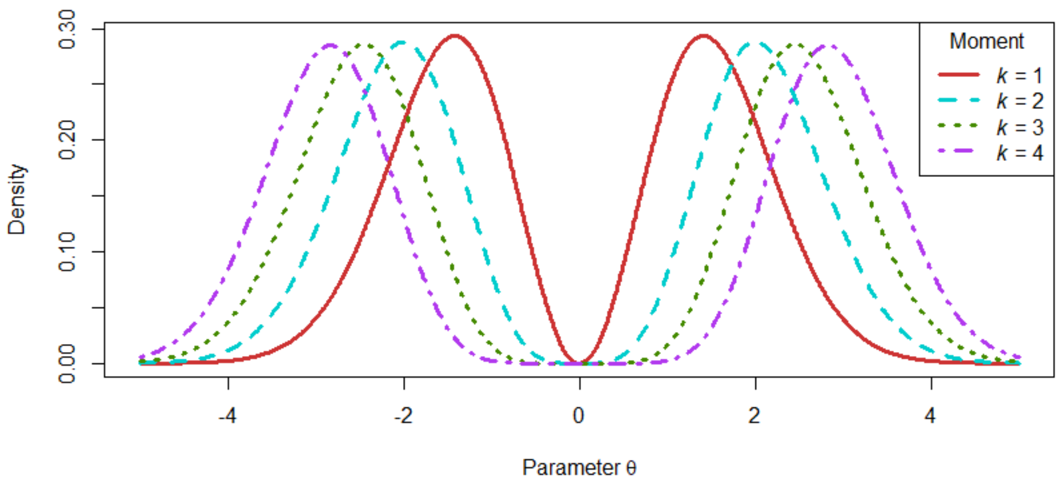

Figure 12: MOM priors with $N(\mu = 0, \sigma^2 = 1)$ base densities, for different order moments. Assuming θ_0 = 0.

https://doi.org/10.26219/heal.aueb.6623

As seen in Fig. 11 and 12, the shapes of the priors are heavily influenced by the moment parameter k, with larger values resulting in wider regions of low density around $\theta_0 = 0$. Although moment priors do not provide exponential accumulation of evidence for the null hypothesis when it is true, they somewhat improve the balance.

Beyond this point, we adopt 1st moment priors, with normal base densities $N(0,\ \tau\sigma^2)$, where τ is a dispersion parameter controlling the shape of the distribution near 0 (Fig. 13), and σ^2 represents the variance in the data. This parameterization is recommended by the authors (Johnson & Rossel, 2010). Under these specifications, and for testing hypothesis H_0: $\theta = 0$, the moment prior is given by

$$p_M(\theta|\tau,\sigma^2) = \frac{\theta^2}{\tau\sigma^2} N(\theta|0,\tau\sigma^2). \tag{4.1.2}$$

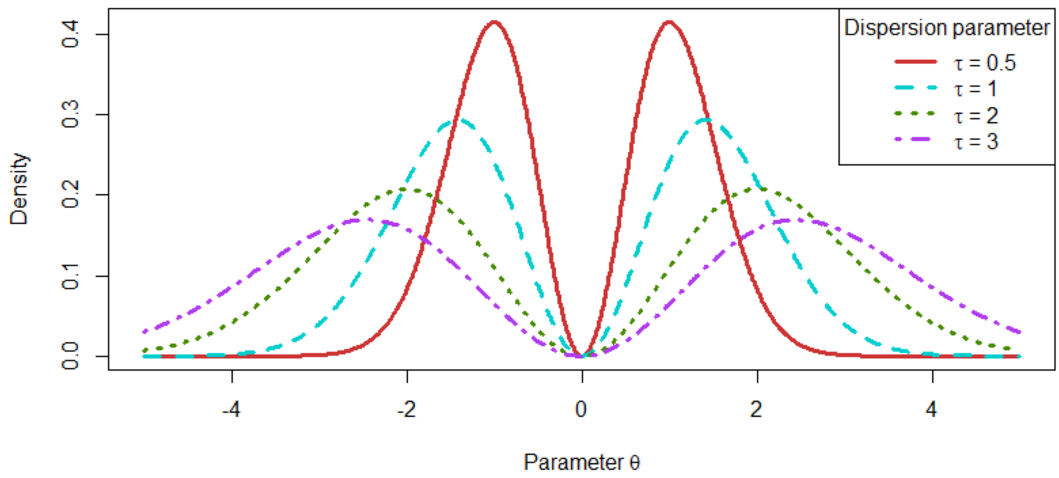

Figure 13: MOM priors, for different dispersion parameter τ values. Assuming base densities $N(\mu = 0,\ \sigma^2 = 1)$, moment order $k = 1$, and $\theta_0 = 0$.

Inverse Moment Prior (iMOM)

For parameters k, v, $\tau > 0$, the inverse moment priors are given by

$$p_I(\theta) = \frac{k\tau^{v/2}}{\Gamma(v/2k)} (\theta-\theta_0)^{-v-1} \exp\left[-\frac{\tau^k}{(\theta-\theta_0)^{2k}}\right]. \tag{4.1.3}$$

https://doi.org/10.26219/heal.aueb.6623

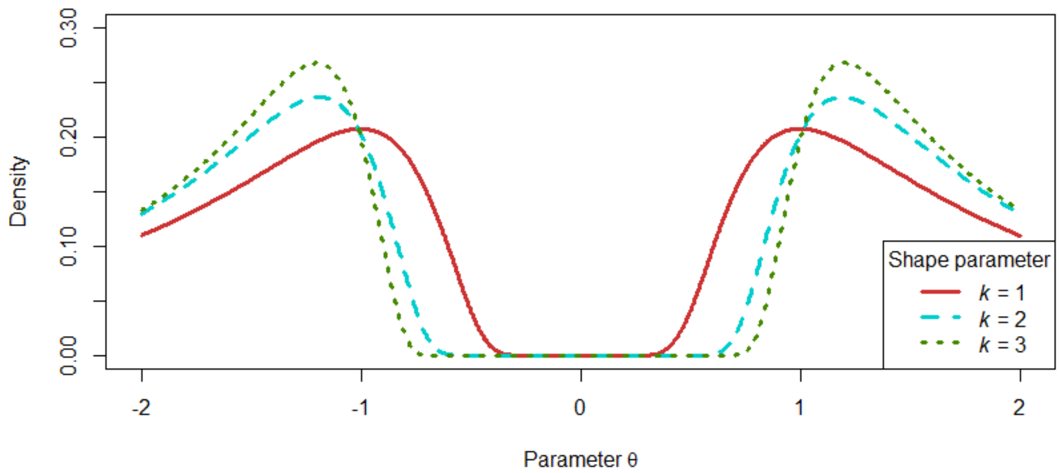

Figure 14: iMOM priors for different parameters k. Assuming ν = 1, τ = 1, and θ_0 = 0.

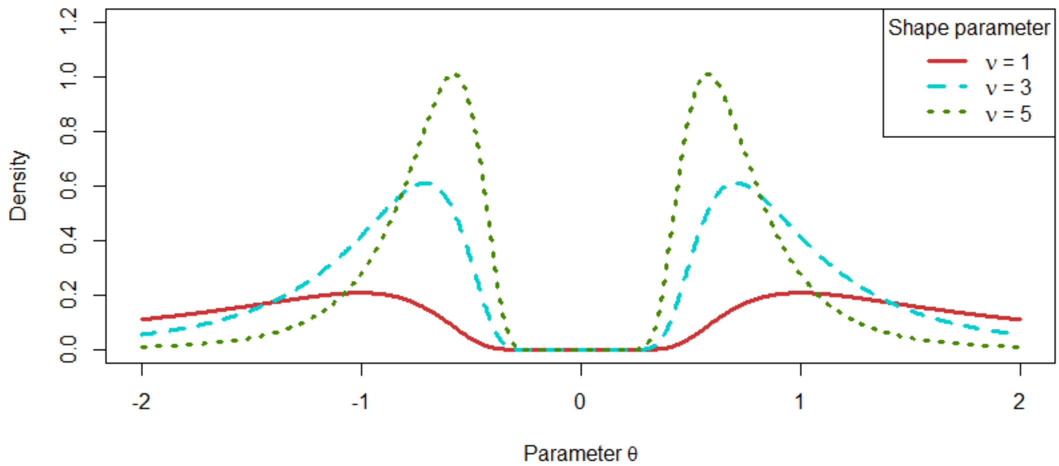

Figure 15: iMOM priors for different parameters ν. Assuming k = 1, τ = 1, and θ_0 = 0.

When approaching $\theta_0 = 0$, inverse moment prior densities behave like Inverse-Gamma distributions when approaching 0 (Fig. 14, 15). Their parameters relate to their shape, and determine properties such as the convergence rate of the resulting Bayes factors. Inverse moment priors provide exponential accumulation of evidence for the null hypothesis when it is true, fully restoring the balance. Adopting the default specifications and

https://doi.org/10.26219/heal.aueb.6623

parameterization of the authors (Johnson & Rossell, 2010), we set $k = \nu = 1$, so that for testing H_0: $\theta = 0$, inverse moment priors (Fig. 16) take the form

$$p_I(\theta|\tau,\sigma^2) = \frac{\sqrt{\tau\sigma^2}}{\sqrt{\pi}\,\theta^2}\exp\left(-\frac{\tau\sigma^2}{\theta^2}\right). \tag{4.1.4}$$

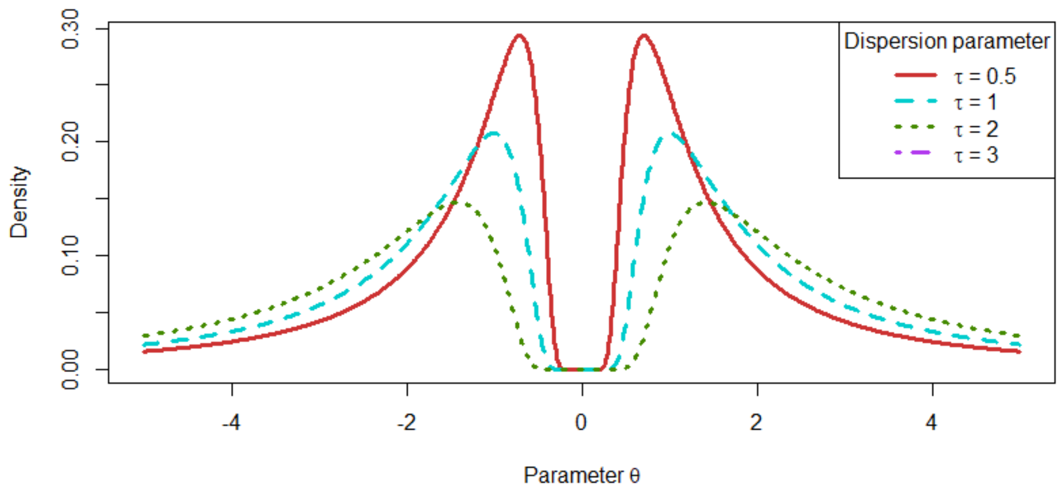

Figure 16: iMOM priors for different dispersion parameters τ. Assuming σ^2 = 1.

4.1.1 Applications to Bayesian Variable Selection

The moment and the inverse moment priors of Johnson and Rossell (2010) can be generalized for testing vector parameters. This can be useful in regression variable selection problems. Assume a response $Y \sim N(X\theta,\ \sigma^2 I_n)$, where $X_{n\times p}$ is a design matrix, and θ is a p-dimensional covariate coefficient vector. With base prior $p_B(\theta) = N_p(\theta \mid 0,\ \tau\sigma^2\Sigma)$, where $\Sigma_{p\times p}$ is a positive definite matrix, the multivariate moment prior for testing H_0: $\theta = 0$ under default specifications, is given by

$$p_M(\theta|\tau,\sigma^2,\Sigma) = \frac{\theta^T\Sigma^{-1}\theta}{p\tau\sigma^2} N_p(\theta|0,\tau\sigma^2\Sigma). \tag{4.1.5}$$

 https://doi.org/10.26219/heal.aueb.6623

Likewise, the multivariate inverse moment prior is given by

$$p_I(\theta|\tau,\sigma^2,\Sigma)=\left|\frac{\Sigma^{-1}}{\tau\sigma^2}\right|^{1/2}\frac{\Gamma(p/2)}{\pi^{(p+1)/2}}\left(\frac{\theta^T\Sigma^{-1}\theta}{\tau\sigma^2}\right)^{-\frac{1+p}{2}}\exp\left(-\frac{\tau\sigma^2}{\theta^T\Sigma^{-1}\theta}\right). \tag{4.1.6}$$

Under these formulations, the multivariate moment and inverse moment prior densities may be used for variable selection. One of their noted features is that they become zero only when *all* components of the parameter vector θ are zero. As the authors explained in a subsequent paper (Johnson & Rossell, 2012), such a characteristic imposes little penalty to models that contain covariates with unimportant effects (i.e.: with coefficients very close to zero). This motivated the development of adjusted product moment and product inverse moment priors, that shrink to zero density when *any* of the θ parameter vector's components approach zero. Such a feature assigns low likelihood to an entire model when it contains covariates with trivial effects, in favor of models that are more parsimonious. Under default specifications for testing H_0: $\theta = 0$, the product moment (pMOM) prior density is given by

$$p_M(\theta|\tau,\sigma^2)=\prod_{i=1}^{p}\frac{\theta_i^2}{\tau\sigma^2}N(\theta_i|0,\tau\sigma^2), \tag{4.1.7}$$

and the product inverse moment (piMOM) prior density is given by

$$p_I(\theta|\tau,\sigma^2)=\prod_{i=1}^{p}\frac{\sqrt{\tau\sigma^2}}{\sqrt{\pi}\,\theta_i^2}\exp\left(-\frac{\tau\sigma^2}{\theta_i^2}\right). \tag{4.1.8}$$

Although their usefulness may not be immediately obvious, variable selection problems can have a direct application to clinical trial hypothesis tests. In particular, the Bayesian t-test for equivalence can be expressed as a linear regression problem where the response variable Y is standardized, and the standardized effect d is the coefficient of a dummy variable indicating whether a subject belongs to the intervention group. This will be explored in further detail in the next section.

https://doi.org/10.26219/heal.aueb.6623

4.2 Description of the Problem

Our goal is to examine how different types of Bayes factors behave when employed for clinical trial-related hypothesis tests in practice. A way of performing such an assessment is through a simulation study. Since the data for every test are simulated, we have full knowledge of the mechanisms that generated them, and can thus easily evaluate each Bayes factor's ability of detecting the true hypothesis. Although all three tests of superiority, equivalence, and non-inferiority are of interest, tests of equivalence using local and non-local priors receive increased attention in our experiment. While theoretically possible to do so, non-local priors will not be employed in tests of superiority and non-inferiority, for reasons that will be discussed further below.

For the simulation study, we will generate continuous observations for independent control and treatment groups of equal size. Specifically, the observations under both groups are generated from a Normal distribution of variance $\sigma^2 = 1$, with a mean $\mu_C = 0$ in the control group, and $\mu_T = d$, denoting the standardized effect size, in the treatment group. For the alternative hypothesis of $d \neq 0$, we take into consideration the categorization of Cohen (1988), and therefore test for effect sizes of magnitude 0.2 (small), 0.5 (medium), and 0.8 (large).

Data generation and analysis is carried out using the R statistical software (R Core Team, 2020, and RStudio Team, 2020), accompanied by a variety of libraries (Linde and van Ravenzwaaij, 2019, Morey and Rouder, 2018, Rossell et al., 2020, Bååth, 2018, Wickham, 2016, and Wickham, 2020).

