A Sequential Approach to Estimating Bayesian P-values through Simulation Studies

Johnpaul Chiemelie Anamege

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation

This Thesis is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.
A SEQUENTIAL APPROACH TO ESTIMATING BAYESIAN P-VALUES IN SIMULATION STUDIES

by

Johnpaul Chiemelie Anamage

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

Major: Mathematical Sciences

The University of Memphis
August 2024
Acknowledgments

The completion of this thesis would not have been possible without the support and guidance of many individuals to whom I am deeply grateful.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Ebenezer O. George, for his invaluable guidance, continuous support, and patience throughout my research journey. His profound knowledge and insightful feedback have been instrumental in shaping this work.

My heartfelt appreciation goes to the members of my thesis committee, Dr. Baris Kopruluoglu and Dr. Majid Noroozi, for their time, constructive criticism, and encouragement. Their diverse perspectives have greatly enhanced the quality of this research.

I am deeply indebted to my family for their unwavering support and endless patience. Thank you for always believing in me and encouraging me to pursue my dreams. Your love and support have been my anchor throughout this journey.

This accomplishment would not have been possible without all of you. Thank you.
Abstract

This thesis introduces a novel sequential approach for calculating Bayesian P-values, addressing limitations of traditional frequentist methods. Using Monte Carlo simulations, we investigate the properties of sequentially calculated Bayesian P-values under various conditions, comparing them to frequentist counterparts.

Our findings demonstrate advantages including incorporation of prior knowledge, continuous evidence measurement, and adaptability to accumulating data. We explore computational challenges, propose optimized algorithms, and address the impact of prior specification, providing guidelines for choosing appropriate priors.

This research contributes to the ongoing discussion about P-values in scientific inference and offers a practical framework for researchers adopting Bayesian methods in sequential analyses. The proposed approach has potential applications across various disciplines, particularly where data is collected sequentially, or early decision-making is crucial. Our work bridges the gap between frequentist and Bayesian methods, offering a more nuanced tool for statistical inference.

Keywords: Bayesian P-values, sequential analysis, Monte Carlo simulation, statistical inference, computational statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Literature Review</td>
<td>5</td>
</tr>
<tr>
<td>Bayesian P-values</td>
<td>5</td>
</tr>
<tr>
<td>General Framework</td>
<td>6</td>
</tr>
<tr>
<td>Properties</td>
<td>16</td>
</tr>
<tr>
<td>3 Research Objective</td>
<td>18</td>
</tr>
<tr>
<td>Sequential Bayesian Analysis</td>
<td>19</td>
</tr>
<tr>
<td>4 Examples and Applications</td>
<td>22</td>
</tr>
<tr>
<td>5 Advantages and Limitations</td>
<td>38</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>41</td>
</tr>
<tr>
<td>References</td>
<td>43</td>
</tr>
<tr>
<td>A R Code</td>
<td>48</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Null hypothesis $H_0$ true: $\mu = 0$ and $n = 100$</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>Alternative hypothesis $H_1$ with $n = 100$ and small effect: $\mu = 0.1$</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>Alternative hypothesis $H_1$ with $n = 100$ and medium effect: $\mu = 0.5$</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>Alternative hypothesis $H_1$ with $n = 100$ and large effect: $\mu = 2$</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>Null hypothesis $H_0$ true: $\mu = 0$ and $n = 1000$</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>Alternative hypothesis $H_1$ with $n = 1000$ and small effect: $\mu = 0.1$</td>
<td>34</td>
</tr>
<tr>
<td>7</td>
<td>Alternative hypothesis $H_1$ with $n = 1000$ and medium effect: $\mu = 0.5$</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>Alternative hypothesis $H_1$ with $n = 1000$ and large effect: $\mu = 2$</td>
<td>36</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Hypothesis testing is a fundamental aspect of statistical inference. There are different methods for evaluating hypothesis tests, the most common of which involves interpreting P-values. P-values play a crucial role in determining the significance of research findings. The use of P-values to assess the statistical significance of experimental results has historically been a cornerstone of statistical hypothesis testing. A P-value is the probability of observing an effect that is at least as extreme as the one obtained in sample data, assuming the null hypothesis is true [Berger and Delampady, 1987]. Traditionally, P-values are calculated using frequentist methods, which rely on the sampling distribution of a test statistic over repeated experiments under the null hypothesis. However, in recent years, there has been growing interest in Bayesian approaches to hypothesis testing. This introduction will discuss the limitations of frequentist P-values as compared to Bayesian P-values.