4.2.1 Equivalence Testing

When testing for equivalence, we have the choice of expressing the null hypothesis as either a point (H_0: $\delta = 0$), or an interval (H_0: $\delta_L < \delta < \delta_U$). As mentioned previously, choosing an interval is a necessity under the frequentist

https://doi.org/10.26219/heal.aueb.6623

paradigm, due to the inability of providing confirmatory evidence for the (point) null hypothesis of equivalence. As this is no problem for Bayes factors, it makes sense to formulate the null as a point (H_0: $\delta = 0$), thereby avoiding the largely subjective specification of an equivalence margin around zero.

With the null hypothesis as a point, our goal is to compare the performance of Bayes factors with local versus non-local priors on the standardized effect difference. This will be achieved by performing tests under two scenarios:

Scenario 1: Local Priors

Under this scenario, we will test for equivalence using the JZS Bayes factor, which places a local alternative Cauchy prior on the effect size d. This can be implemented in the R libraries 'baymedr' (Linde and van Ravenzwaaij, 2019), and 'BayesFactor' (Morey and Rouder, 2018), both of which place a scale of $r = 1/\sqrt{2}$ on the Cauchy prior as the default choice. A Cauchy distribution with scale r has half its mass within the interval ($-r$, r), so apart from the default choice of $r = 1/\sqrt{2}$, we will use a 'narrow' prior with $r = 0.3$, and a 'wide' one with $r = 1.1$, as a robustness check (Fig. 31, Appendix). The three JZS Bayes factors resulting from these priors will be accompanied by the Schwarz criterion Bayes factor approximation (BIC), and the minimum Bayes factor (MBF). The latter provides an objective lower bound for the purposes of comparison.

Scenario 2: Non-local Priors

After testing for equivalence using local priors, we will repeat the tests using the non-local moment (MOM) and inverse moment (iMOM) alternative priors of Johnson and Rossell (2010, 2012). This can be implemented in the R library 'mombf' (Rossell et al., 2020) in the context of Bayesian variable selection for linear regression models.

Consider standardized observations Y_i, and a dummy variable X_i taking on a value of 0 when subject i belongs to the control group, and a value of 1

https://doi.org/10.26219/heal.aueb.6623

when it belongs to the treatment group. Assuming equal variances between control and treatment group, an equivalence test on the standardized effect difference d is like performing variable selection on the linear model

$$Y_i = b_0 + dX_i + \varepsilon_i, \quad \text{where} \quad \varepsilon_i \sim N(0, \sigma^2). \tag{4.2.1}$$

The 'mombf' library places a multivariate pMOM or piMOM prior on the covariate coefficient vector of the model, and performs variable selection by computing the Bayes factor between a fitted full model, and a submodel where one or more covariates have been excluded. In our case, the full covariate coefficient vector is $[b_0, d]^T$, and the component corresponding to the excluded covariate is d. Unlike the JZS Bayes factor, which places an improper Jeffreys prior $p(\sigma^2) \propto 1/\sigma^2$ on the data variance σ^2, the 'mombf' library places a "minimally informative" Inverse-Gamma prior. The density of an Inverse-Gamma distribution is proportional to

$$f(\sigma^2 | a, b) \propto (\sigma^2)^{-a-1} \exp\left(-\frac{b}{\sigma^2}\right) \to \frac{1}{\sigma^2}, \quad \text{for} \quad (a, b) \to (0,0). \tag{4.2.2}$$

It is, therefore, not surprising that Inverse-Gamma distributions with small parameter values a and b are popular choices for priors placed on variances.

Finally, we need to specify the dispersion parameter τ of the pMOM and piMOM priors. The recommended choices of the authors are $\tau = 0.348$ for the pMOM, and $\tau = 0.133$ for the piMOM. Such choices are justified on the grounds that they lead to priors with only 1% of their probability mass in the interval (–0.2, 0.2), thus tuned towards detecting effects of magnitude greater than 0.2 (Fig. 32, Appendix). We adopt this default specification, but also implement our own values, with $\tau = 0.021$ for the pMOM, and $\tau = 0.015$ for the piMOM prior. These choices, mostly tuned visually (Fig. 33, Appendix), place 15% of the probability mass in the interval (–0.13, 0.13) to the former, and 5% of the probability mass in the interval (–0.09, 0.09) to the latter distribution. Each analysis is accompanied by the default JZS Bayes factor

https://doi.org/10.26219/heal.aueb.6623

(with Cauchy scale $r = 1/\sqrt{2}$), and the Schwarz criterion approximation (BIC). Under both scenarios, we test for detecting effects of size $d = 0$, 0.2, 0.5, and 0.8.

4.2.2 Non-inferiority and Superiority Testing

For non-inferiority testing, we face a similar choice of formulating the null hypothesis as either a point (H_0: $\delta = -m$), or an interval (H_0: $\delta < -m$). In the former case, an alternative prior p_1 with support on $(-m, +\infty)$ that doesn't smoothly shrink to zero when approaching $-m$[14] would fulfil the definition of a local prior, and could potentially cause the usual problems of evidence accumulation rate imbalance (Johnson and Rossell, 2010). While this could be fixed by employing a non-local alternative prior on $(-m, +\infty)$, a simpler solution would be to express the null as an interval, and use truncated priors on the regions defined by each hypothesis. Under this formulation, the definition of local alternative prior does not hold, and there is no apparent reason to expect an imbalance in the rate of evidence accumulation. For this test, we will employ the JZS Bayes factor with narrow, standard, and wide Cauchy priors (Fig. 34, Appendix), as defined in the equivalence test of Scenario 1. We will test for effects of $d = 0$, ±0.2, ±0.5, and ±0.8, and the non-inferiority margin will be $-m = -0.1$.

The setup for the superiority test will follow the same logic, employing JZS Bayes factors with Cauchy priors truncated at $(-\infty, 0)$ and $[0, +\infty)$ for the null and alternative hypotheses respectively (Fig. 35, Appendix). The tested effects will be $d = \pm0.2$, ±0.5, and ±0.8. Both non-inferiority and superiority tests can be implemented using the R libraries 'baymedr' and 'BayesFactor'.

14 Meaning that there exists an $\varepsilon > 0$, such that for any $\delta > 0$ there exists a point $x > -m$ such that $|x+m|<\delta$ and $p_1(x) > \varepsilon$.

 https://doi.org/10.26219/heal.aueb.6623

4.2.3 Bayesian Power Analysis

As previously mentioned, the assessment of operational characteristics is a crucial aspect of clinical trial hypothesis tests, and a major reason for the dominance of frequentist methods in the field. However, operational characteristics can be assessed under the Bayesian framework as well. For each test of the simulation study, the type I error rate α and power $1 - \beta$ will be determined through repeated Monte Carlo simulations, in line with the suggestion of Schönbrodt and Wagenmakers (2018). What is required is a decision rule for accepting or rejecting the null hypothesis, so we choose to accept the null for $BF_{01} \geq 1$, and reject it for $BF_{01} < 1$. While this is not a threshold that requires decisive evidence, the choice is regarded as valid by the authors. For every type of test, operational characteristics will be determined through the following process:

For any given (total) sample size N, and true effect d, we will generate observations for the control and treatment groups, and perform a test using the Bayes factor of interest. This will be repeated $B = 10{,}000$ times, resulting in a large sample of Bayes factor values. Based on that sample, we will calculate each Bayes factor type's probability of false rejection of the null (α), or correct acceptance of the alternative ($1 - \beta$). The sample sizes that give the desired operational characteristics $\alpha = 0.05$, 0.025 and $1 - \beta = 0.9$, 0.95 will be estimated through trial and error.

An advantage of this approach, according to the authors, is that it can account for expected effect size uncertainty, unlike frequentist power calculations that condition on a fixed effect size. For instance, we may suspect that the effect size follows some distribution, e.g.: $d \sim N(\mu, \sigma^2)$, and we want this variability to be reflected in our power calculations. The solution is to generate a different effect size from this distribution, for each of the $B = 10{,}000$ Monte Carlo simulations. In our tests of equivalence, we use this approach as well, under the assumption that $d \sim N(0, 1.5^2)$.

https://doi.org/10.26219/heal.aueb.6623

4.3 Results and Conclusions

Presenting the values of different Bayes factors at the end of every test would not provide much insight. Instead, our analysis is based on the following steps:

1. Simulate B = 1,000 datasets of (total) sample size N.
2. Calculate all the different Bayes factor types for each dataset.
3. Derive a measure of centrality for every Bayes factor type.
4. Repeat for incrementally larger sample sizes N, up to a reasonable limit.

The measure of centrality that we use is the median, due to its robustness. Most graphs are presented in the $2log_e(BF_{01})$ scale, as recommended by Kass and Raftery (1995).

4.3.1 Equivalence Testing

In this section we will test H_0: $d = 0$ vs H_1: $d \neq 0$. Let's first consider local prior Bayes factors, when the real standardized effect difference is $d = 0$.

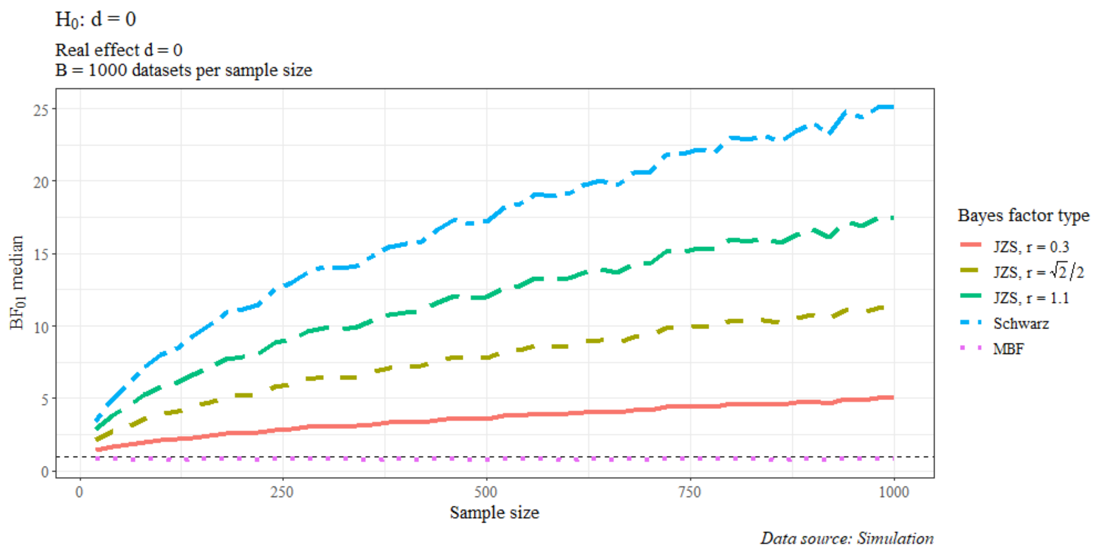

Figure 17: Equivalence testing with local priors.

 https://doi.org/10.26219/heal.aueb.6623

As observed in Fig. 17, the Schwarz approximation provides the most optimistic evidence in support of the null hypothesis, followed by the JZS Bayes factors, in order of decreasing Cauchy prior scale r. For most Bayes factor types, evidence for the null becomes strong at sample sizes of about $N = 1{,}000$. In contrast, the JZS Bayes factor with scale $r = 0.3$ fails to provide strong evidence, as it increases very slowly. The overall picture suggests that Bayes factors with Cauchy (local) priors are hesitant to support the null hypothesis when it is true, something in line with the conclusions of Johnson and Rossell (2010).

Let us compare this with the behavior of non-local prior Bayes factors with default specifications, this time with both graphs on the $2log_e(BF_{01})$ scale. Comparing Fig. 18, and 19, we see that the Bayes factors with default non-local priors show extreme support for the null hypothesis from small sample sizes, with the MOM Bayes factor initially in the lead, but eventually overtaken by the iMOM Bayes factor. This could follow from the fact that the default iMOM prior decreases a lot faster than the default MOM prior as it approaches the null value of $d = 0$ (Fig. 32, Appendix).

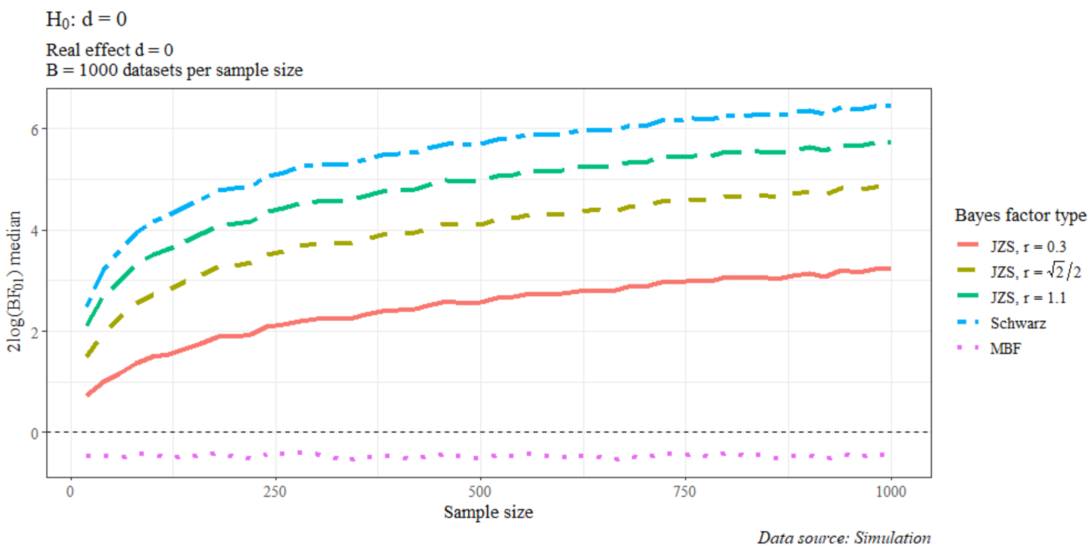

Figure 18: Equivalence testing with local priors.

https://doi.org/10.26219/heal.aueb.6623

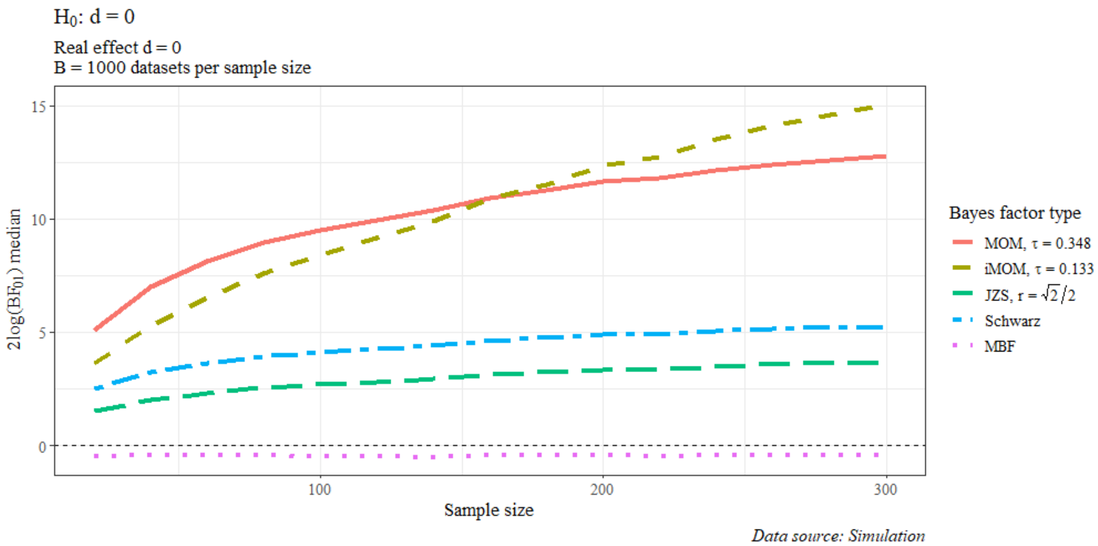

Figure 19: Equivalence testing with default non-local priors.