Despite their widespread use, P-values have several limitations and are often criticized for their potential to mislead researchers [Wagenmakers, 2007]. A common misconception of P-values is that they measure the probability that the null hypothesis is true. For example, when testing a hypothesis, we may run into a solution that produces a P-value of 0.05. This probability is often misinterpreted to mean that there is a five percent chance that the null hypothesis is true. This is in fact incorrect. A P-value, in frequentist context, does not represent the probability that the null hypothesis itself is true. A key limitation of P-values is that they do not measure the size or importance of an effect. In the context of statistical analysis, the size of an effect, also known as effect size, refers to the size or strength of the relationship between variables or the difference between groups [Cohen, 1988]. Effect size is a quantitative measure that is independent of sample size, unlike P-values. A very small effect can have a significant P-value with a large enough sample size. It is important to keep in mind that practical significance is very different from statistical significance. Another major limitation of P-values is their sensitivity to sample
size. With large sample sizes, even small and practically insignificant differences can lead to statistically significant P-values [Berger and Sellke, 1987]. Conversely, with small sample sizes, large and potentially important differences may fail to reach statistical significance. Increasing the sample size makes it more likely to obtain a “significant” P-value even if the effect size is small.

While P-values can be a useful tool, over-reliance on them without considering these limitations can lead to misinterpretations and flawed conclusions [Wasserstein and Lazar, 2016]. Although many statisticians advocate for a more holistic interpretation of results that goes beyond solely focusing on P-values [Greenland et al., 2016], many others prefer to use alternative methods for calculating P-values. These alternative approaches were created to address some of the limitations of frequentist P-values. One such alternative is Sequential Bayesian P-values. The purpose of the paper is to introduce, with examples and applications, a Sequential approach to computing Bayesian P-values.

A Bayesian P-value calculates the probability that the null hypothesis is true given that the test statistic is as extreme as or more extreme than the one observed [Sellke et al., 2001]. In other words, Bayesian P-values quantify the evidence against the null hypothesis taking both the data and the prior information into consideration. A key challenge in calculating Bayesian P-values is that they require specifying a prior distribution for the parameters of interest. A prior distribution represents the relative plausibility of different parameter values before seeing the data. It is the distribution of a set of data based on prior knowledge about the data. In many situations, it can be difficult to specify an informative prior. A common alternative to this involves specifying an objective, non-informative, or ”default” prior. However, the choice of default prior can have a substantial impact on the resulting Bayesian inferences, including the Bayesian P-value. To address this issue, several authors have proposed using a sequential approach to specifying the prior distribution [Berger and Pericchi, 1996]. Under this approach, statisticians use a small fraction of the data to update a non-informative initial prior which is employed to obtain a data-dependent default prior. This default prior is then applied in computing
the Bayesian P-value using the remaining data. By utilizing the data to guide the choice of prior, this approach can help to ensure that the prior is both objective and well-calibrated to the observed data.

Building on this idea, Berger and Pericchi introduced the concept of “intrinsic Bayes factors,” which can be defined as the ratio of the marginal likelihoods under the null and alternative hypotheses, averaged over all possible training samples of a given size [Berger and Pericchi, 1996]. Berger and Pericchi showed that the intrinsic Bayes factor provides a general solution to the Jeffreys-Lindley paradox and has several desirable properties, including consistency and asymptotic normality. Perez and Pericchi extended this approach to the calculation of Bayesian P-values, demonstrating how to use a training sample to compute a lower bound on the P-value that is independent of the choice of noninformative prior [Perez and Pericchi, 2014].

Despite these developments, the use of sequential methods to compute Bayesian P-values remains relatively uncommon in practice. One reason for this may be the perceived complexity and computational burden of these approaches relative to standard frequentist procedures. However, with the advent of modern Bayesian computing techniques such as Markov chain Monte Carlo (MCMC) methods, it is now feasible to perform sequential Bayesian approaches even for complex models and large datasets.