Our non-local prior Bayes factors with custom specifications (Fig. 20) eventually provide strong support for the null at large sample sizes. However, when it comes to smaller sample sizes, they are a lot more hesitant to support the null compared to the ones with default specifications. Likewise, the iMOM prior Bayes factor eventually overtakes the MOM prior Bayes factor, at sample sizes of around $N = 400$.

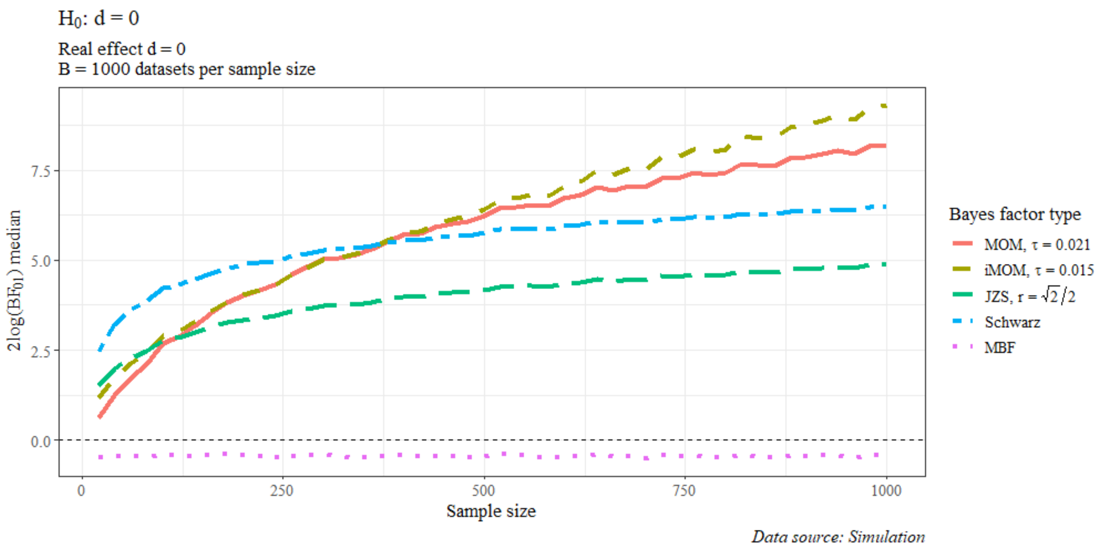

Figure 20: Equivalence testing with custom non-local priors.

 https://doi.org/10.26219/heal.aueb.6623

Let us now demonstrate the results for the case where the real standardized effect size is $d = 0.2$.

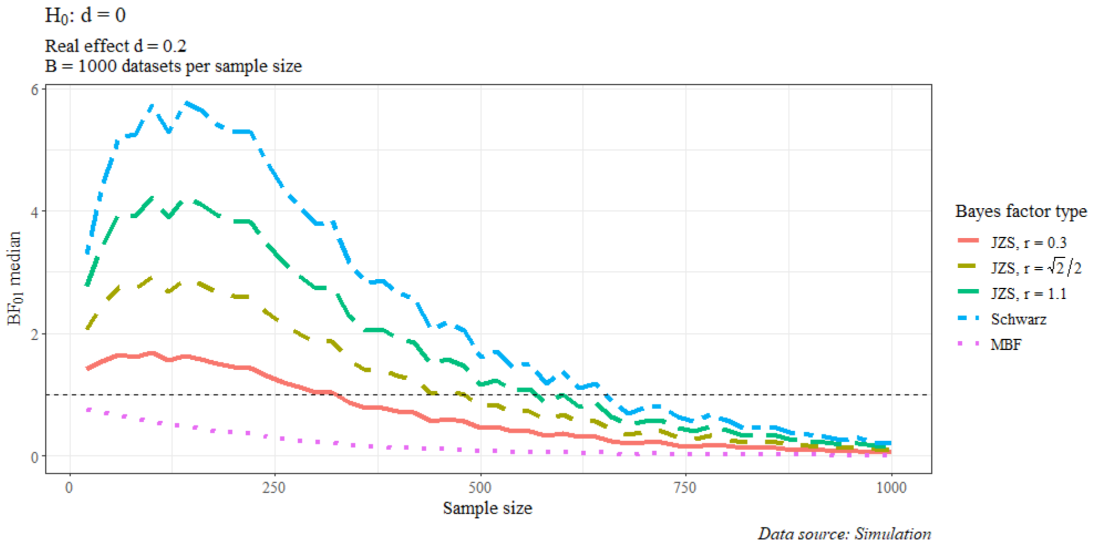

Figure 21: Equivalence testing with local priors.

In Fig. 21, all Bayes factors incorrectly point towards H_0, but they eventually start to favor H_1 at larger samples. The evidence against the null becomes strong at about $N = 1{,}000$, which is easily visible in the $2log_e(BF_{01})$ scale (Fig. 22). The behavior of Bayes factors becomes almost indistinguishable as the real effect increases to $d = 0.5$ and $d = 0.8$ (Fig. 36 – 39, Appendix), and they require much smaller samples to reach strong levels of evidence.

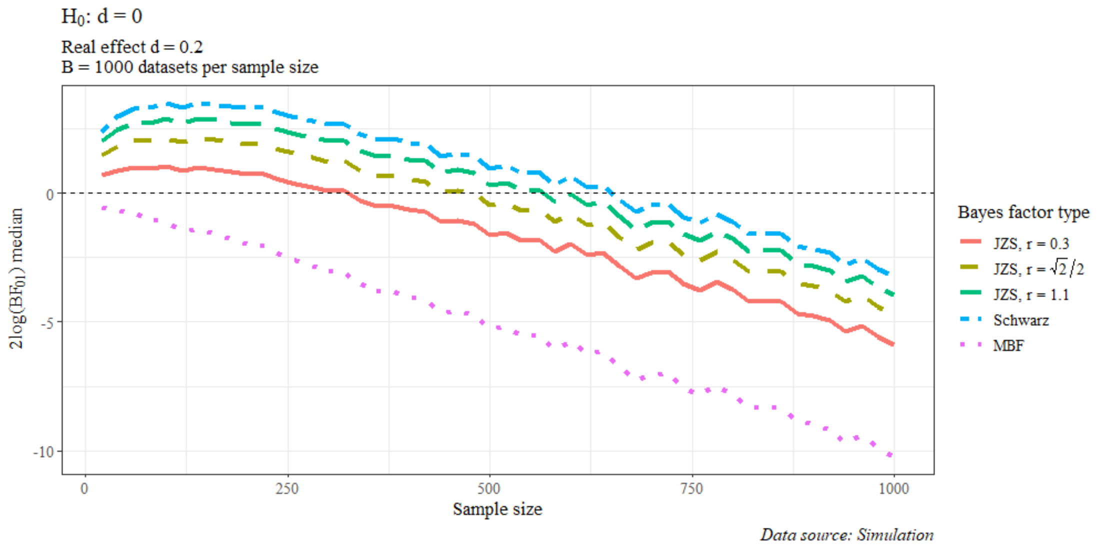

Figure 22: Equivalence testing with local priors.

 https://doi.org/10.26219/heal.aueb.6623

In the problem where $d = 0.5$, the custom non-local Bayes factors seem to be slightly lower than the standard JZS Bayes factor (Fig. 23). This minor improvement at detecting the true alternative hypothesis also manifests when the true effect is $d = 0.2$, and 0.8 (Fig. 40, 41, Appendix).

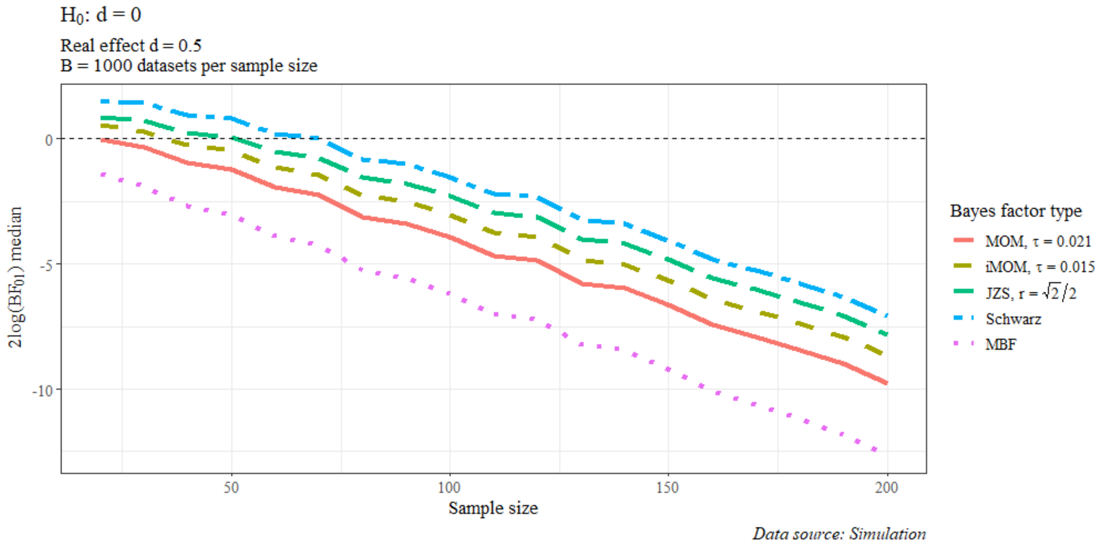

Figure 23: Equivalence testing with custom non-local priors.

Conversely, non-local Bayes factors with default specifications seem predisposed to falsely supporting the null hypothesis, in light of anecdotal data (Fig. 24). It may be easier to understand the magnitude of this behavior for the same test on the linear scale (Fig. 42, Appendix).

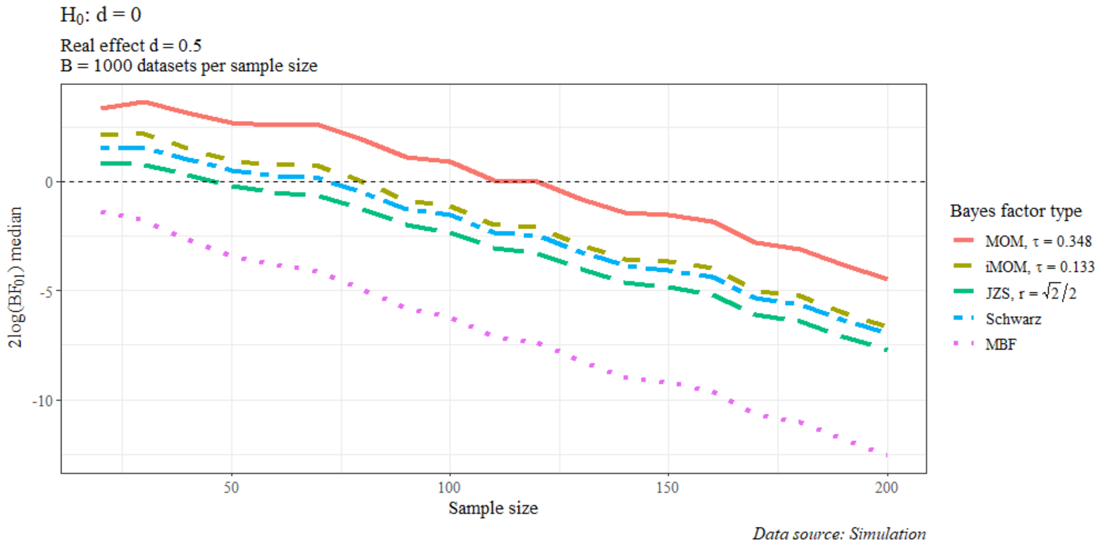

Figure 24: Equivalence testing with default non-local priors.

 https://doi.org/10.26219/heal.aueb.6623

This phenomenon is also present when the true effect is large (d = 0.8) (Fig. 43, Appendix), or small (d = 0.2) (Fig. 25).

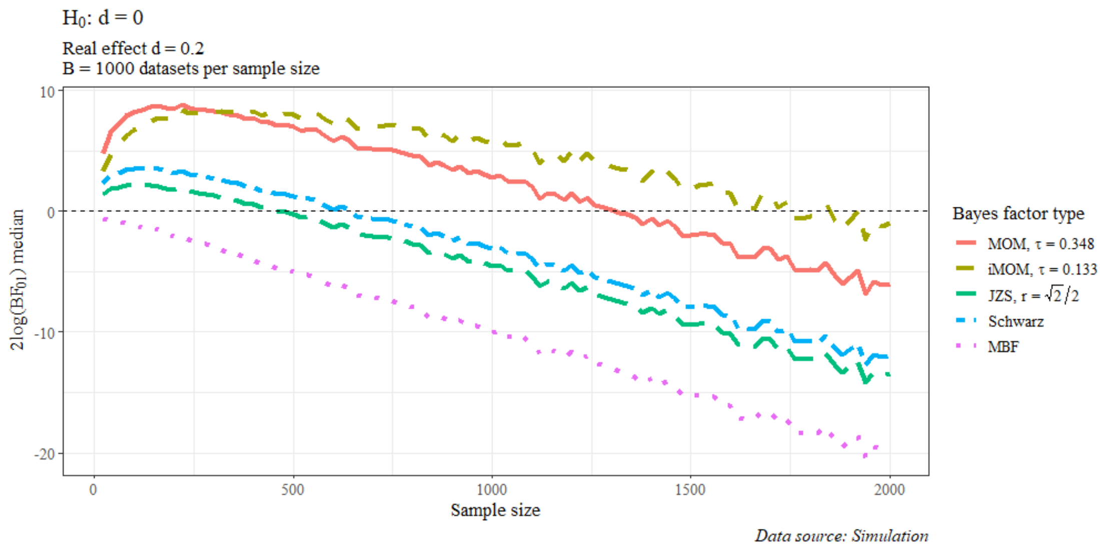

Figure 25: Equivalence testing with default non-local priors.

4.3.2 Non-Inferiority Testing

In this section we will test H_0: $d < -0.1$ vs H_1: $d \geq -0.1$. To analyze the behavior of JZS Bayes factors with truncated Cauchy priors in non-inferiority tests, let's examine the case in which the real standardized effect is $d = -0.5$.

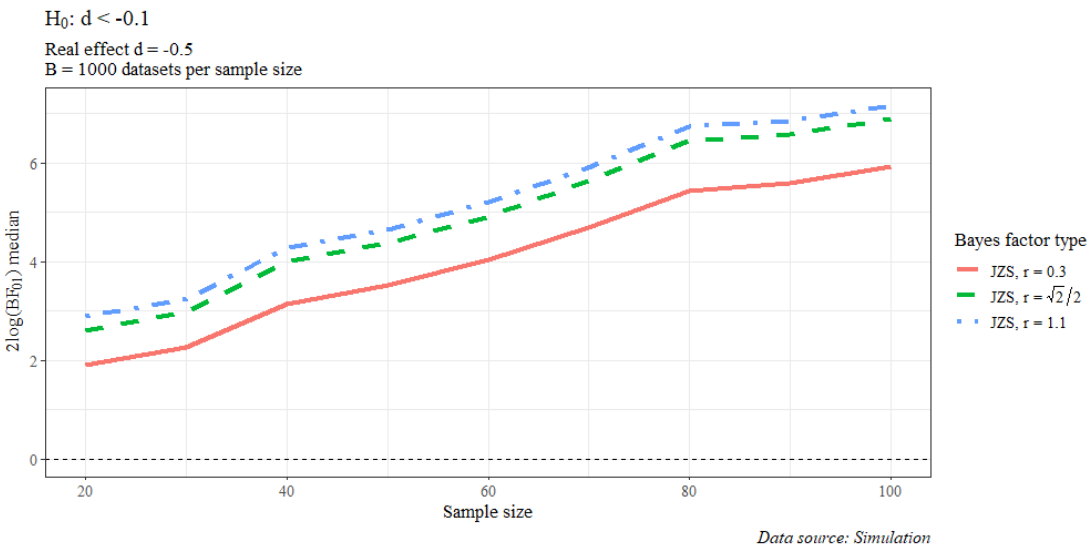

Figure 26: Non-inferiority testing.

https://doi.org/10.26219/heal.aueb.6623

In Fig. 26 the evidence seems to become decisive at a sample size of around N = 100, and larger Cauchy prior scales r result in portraying the evidence more favorably. As expected, effects of larger magnitudes result in Bayes factors which strongly support the true hypotheses at smaller samples (Fig. 27, and Fig. 44 – 48, Appendix).