The goal of this paper is to provide a practical guide to computing sequential Bayesian P-values and to compare their performance with frequentist P-values and standard Bayesian P-values based on informative priors. We begin by reviewing the theoretical foundations of the sequential Bayesian approach and describing some commonly used methods for selecting training samples and combining the resulting P-values. We then present a few examples illustrating the application of these methods to simulated datasets. Through these examples, we aim to demonstrate the potential benefits of sequential Bayesian P-values, including their ability to provide robust, well-calibrated inferences that are less sensitive to the choice of prior than standard Bayesian methods. At the same time, we will discuss some of the challenges and
limitations of the sequential approach, such as the need to specify the size of the training sample and the potential for the results to depend on the order in which the data is processed.

Ultimately, we hope that this paper will help to demystify sequential Bayesian methods and encourage researchers to consider using them as a complementary approach alongside frequentist and standard Bayesian techniques. By doing so, we believe that researchers can obtain a more comprehensive and reliable assessment of the strength of evidence provided by their data, leading to more robust and reproducible scientific findings.
Bayesian P-values

Bayesian P-values offer an alternative approach to hypothesis testing that addresses some of the limitations of frequentist P-values. For example, Bayesian P-values have a more intuitive interpretation than frequentist P-values. They directly quantify the probability of the null hypothesis being true, given the test statistic is, at least, as extreme as the observed [Wagenmakers, 2007]. This makes it easier for researchers to read and understand the strength of evidence against the null hypothesis. Bayesian P-values are also less sensitive to sample size than frequentist P-values. The use of prior knowledge can help stabilize the results and reduce the influence of sample size on the conclusions [Berger and Sellke, 1987].

The preference for Bayesian P-values over classical P-values among statisticians stems from fundamental differences in their interpretations and underlying assumptions. The key distinction lies in the direct probabilistic assessment of the null hypothesis provided by Bayesian methods. Specifically, the Bayesian P-value computes \( P(H_0 \mid T \geq t_{obs}) \), which represents the probability of the null hypothesis being true given the observed data. This approach aligns more closely with the intuitive question researchers often seek to answer: "What is the probability that my hypothesis is true given the data I have observed?" [Gelman et al., 2013]. In contrast, the classical P-value calculates \( P(T \geq t_{obs} \mid H_0) \), which assumes the truth of the null hypothesis and quantifies the probability of observing data as extreme as or more extreme than what was actually observed under this assumption [Hubbard and Bayarri, 2003]. This fundamental difference in perspective (Bayesian methods considering the probability of the hypothesis given the data, versus frequentist methods considering the probability of the data given the hypothesis) underlies the growing preference for Bayesian approaches in many statistical applications [Kruschke and Liddell, 2018]. The Bayesian framework thus offers a more direct and intuitive way of addressing the central question of hypothesis testing, potentially leading to more nuanced and accurate
interpretations of scientific findings.

Bayesian P-values are closely related to Bayes factors, which are another commonly used measure of evidence in Bayesian hypothesis testing. Bayes factors quantify the relative likelihood of the data under the null and alternative hypothesis, and they can be used to calculate the Bayesian P-values [Wagenmakers, 2007]. One type of Bayes factor is called Intrinsic Bayes factors or IBFs. This method was proposed as a solution to the earlier mentioned Jeffreys-Lindley paradox and several other problems associated with improper priors [Berger and Pericchi, 1996]. IBFs are closely related to Bayesian P-values as it is obtained by calculating the probability of obtaining a Bayes factor as extreme as or more extreme than the one observed, based on the posterior distribution of the Bayes factor [Moreno et al., 1998].

**General Framework**

Following the Bayes theorem, the general framework we will utilize for calculating Bayesian P-values can be described using the formula below:

\[
P(H_0 \mid t_{obs}) = \frac{P(H_0) \times P(T \geq t_{obs} \mid H_0)}{P(T \geq t_{obs})}
\]

where \(P(H_0 \mid T \geq t_{obs})\) represents the posterior probability and is our value of interest. \(P(H_0)\) represents the prior probability which for the purpose of this paper, we will represent using the symbol \(\pi_0\). \(P(T \geq t_{obs} \mid H_0)\) represents the likelihood or classical P-value under the null hypothesis while \(P(T \geq t_{obs})\) represents the normalizing constant commonly known as the marginal likelihood or the prior predictive probability of the data. In this paper, we aim to investigate the performance of Bayesian P-values through simulation studies using the R statistical software.
Sample Size