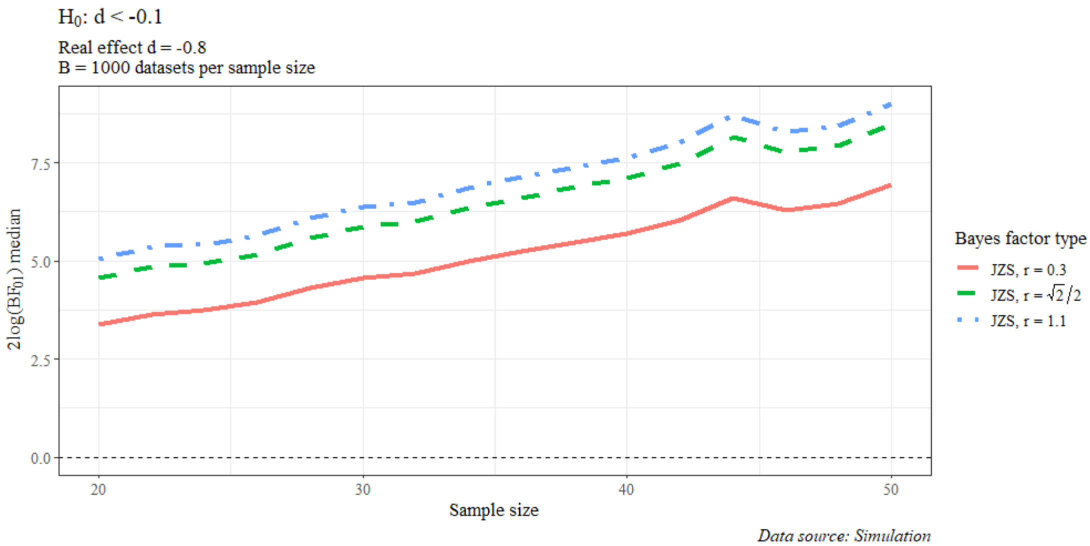

Figure 27: Non-inferiority testing.

4.3.3 Superiority Testing

In this section we will test H_0: d < 0 vs H_1: $d \geq 0$. Like non-inferiority tests, superiority tests with JZS Bayes factors are straightforward. As a typical example, we present the tests where the real effects are d = –0.2 and d = 0.8 (Fig. 28, 29), while the rest are provided in the Appendix (Fig. 49 – 52, Appendix).

As expected, effects of greater magnitude require samples of smaller size in order for the Bayes factors to reach strong levels of evidence. Another interesting behavior, also present in non-inferiority testing, is that larger Cauchy prior scales r lead to Bayes factors that are more favorable towards either hypothesis, conditional on all Bayes factors being on the same side of the horizontal line y = 0. This can be clearly seen in Fig. 28 and 29, where the

https://doi.org/10.26219/heal.aueb.6623

relative positions of the three Bayes factors are 'mirrored' across the line defined by $y = 0$.

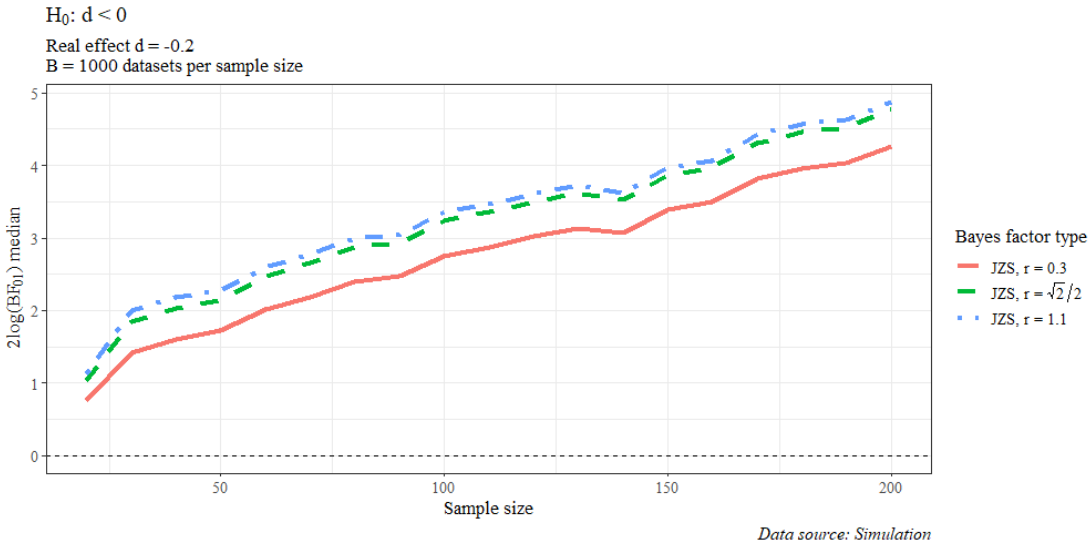

Figure 28: Superiority testing.

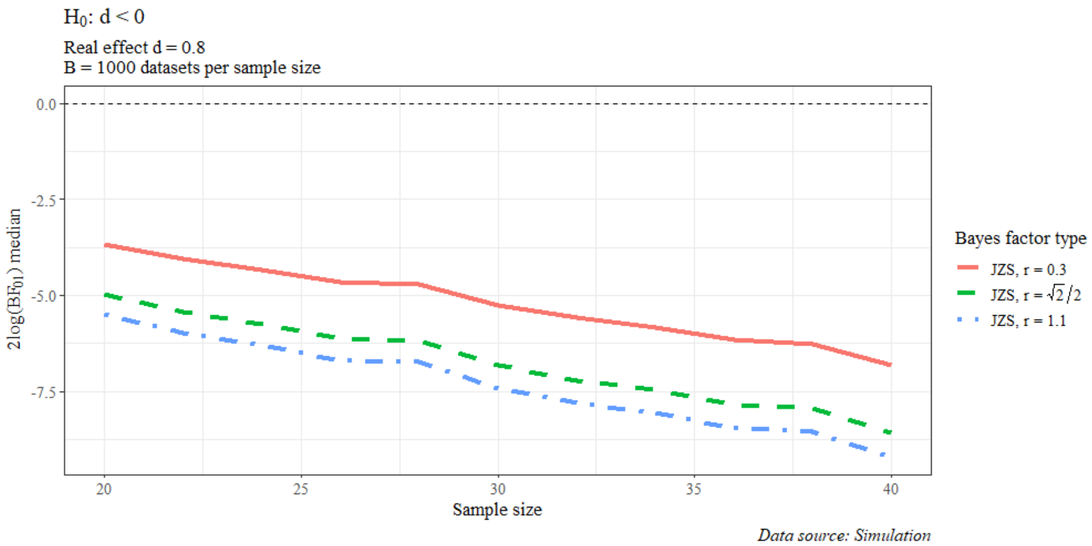

Figure 29: Superiority testing.

4.3.4 Power Analysis

As previously stated, the operational characteristics of every Bayes factor have been examined exclusively through simulations. By using the value of

https://doi.org/10.26219/heal.aueb.6623

one (1) as a decision threshold[15] our analysis leads to taking decisions even when the evidence is anecdotal (e.g.: accepting H_0 when BF_{01} = 1.3). We could have chosen a stronger threshold for decision-making, such as accepting H_0 when $BF_{01} > 10$, and rejecting it when $BF_{01} < 0.1$, but this approach would have created an inconclusive non-decision region between 0.1 and 10. As the option of no decision is not available in traditional power analysis, we avoided this approach, in order to keep our Bayesian results in line with the prevailing philosophy.

For equivalence testing using local priors, results for every Bayes factor (including the Schwarz approximation) can be seen in Tables 5 – 8.

JZS Bayes factor		**$r = 0.3$**	**$r = 1/\sqrt{2}$**	**$r = 1.1$**
Real effect size	**Operational Characteristic**	**Sample Size**	**Sample Size**	**Sample Size**
$d = 0.0$ (H_0)	$\alpha < 0.05$	$N = 900$	$N = 190$	$N = 100$
	$\alpha < 0.025$	$N = 3{,}500$	$N = 720$	$N = 340$
$d = 0.2$ (H_1)	$1 - \beta > 0.90$	$N = 1{,}100$	$N = 1{,}350$	$N = 1{,}500$
	$1 - \beta > 0.95$	$N = 1{,}400$	$N = 1{,}650$	$N = 1{,}900$
$d = 0.5$ (H_1)	$1 - \beta > 0.90$	$N = 140$	$N = 176$	$N = 200$
	$1 - \beta > 0.95$	$N = 180$	$N = 220$	$N = 250$
$d = 0.8$ (H_1)	$1 - \beta > 0.90$	$N = 50$	$N = 60$	$N = 70$
	$1 - \beta > 0.95$	$N = 70$	$N = 80$	$N = 90$
$d \sim N(0, 1.5^2)$ (H_1)	$1 - \beta > 0.90$	$N = 420$	$N = 500$	$N = 750$
	$1 - \beta > 0.95$	$N = 2{,}200$	$N = 2{,}240$	$N = 3{,}400$

Table 5: Equivalence testing power analysis for the JZS Bayes factor with Cauchy prior scales r = 0.3, 1/√2, and 1.1.

15 Accepting H_0 when $BF_{01} \geq 1$, and rejecting it when $BF_{01} < 1$.

https://doi.org/10.26219/heal.aueb.6623

Schwarz criterion Bayes factor approximation		
Effect size	**Operational Characteristic**	**Sample Size**
d = 0.0 (H_0)	α < 0.05	N = 60
	α < 0.025	N = 190
d = 0.2 (H_1)	1 – β > 0.90	N = 1,650
	1 – β > 0.95	N = 1,950
d = 0.5 (H_1)	1 – β > 0.90	N = 210
	1 – β > 0.95	N = 260
d = 0.8 (H_1)	1 – β > 0.90	N = 70
	1 – β > 0.95	N = 100
$d \sim N(0, 1.5^2)$ (H_1)	1 – β > 0.90	N = 700
	1 – β > 0.95	N = 3,800

Table 6: Equivalence testing power analysis for the Schwarz Bayes factor approximation.

MOM Bayes factor		**τ = 0.021**	**τ = 0.348**
Effect size	**Operational Characteristic**	**Sample Size**	**Sample Size**
d = 0.0 (H_0)	α < 0.05	N = 350	N = 26
	α < 0.025	N = 690	N = 50
d = 0.2 (H_1)	1 – β > 0.90	N = 1,500	N = 2,700
	1 – β > 0.95	N = 1,900	N = 3,200
d = 0.5 (H_1)	1 – β > 0.90	N = 140	N = 310
	1 – β > 0.95	N = 190	N = 370
d = 0.8 (H_1)	1 – β > 0.90	N = 50	N = 100
	1 – β > 0.95	N = 60	N = 130
$d \sim N(0, 1.5^2)$ (H_1)	1 – β > 0.90	N = 650	N = 1,600
	1 – β > 0.95	N = 4,500	N = 8,000

Table 7: Equivalence testing power analysis for the MOM Bayes factor with custom (τ = 0.021) and default (τ = 0.348) specifications.

https://doi.org/10.26219/heal.aueb.6623

iMOM Bayes factor		$\tau = 0.015$	$\tau = 0.133$
Effect size	Operational Characteristic	Sample Size	Sample Size
$d = 0.0$ (H_0)	$\alpha < 0.05$	$N = 350$	$N = 50$
	$\alpha < 0.025$	$N = 750$	$N = 100$
$d = 0.2$ (H_1)	$1 - \beta > 0.90$	$N = 1{,}500$	**[16]
	$1 - \beta > 0.95$	$N = 1{,}900$	***[17]
$d = 0.5$ (H_1)	$1 - \beta > 0.90$	$N = 150$	$N = 280$
	$1 - \beta > 0.95$	$N = 200$	$N = 340$
$d = 0.8$ (H_1)	$1 - \beta > 0.90$	$N = 60$	$N = 80$
	$1 - \beta > 0.95$	$N = 70$	$N = 100$
$d \sim N(0, 1.5^2)$ (H_1)	$1 - \beta > 0.90$	$N = 650$	$N = 2{,}500$
	$1 - \beta > 0.95$	*[18]	****[19]

Table 8: Equivalence testing power analysis for the iMOM Bayes factor with custom ($\tau = 0.015$) and default ($\tau = 0.133$) specifications.

A notable finding is that the custom specification iMOM Bayes factor fails to reach power greater than 0.92, even for very large sample sizes (e.g.: $N >$ 10,000) when d is assumed to be a random variable, generated from a $N(0, 1.5^2)$ distribution (Table 8). A potential explanation for this phenomenon, is that small effects (generated from the $N(0, 1.5^2)$ distribution) are assigned very low probability density under the custom iMOM prior. Hence the Bayes factors end up favoring the null hypothesis in the absence of extremely large samples. This behavior is more pronounced with the default iMOM Bayes factor, which even results in computational precision errors when the software attempts to compute Bayes factors for small effects.

16 For sample size $N \approx 2{,}820$, power reaches $1 - \beta \approx 0.669$.

17 Greater sample size than $N \approx 2{,}820$, leads to technical failure of the 'mombf' library to compute Bayes factors.

18 For sample size $N > 10{,}000$, power stays fixed at about $1 - \beta \approx 0.92$.

19 Similarly, Bayes factors fail to be computed for sample sizes greater than $N \approx 2{,}500$.

https://doi.org/10.26219/heal.aueb.6623

For non-inferiority and superiority testing (Tables 9, 10), results show very little variability across prior scales of the JZS Bayes factors. The overall conclusion is that large effects require smaller samples for similar levels of power.

JZS Bayes factor		$r = 0.3$	$r = 1/\sqrt{2}$	$r = 1.1$
Effect Size	**Operational Characteristic**	**Sample Size**	**Sample Size**	**Sample Size**
$d = -0.2$ (H_0)	$\alpha < 0.05$	$N = 900$	$N = 1{,}000$	$N = 1{,}000$
	$\alpha < 0.025$	$N = 1{,}350$	$N = 1{,}450$	$N = 1{,}500$
$d = -0.5$ (H_0)	$\alpha < 0.05$	$N = 70$	$N = 70$	$N = 70$
	$\alpha < 0.025$	$N = 100$	$N = 100$	$N = 100$
$d = -0.8$ (H_0)	$\alpha < 0.05$	$N = 24$	$N = 22$	$N = 22$
	$\alpha < 0.025$	$N = 40$	$N = 30$	$N = 30$
$d = 0.0$ (H_1)	$1 - \beta > 0.90$	$N = 850$	$N = 750$	$N = 750$
	$1 - \beta > 0.95$	$N = 1{,}300$	$N = 1{,}200$	$N = 1{,}200$
$d = 0.2$ (H_1)	$1 - \beta > 0.90$	$N = 80$	$N = 90$	$N = 90$
	$1 - \beta > 0.95$	$N = 130$	$N = 130$	$N = 130$
$d = 0.5$ (H_1)	$1 - \beta > 0.90$	$N = 20$	$N = 20$	$N = 20$
	$1 - \beta > 0.95$	$N = 30$	$N = 36$	$N = 36$
$d = 0.8$ (H_1)	$1 - \beta > 0.90$	$N = 10$	$N = 10$	$N = 10$
	$1 - \beta > 0.95$	$N = 16$	$N = 14$	$N = 16$

Table 9: Non-inferiority testing power analysis for the JZS Bayes factor with Cauchy prior scales r = 0.3, 1/√2, and 1.1.

https://doi.org/10.26219/heal.aueb.6623

JZS Bayes factor power analysis		**$r = 0.3$**	**$r = 1/\sqrt{2}$**	**$r = 1.1$**
Effect Size	**Operational Characteristic**	**Sample Size**	**Sample Size**	**Sample Size**
$d = -0.2$ (H_0)	$\alpha < 0.05$	$N = 260$	$N = 260$	$N = 260$
	$\alpha < 0.025$	$N = 380$	$N = 380$	$N = 390$
$d = -0.5$ (H_0)	$\alpha < 0.05$	$N = 48$	$N = 46$	$N = 50$
	$\alpha < 0.025$	$N = 64$	$N = 66$	$N = 70$
$d = -0.8$ (H_0)	$\alpha < 0.05$	$N = 18$	$N = 18$	$N = 20$
	$\alpha < 0.025$	$N = 26$	$N = 26$	$N = 30$
$d = 0.2$ (H_1)	$1 - \beta > 0.90$	$N = 170$	$N = 170$	$N = 170$
	$1 - \beta > 0.95$	$N = 280$	$N = 270$	$N = 280$
$d = 0.5$ (H_1)	$1 - \beta > 0.90$	$N = 28$	$N = 26$	$N = 28$
	$1 - \beta > 0.95$	$N = 46$	$N = 46$	$N = 48$
$d = 0.8$ (H_1)	$1 - \beta > 0.90$	$N = 10$	$N = 10$	$N = 10$
	$1 - \beta > 0.95$	$N = 18$	$N = 18$	$N = 20$

Table 10: Superiority testing power analysis for the JZS Bayes factor with Cauchy prior scales $r = 0.3$, $1/\sqrt{2}$, and 1.1.

4.4 Closing Remarks

We began this chapter by introducing two classes of non-local priors, and explaining how they can be utilized for Bayesian variable selection. Based on formulating the two group means comparison as a variable selection problem, we planned and implemented our simulation experiment. While the behavior of Bayes factors in non-inferiority and superiority tests received attention, the main focus of the simulation study was the use of local and non-local priors in equivalence testing. In the chapter's final sections, we presented graphs and power analysis tables accompanied by commentary, for a variety of problems. The main conclusions and takeaways are reserved for the next and final chapter, which is the epilogue of the thesis.

CHAPTER 5

EPILOGUE

5.1 Main Findings

After completing our study, the following general conclusions emerge:

- In equivalence testing, a higher Cauchy prior scale *r* for the JZS Bayes factor leads to greater tendency to support the null hypothesis. This is expected, as higher scale *r* assigns lower density to $d = 0$ under the alternative hypothesis.
- The MOM and iMOM Bayes factors do indeed speed up the accumulation of evidence for H_0 when it is true. However, correct adjustment of parameters is of extreme importance if we want to avoid misleading conclusions when true effects are present (and hence H_0 is not valid).
- The authors' default MOM and iMOM specifications, explicitly tuned to detect standardized effects $d > 0.2$, should not discourage analysts from experimenting with their own custom specifications. In fact, our position is that since even small effects ($d = 0.2$) are easily detected by the JZS Bayes factor, there is no advantage in employing MOM and iMOM Bayes factors in such cases. Aside from supporting H_0 sooner when it is true, it is not obvious that MOM and iMOM Bayes factors offer any other practical benefit.
- Judging by how differently the various types of Bayes factors behave, it may not be sensible to rely on a single type (or a single set of specifications) when conducting hypothesis tests. Instead, it is prudent

https://doi.org/10.26219/heal.aueb.6623

to use many Bayes factor types in the same test, and take them all into account before drawing a conclusion.

- For tests of non-inferiority and superiority, a lower Cauchy prior scale r is associated with JZS Bayes factors that have a greater tendency to stay close to the horizontal line $y = 0$ ($y = 1$ in the $2log_e(BF_{01})$ scale). In other words, small scales r result in Bayes factors that need greater amounts of data to provide strong evidence, in *any* direction. Conversely, for higher scales r, Bayes factors portray the evidence in *any* direction more favorably.
- With regards to power analysis, high levels of power are achieved with reasonable and realistic sample sizes, for any scenario we have studied in this dissertation. Where this is not the case, tuning of parameters, or switching to a different Bayes factor type can offer a solution.

5.2 Discussion

Simulation experiments aiming to explore the behavior of various Bayes factor types in hypothesis tests rely heavily on computational efficiency and speed, especially when assessment of operational characteristics is a goal. Examples of Bayes factors that are not prohibitively cumbersome to calculate include those using g, hyper-g, and hyper-g/n priors. Although we elected to use versions of the JZS Bayes factor in our experiment, the above choices would have probably been easy to implement, even in the absence of specialized software. On the other hand, the Bayes factor variants described in Section 2.2 are not only harder and more time consuming to calculate, but also lack many desirable properties in terms of 'objective' inference to warrant the effort under time constraints.

Other areas of potential interest that were not explored in this dissertation include:

- Bayes factors for hypothesis testing where the clinical outcome is a categorical measure, instead of a continuous one.

 https://doi.org/10.26219/heal.aueb.6623

- Other kinds of priors, such as the exponential moment (eMOM) non-local prior (Rossell & Telesca, 2017), or a spike-and-slab approach when the equivalence test's null hypothesis is an interval.
- The application of Bayes factors for hypothesis testing on a real clinical dataset, accompanied by sequential monitoring of the evidence.
- Incorporating elements of decision theory into our simulation experiment, and comparing the behaviors of different Bayes factor types with regards to selected loss functions.

Given the above, we cannot claim that this dissertation is even close to an exhaustive study of Bayes factors, or of their implementation in clinical trials. Bayes factors are not as popular as classical hypothesis tests, but the volume of different types and variants has crossed the point where a complete and thorough assessment could fit into any single study. However, we believe that the insights concerning the use of Bayes factors with local and non-local priors can be valuable, since they have not been yet studied (to our knowledge) in the context of clinical trial hypothesis testing. A secondary goal, mostly successful, has been to demonstrate that the operational characteristics of Bayesian tests can be assessed through Monte Carlo simulations.

Furthermore, studying and employing Bayes factors has led to a renewed appreciation for frequentist hypothesis testing. Using and interpreting Bayes factors correctly, requires significantly deeper mathematical knowledge compared to the simple option of calculating a *p*-value. It is easy to see why the frequentist approach is attractive to so many statistical analysts, without discounting the advantages of Bayes factors. It is our view that the popularity of Bayes factors would be greatly boosted by the development of appropriate Bayesian testing software, offering a greater variety of Bayes factor types for different tests. The R libraries used in our simulation study, as well as the point-and-click statistical program JASP (JASP Team, 2020) are steps in the right direction.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623

APPENDIX

FIGURES

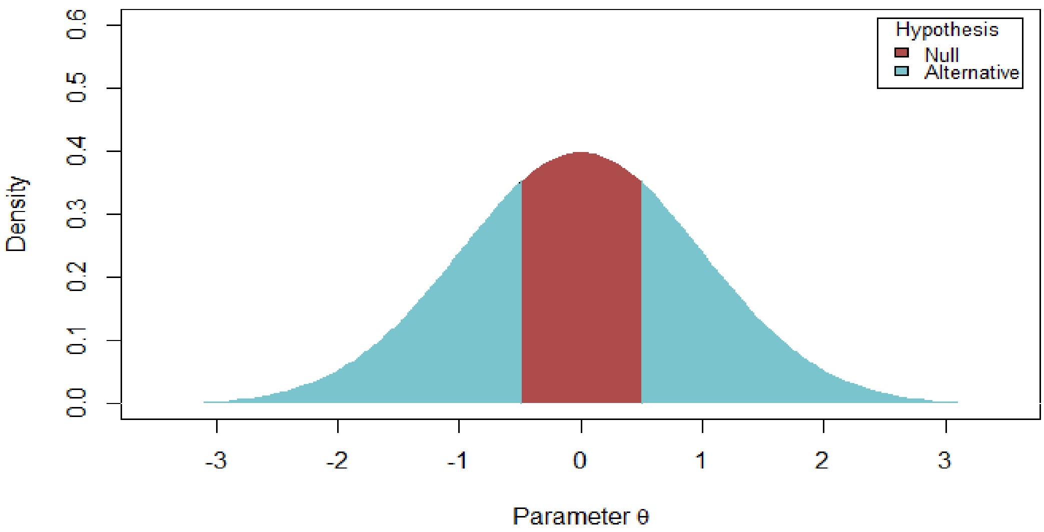

Figure 30: Hypotheses with non-overlapping priors.

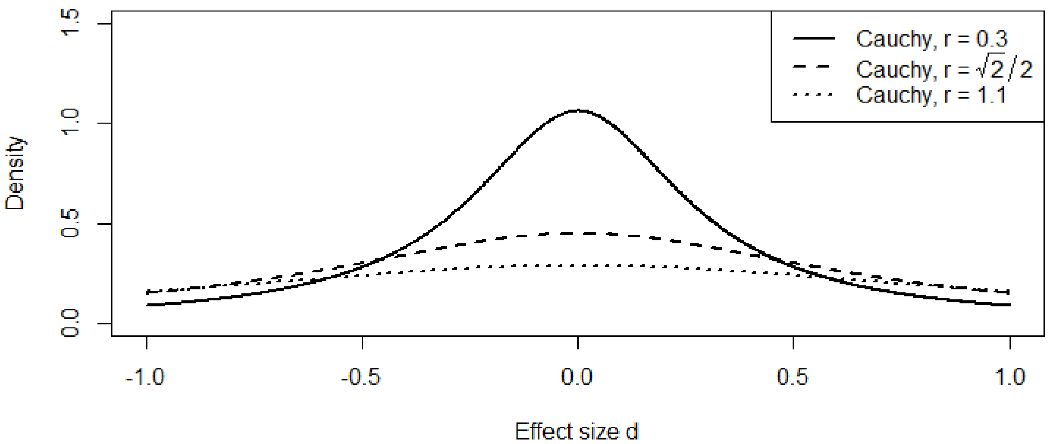

Figure 31: Local priors for equivalence testing.

https://doi.org/10.26219/heal.aueb.6623

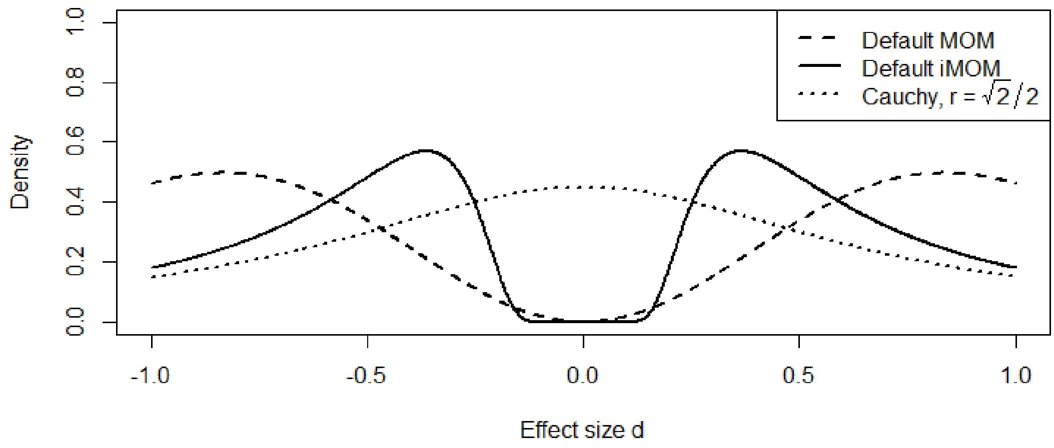

Figure 32: Cauchy and default MOM, iMOM priors.

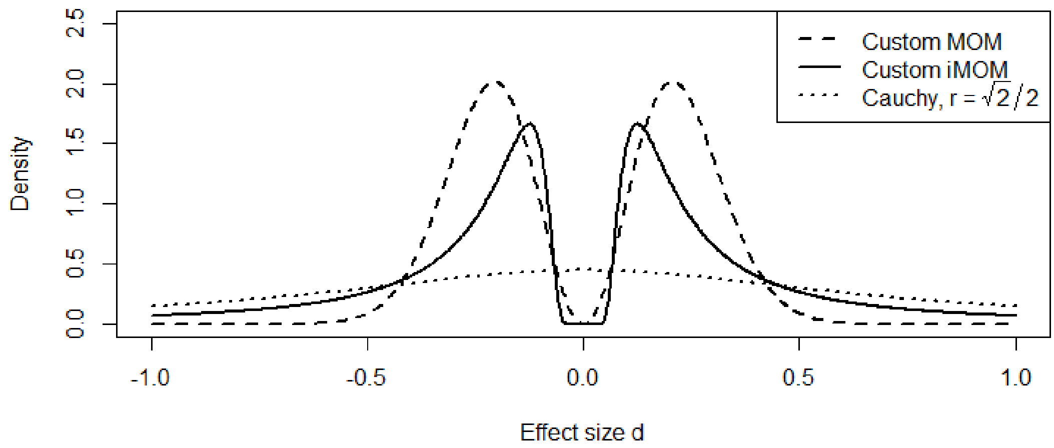

Figure 33: Cauchy and custom MOM, iMOM priors.

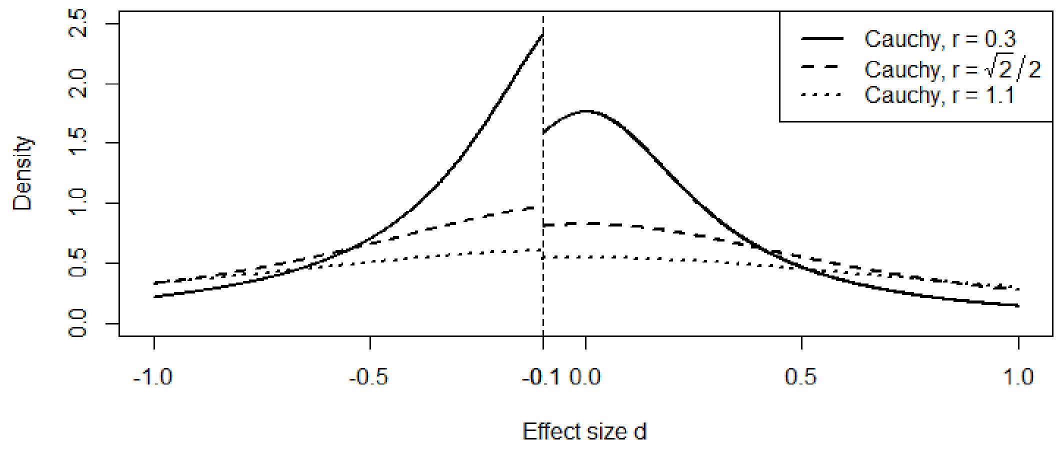

Figure 34: Cauchy priors for non-inferiority testing.

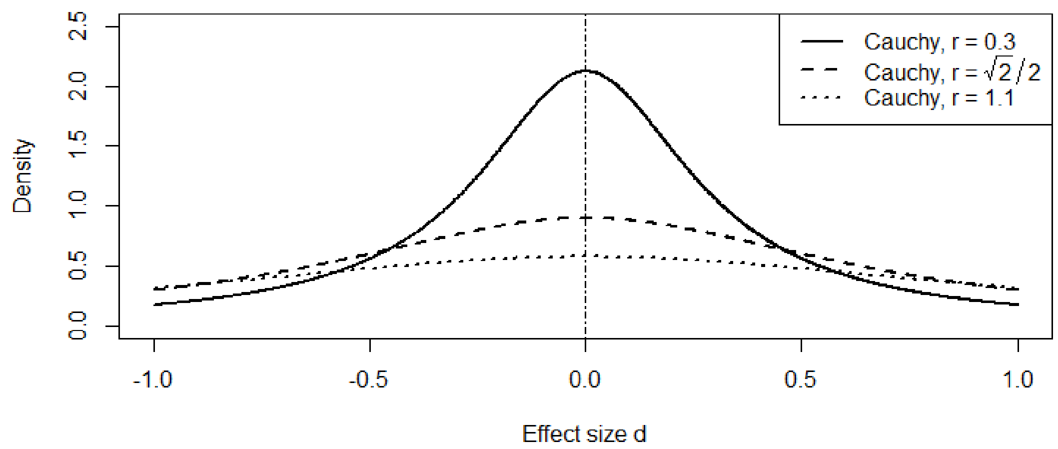

Figure 35: Cauchy priors for superiority testing.

 https://doi.org/10.26219/heal.aueb.6623

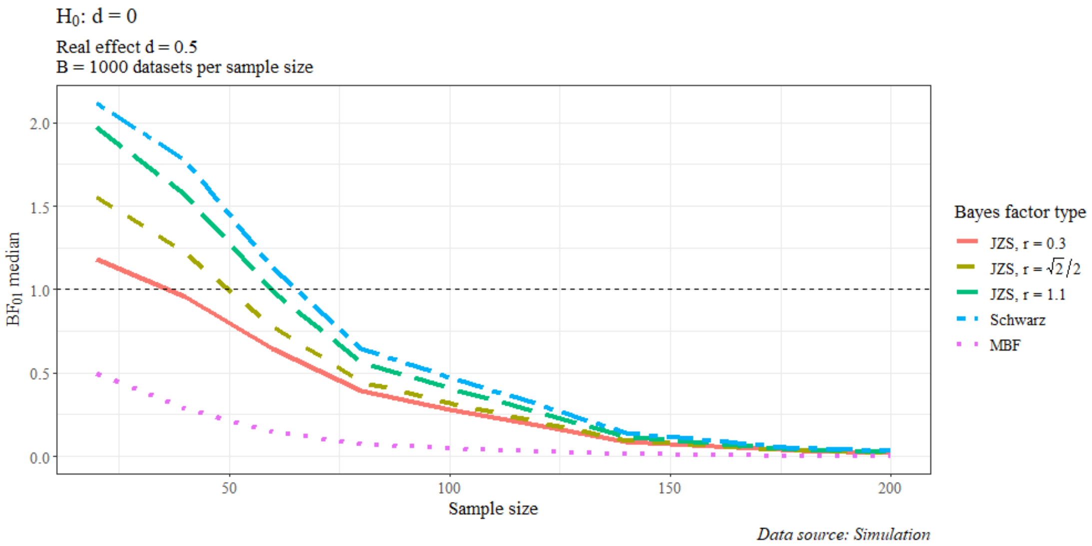

Figure 36: Equivalence testing with local priors.

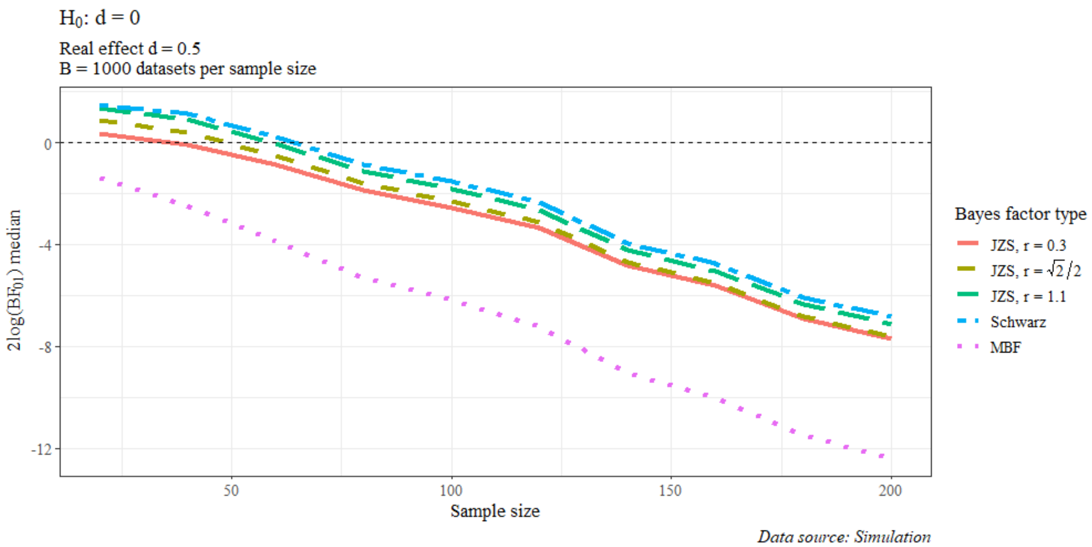

Figure 37: Equivalence testing with local priors.

https://doi.org/10.26219/heal.aueb.6623

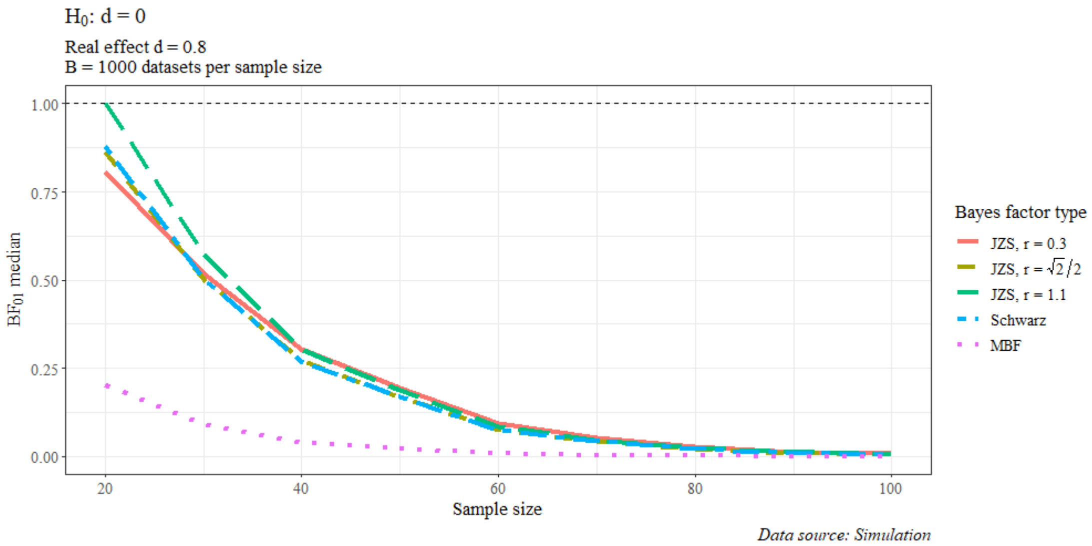

Figure 38: Equivalence testing with local priors.

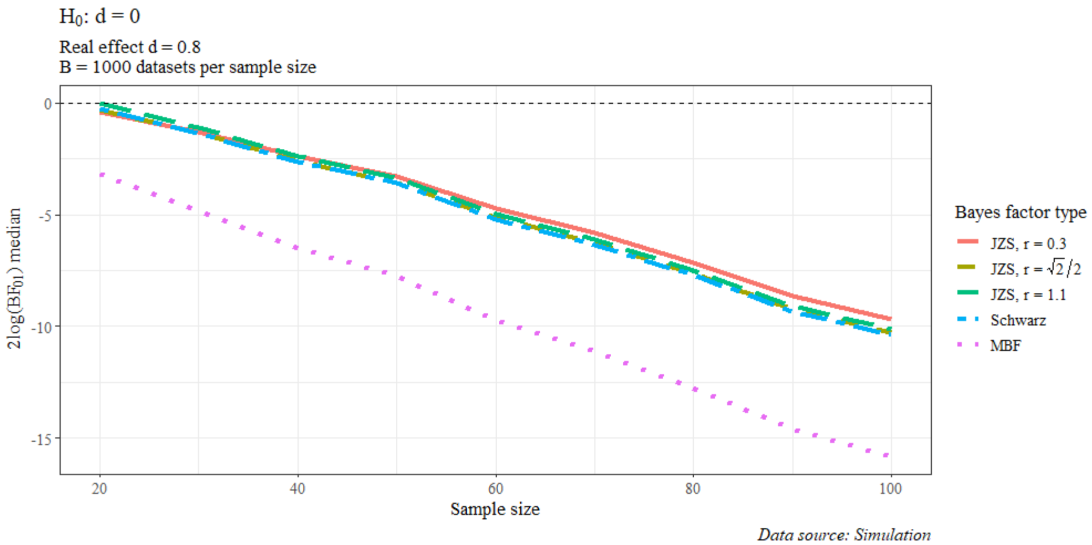

Figure 39: Equivalence testing with local priors.

 https://doi.org/10.26219/heal.aueb.6623

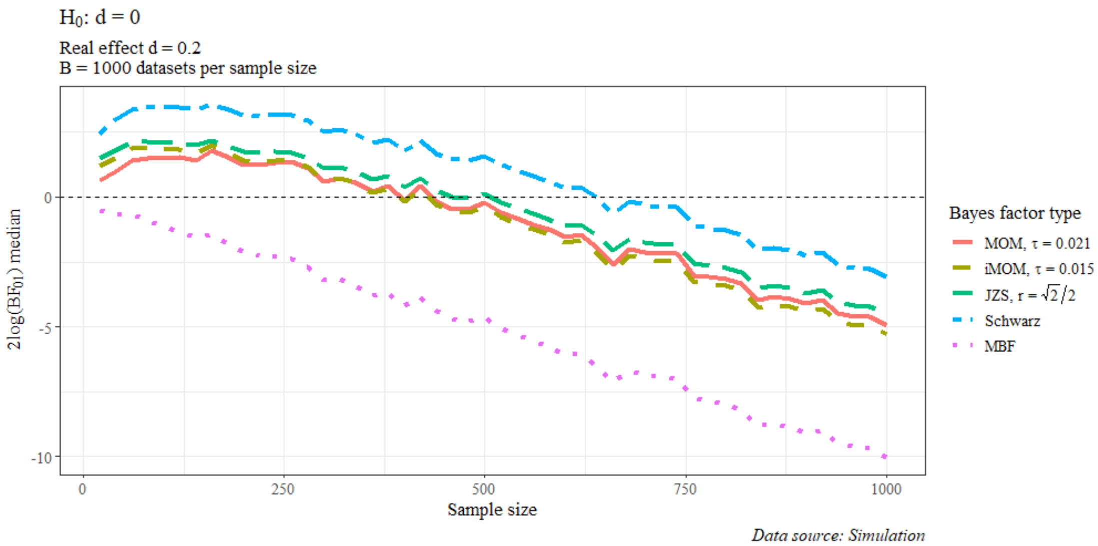

Figure 40: Equivalence testing with custom non-local priors.

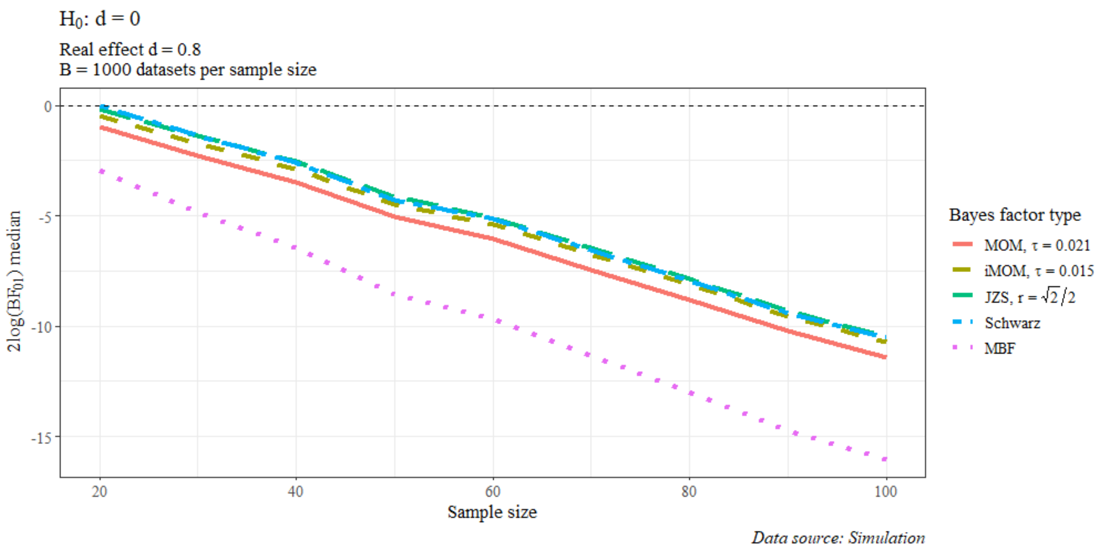

Figure 41: Equivalence testing with custom non-local priors.

https://doi.org/10.26219/heal.aueb.6623

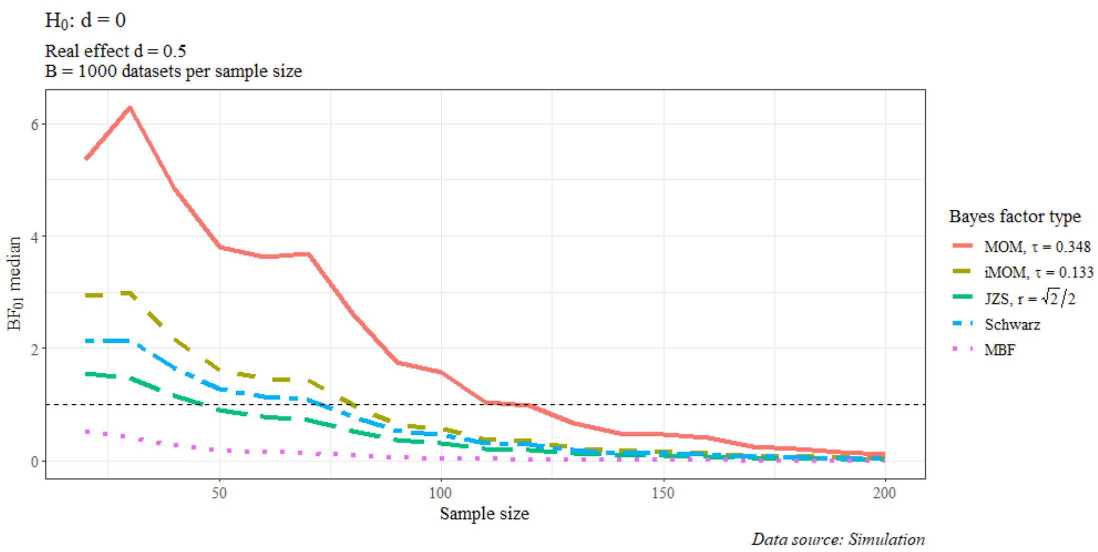

Figure 42: Equivalence testing with default non-local priors.

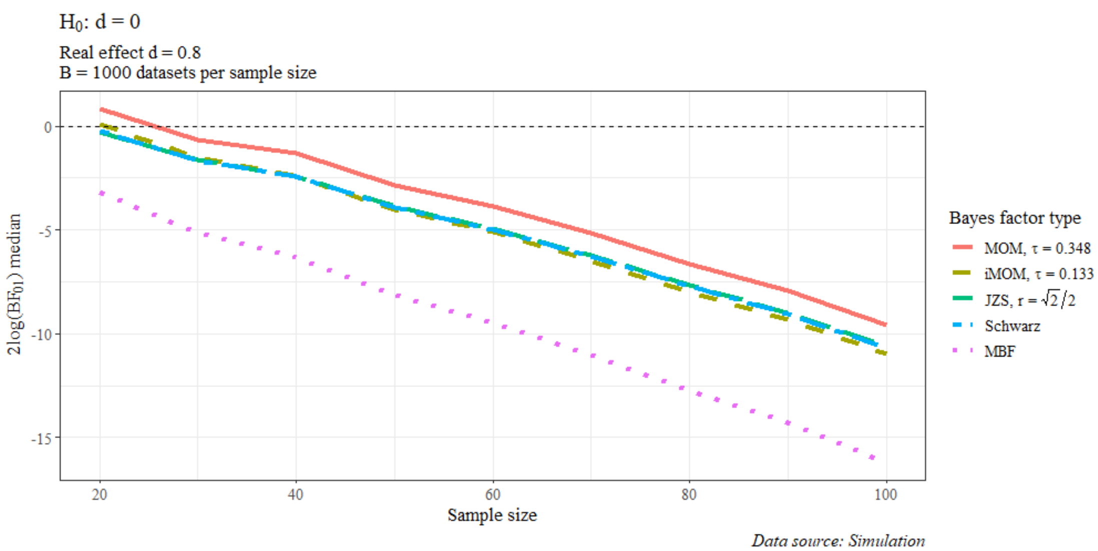

Figure 43: Equivalence testing with default non-local priors.

https://doi.org/10.26219/heal.aueb.6623

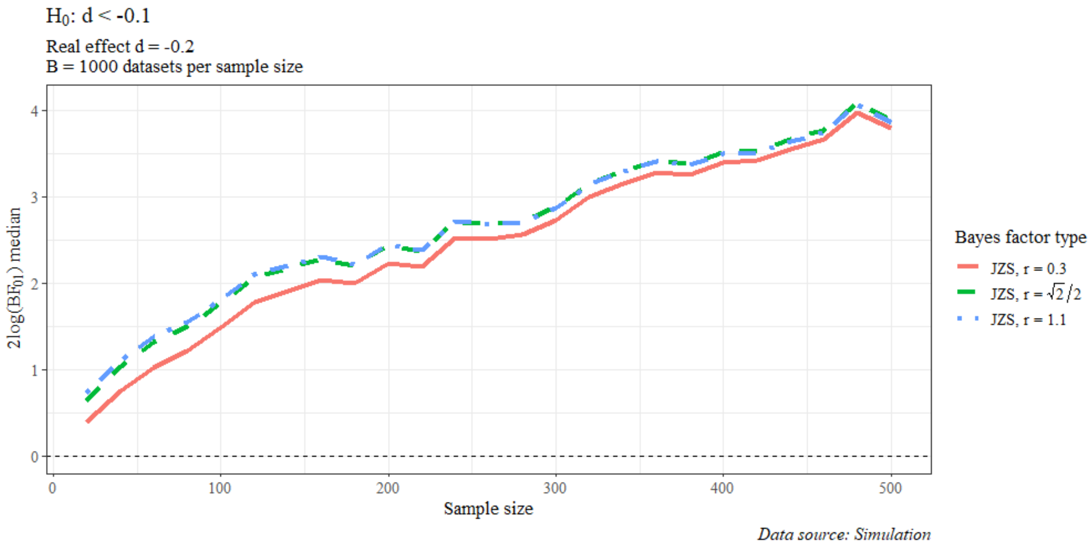

Figure 44: Non-inferiority testing.

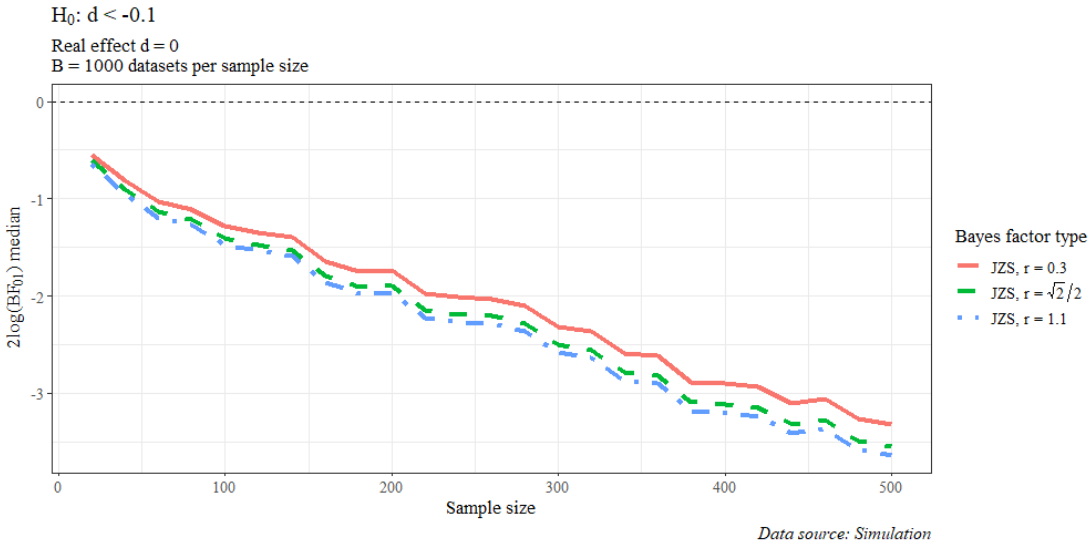

Figure 45: Non-inferiority testing.

https://doi.org/10.26219/heal.aueb.6623

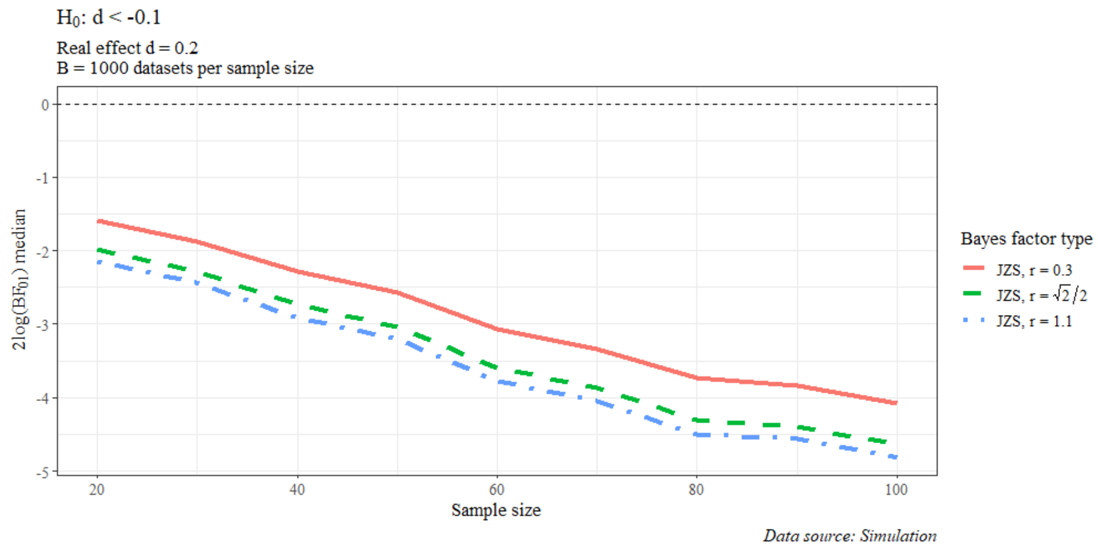

Figure 46: Non-inferiority testing.

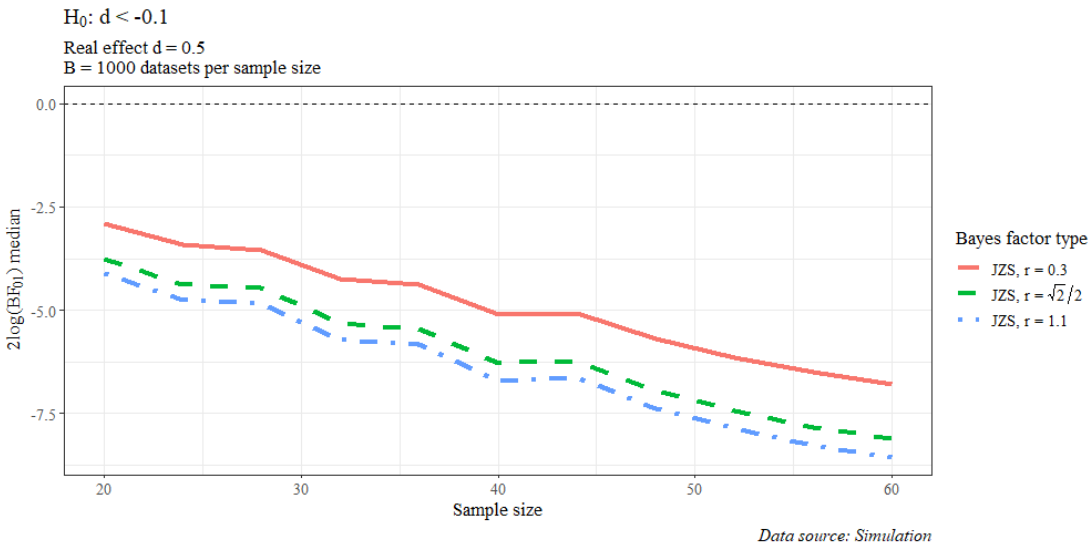

Figure 47: Non-inferiority testing.

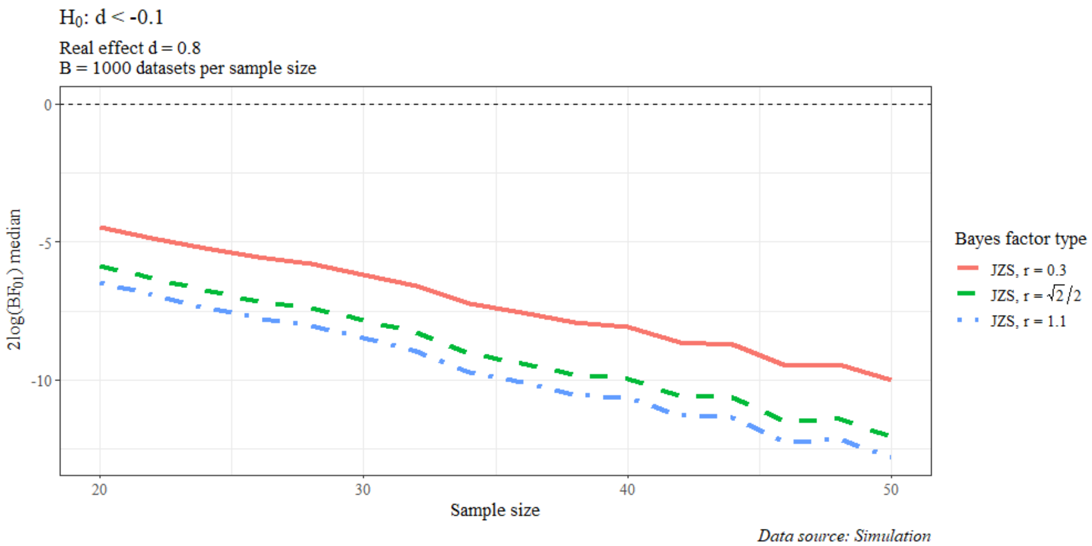

Figure 48: Non-inferiority testing.

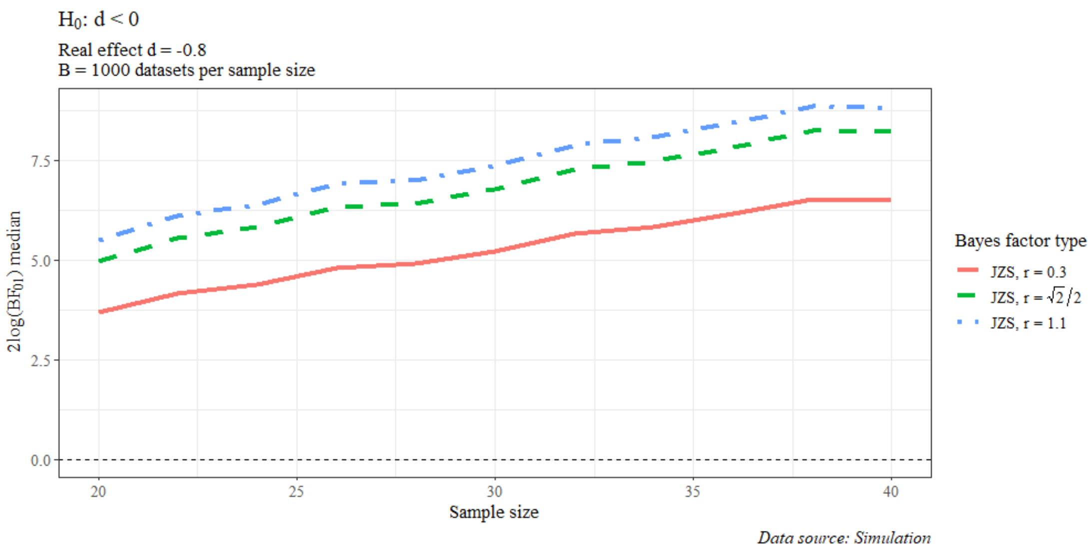

Figure 49: Superiority testing.

https://doi.org/10.26219/heal.aueb.6623

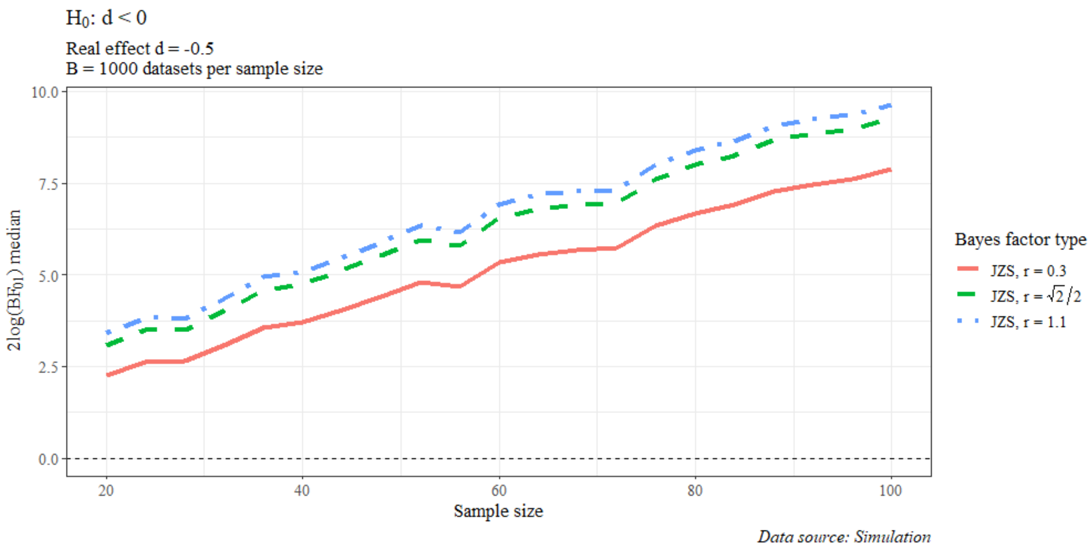

Figure 50: Superiority testing.

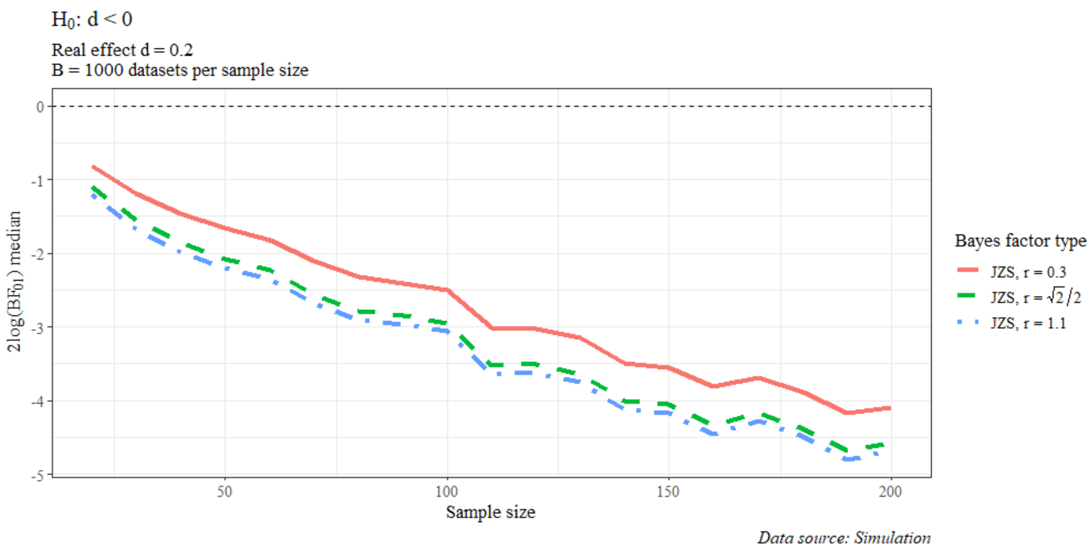

Figure 51: Superiority testing.

https://doi.org/10.26219/heal.aueb.6623

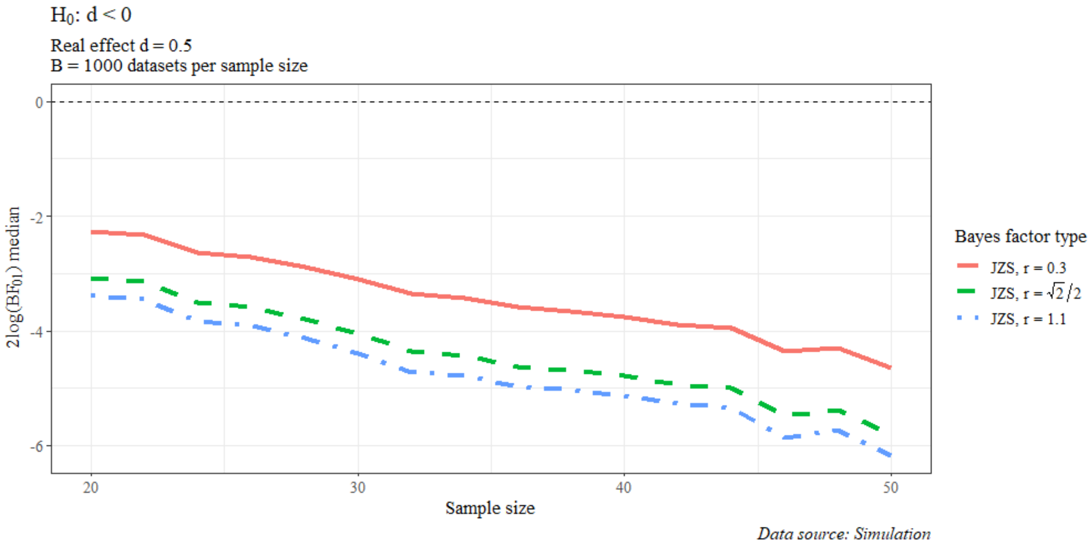

Figure 52: Superiority testing.

https://doi.org/10.26219/heal.aueb.6623

REFERENCES

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1), 111-128.

Albers, C. J., Kiers, H. A., & van Ravenzwaaij, D. (2018). Credible Confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology,* 4(1), 31.

Altman, D. G., & Bland, J. M. (1995). Statistics Notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003), 485.

Armitage, P., Berry, G., & Matthews, J. N. S. (2008). *Statistical Methods in Medical Research*. John Wiley & Sons.

Ashby, D., & Smith, A. F. (2000). Evidence-based medicine as Bayesian decision-making. *Statistics in Medicine*, 19(23), 3291-3305.

Bååth, R. (2018). beepr: Easily Play Notification Sounds on any Platform. R package version 1.3. https://CRAN.R-project.org/package=beepr.

Bartlett, M. S. (1957). Comment on 'A statistical paradox' by DV Lindley. *Biometrika*, 44(1-2), 533-534.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1), 58-80.

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), 1550-1577.

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London,* (53), 370-418.

Berger, J. O., & Pericchi, L. R. (1996a). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433), 109-122.

https://doi.org/10.26219/heal.aueb.6623

Berger, J. O., & Pericchi, L. R. (1996b). The intrinsic Bayes factor for linear models. *Bayesian statistics*, 5, 25-44.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397), 112-122.

Berger, J. O. & Wolpert, R. L. (1988). *The Likelihood Principle*, Monograph Series 6. IMS, Hayward, CA.

Berger, J. O., Brown, L. D., & Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, 22(4), 1787-1807.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3), 385-402.

Berger, J.. (2008). A comparison of testing methodologies. *PHYSTAT LHC Workshop on Statistical Issues for LHC Physics*, PHYSTAT 2007 - Proceedings. 8-19.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian Theory*. John Wiley & Sons.

Berry, D. A. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1), 27-36.

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research®,* 468(3), 885-892.

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. In *Compstat* (pp. 111-116). Physica, Heidelberg.

Campbell, G. (2017). Bayesian methods in clinical trials with applications to medical devices. *Communications for Statistical Applications and Methods*, 24(6), 561-581.

Carlin, B. P., and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Boca Raton: Chapman & Hall/CRC.

Casella, G., & Berger, R. L. (2002). *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury.

https://doi.org/10.26219/heal.aueb.6623

Chan, A. W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J. A., Doré, C. J., Parulekar, W. R., Summerskill, W. S. M., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D., & Moher, D. (2013). SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Annals of Internal Medicine*, 158(3), 200-207.

Chang, M., & Boral, A. (2008). ABC of Bayesian approaches to drug development. *Pharmaceutical Medicine*, 22(3), 141-150.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432), 1313-1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453), 270-281.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). Things I have learned (so far). In *Annual Convention of the American Psychological Association*, 98th, Aug, 1990, Boston, MA, US; Presented at the aforementioned conference. American Psychological Association.

Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2), 627-679.

Dawid, A. P., Stone, M., & Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(2), 189-213.

DiCiccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439), 903-915.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4), 409-425.

https://doi.org/10.26219/heal.aueb.6623

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193.

European Medicines Agency, Committee for Medicinal Products for Human Use (2014). Guideline on the investigation of subgroups in confirmatory clinical trials.

European Medicines Agency, Committee for Medicinal Products for Human Use (2017). Guideline on multiplicity issues in clinical trials.

Fienberg, S. E. (2006). Does it make sense to be an "objective Bayesian"? (comment on articles by Berger and by Goldstein). *Bayesian Analysis*, 1(3), 429-432.

Fisher, R. A. (1973). *Statistical Methods and Scientific Inference*. 3rd ed. Macmillan, New York.

Fisher, R. A. (1992). Statistical Methods for Research Workers. In *Breakthroughs in* Statistics. Springer, New York, NY.

Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2015). *Fundamentals of Clinical Trials*. Springer.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365), 153-160.

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3), 501-514.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association,* 85(410), 398-409.

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4), 967-1033.

Gelman, A., & Meng, X. L. (1994). Path sampling for computing normalizing constants: identities and theory. University of Chicago Department of Statistics Technical Report, (377).

 https://doi.org/10.26219/heal.aueb.6623

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis,* 3(3), 445-449.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* (6), 721-741.

Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach.* 3rd ed. CRC press.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, 59(3), 252-257.

Good, I. J. (1985). Weight of evidence: A brief survey. *Bayesian statistics*, (2), 249-270.

Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419), 597-606.

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology,* 137(5), 485-496.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine,* 130(12), 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine* 130(12), 1005-1013.

Goodman, S. N. (2001). Of P-values and Bayes: A Modest Proposal. *Epidemiology*, 12(3), 295-297.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in Hematology*, 45(3), 135-140.

Gronau, Q. F., Ly, A., & Wagenmakers, E. J. (2020). Informed Bayesian t-tests. *The American Statistician*, 74(2), 137-143.

Han, C., & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96(455), 1122-1132.

Held, L. & Ott, M. (2018). On *p*-values and Bayes Factors. *Annual Review of Statistics and its Applications*, 5(1), 393-419.

https://doi.org/10.26219/heal.aueb.6623

Hodges Jr, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2), 261-268.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. Chicago.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(4), 267-268.

Hubbard, R., & Bayarri, M. J. (2003). P values are not error probabilities. *Institute of Statistics and Decision Sciences,* Working Paper, (03-26), 27708-0251.

Huntjens, R. J., Peters, M. L., Woertman, L., Bovenschen, L. M., Martin, R. C., & Postma, A. (2006). Inter-identity amnesia in dissociative identity disorder: A simulated memory impairment?. *Psychological Medicine*, 36(6), 857.

Ibrahim, J. G., Chen, M. H., & Chu, H. (2012). Bayesian methods in clinical trials: a Bayesian analysis of ECOG trials E1684 and E1690. *BMC Medical Research Methodology*, 12(1), 183.

Inoue, L. Y., Berry, D. A., & Parmigiani, G. (2005). Relationship between Bayesian and frequentist sample size determination. *The American Statistician*, 59(1), 79-87.

International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (1998). ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials E9.

Jack Lee, J., & Chu, C. T. (2012). Bayesian clinical trials in action. *Statistics in Medicine*, 31(25), 2955-2972.

JASP Team, (2020). JASP (Version 0.14)[Computer Software]. https://jasp-stats.org/.

Jeffreys, H. (1939). *Theory of Probability,* 1st ed. The Clarendon Press, Oxford.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2), 143-170.

Johnson, V. E., & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498), 649-660.

Kalbfleisch, J. G., & Sprott, D. A. (1976). On tests of significance. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,* (Eds) W. L. Harper and C. A. Hooker, Dordrecht: Reidel, 259-270.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795.

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 42(5), 551-560.

Kim, J. H., & Robinson, A. P. (2019). Interval-based hypothesis testing and its applications to economics and finance. *Econometrics*, 7(2), 21.

Klugkist, I., Laudy, O., & Hoiktink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods,* 10(4), 477.

Lavine, M., & Schervish, M. J., (1999). Bayes factors: What they are and what they are not. *The American Statistician,* 53(2), 119-122.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. Journal of the American Statistical Association, 92(438), 648-655.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 410-423.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4), 906-917.

Linde, M., & van Ravenzwaaij, D. (2019). baymedr: Computation of Bayes Factors for Common Biomedical Designs. R package version 0.1.0. https://CRAN.R-project.org/package=baymedr.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2), 187-192.

https://doi.org/10.26219/heal.aueb.6623

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428), 1535-1546.

Mahan, V. L. (2014). Clinical Trial Phases. *International Journal of Clinical Medicine*, 5(21), 1374.

Meng, X. L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831-860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21(6), 1087-1092.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406.

Morey, R. D. & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1), 3-26.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrica,* 20A(1/2), 175-240.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,* 231(694-706), 289-337.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS* (Vol. 698). John Wiley & Sons.

O' Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 99-118.

Page, R., & Satake, E. (2017). Beyond P Values and Hypothesis Testing: Using the Minimum Bayes Factor to Teach Statistical Inference in

 https://doi.org/10.26219/heal.aueb.6623

Undergraduate Introductory Statistics Courses. *Journal of Education and Learning*, 6(4), 254-266.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* 50(302), 157-175.

Perrakis, K., Ntzoufras, I., & Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Computational Statistics & Data Analysis*, 77, 54-69.

Peto, R., Pike, M., Armitage, P., Breslow, et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer,* 34(6), 585-612.

Pocock, S. J. (2013). *Clinical Trials: A Practical Approach*. John Wiley & Sons.

Quan, H., Zhang, B., Lan, Y., Luo, X., & Chen, X. (2019). Bayesian hypothesis testing with frequentist characteristics in clinical trials. *Contemporary Clinical Trials*, 87, 105858.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raftery, A. (1996). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman and Hall, pp. 165-187.

Raftery, A., Hoeting, J., & Madigan, D. (1993). Model selection and accounting for model uncertainty in linear regression models. *Techical Report No. 262*, Department of Statistics, University of Washington.

Raftery, A. E., Madigan, D., & Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics*, 5, 323-349.

Robert, C., & Casella, G. (2013). *Monte Carlo Statistical Methods.* Springer Science & Business Media.

https://doi.org/10.26219/heal.aueb.6623

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer Science & Business Media.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2), 216-232. Chicago.

Rossell, D., Cook, J., D., Telesca, D., Roebuck, P., & Abril, O. (2020). mombf: Bayesian Model Selection and Averaging for Non-Local and Local Priors. R package version 3.0.1. https://CRAN.R-project.org/package=mombf.

Rossell, D., & Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517), 254-265.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225-237.

RStudio Team, (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. https://www.rstudio.com.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128-142.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, 77(378), 325-334.

Smith, A. F., & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2), 213-220.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox?. *Philosophy of Science*, 80(1), 73-93.

Spiegelhalter, D. J., & Smith, A. F. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3), 377-387.

https://doi.org/10.26219/heal.aueb.6623

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1993). Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, 12(15-16), 1501-1511.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M. K. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157(3), 357-387.

Sprenger, J. (2013). Testing a precise null hypothesis: The case of Lindley's paradox. *Philosophy of Science*, 80(5), 733-744.

Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science,* 359-363.

Stigler, S. M. (2005). PS Laplace, Théorie analytique des probabilités, (1812); Essai philosophique sur les probabilités, (1814). In *Landmark Writings in Western Mathematics 1640-1940* (pp. 329-340), Elsevier Science.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

U. S. Food and Drug Administration, Center for Drug Evaluation and Research (2017). Multiple Endpoints in Clinical Trials Guidance for Industry.

U. S. Food and Drug Administration, Office of the Commissioner & Office of Clinical Policy and Programs. (2010). *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials*.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89, 31-50.

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. (2019). Bayes Factors for Superiority, Non-inferiority, and Equivalence Designs. *BMC Medical Research Methodology*, 19(1), 1-12.

Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian Evaluation of Informative Hypotheses,* 181-207. Springer, New York, NY.

https://doi.org/10.26219/heal.aueb.6623

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60(3), 158-189.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25(1), 35-57.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779-804.

Wang, M., & Liu, G. (2016). A simple two-sample Bayesian t-test for hypothesis testing. *The American Statistician*, 70(2), 195-201.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752-760.

Whitehead, J., Valdés-Márquez, E., Johnson, P., & Graham, G. (2008). Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, 27(13), 2307-2327.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wickham, H. (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN.R-project.org/package=tidyr.

Wilkinson, L., (1999). The Task Force on Statistical Inference: Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, 54, 594-604.

World Medical Association. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79(4), 373.

Xuefeng Li & Yunling Xu (2011). Guest Editors' Note: Bayesian Approaches for Medical Product Evaluation—Today and Tomorrow, *Journal of Biopharmaceutical Statistics*, 21(5), 869-870.

https://doi.org/10.26219/heal.aueb.6623

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith*, (Eds) Bayesian Statistics, Proceedings of the First International Meeting, Valencia: University of Valencia Press, 585-603.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions. In Goel, P.; Zellner, A. (Eds) Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. *Studies in Bayesian Econometrics and Statistics*. 6. New York: Elsevier. pp. 233–243.

https://doi.org/10.26219/heal.aueb.6623

https://doi.org/10.26219/heal.aueb.6623