The first step to estimating Bayesian P-values is choosing an adequate sample size \( n \). The choice of sample size can have a significant effect on the final results. Berger and Pericchi introduced the concept of intrinsic Bayes factors, which use a portion of the data to update an improper prior into a proper posterior [Berger and Pericchi, 1996]. They emphasized the importance of choosing a sample size that is large enough to provide a meaningful update to the prior distribution while avoiding the domination of the analysis by the data. This balance is crucial in maintaining the benefits of incorporating prior information while allowing the data to inform the posterior distribution. The impact of a sample size on Bayesian model selection has been a subject of research. [Moreno and Girón, 2008] investigated the performance of Bayesian procedures for variable selection in linear regression models. They found that for large samples, Bayesian procedures tend to outperform frequentist criteria in selecting the true model. This finding suggests that as sample size increases, the advantage of Bayesian methods becomes more pronounced. In 1995, Raftery discussed the use of Bayes factors in social research, highlighting the relationship between sample size and the strength of evidence provided by Bayes factors [Raftery, 1995]. He noted that with large samples, Bayes factors can provide strong evidence for or against a hypothesis, even when the effect size is small. This observation underscores the need for careful interpretation of results, especially in large-sample studies.

The concept of “sample size calibration” was introduced by De Santis (2004), who proposed a method to determine the sample size needed to achieve a desired level of evidence in Bayesian hypothesis testing [De Santis, 2004]. This approach provides a framework for researchers to plan studies with specific goals for the strength of evidence they wish to obtain. In the context of clinical trials, Spiegelhalter et al. (2004) discussed the role of sample size in Bayesian monitoring of clinical trials. They emphasized the flexibility of Bayesian methods in adapting to accumulating data and highlighted how sample size considerations can be integrated into the design and analysis of clinical trials [Spiegelhalter et al., 2004]. Johnson (2013) examined the relationship between Bayes factors and P-values, showing that for large samples,
there is often a discrepancy between the conclusions drawn from these two approaches [Johnson, 2013]. This finding further supports the observation of Moreno and Giron (2008) regarding the dominance of Bayesian procedures in large samples.

A prior effective sample size was explored by Morita et al. in 2008, who proposed methods to quantify the amount of information contained in prior distributions. This approach provides a way to compare the relative contributions of the prior and the data, which is particularly relevant when considering sample size in Bayesian analyses [Morita et al., 2008]. In a comprehensive review, Gelman et al. discussed various aspects of sample size determination in Bayesian analysis, including power analysis and design of experiments [Gelman et al., 2013]. They emphasized the importance of considering not only the sample size but also the informativeness of the data and the goals of the analysis. More recently, Schonbrodt and Wagenmakers introduced the concept of Bayes factor design analysis, which allows researchers to plan sample sizes for Bayesian hypothesis tests. This approach takes into account the desired strength of evidence and the uncertainty in effect size, providing a more nuanced approach to sample size determination in Bayesian studies [Schönbrodt and Wagenmakers, 2018].

In general, literature on sample size considerations in Bayesian inference highlights the complex interplay between prior information, data, and inferential goals. While larger samples generally provide more precise estimates and stronger evidence, the choice of sample size should be guided by the specific context of the study, the nature of the prior information, and the desired balance between prior influence and data-driven inference. As Bayesian methods continue to gain popularity, further research into optimal sample size determination for various Bayesian procedures will likely emerge, providing researchers with more refined tools for study design and analysis.

**Test Statistic**

The next step in estimating Bayesian P-values involves calculating the observed test statistic \( t_{obs} \). A test statistic is a numerical value calculated from sample data that is used to
make decisions about the null hypothesis in hypothesis testing. Test statistics play a crucial role in statistical inference serving as a bridge between observed data and the underlying population parameters of interest. These numerical summaries quantify the evidence against a null hypothesis and form the basis for hypothesis testing and the calculation of P-values. The concept of test statistics has its roots in the early 20th century, with the pioneering work of statisticians like Ronald Fisher, Jerzy Neyman, and Egon Pearson. Fisher introduced the idea of significance testing, which involves calculating a test statistic and comparing it to a known distribution to determine the probability of observing such an extreme value under the null hypothesis. This approach laid the groundwork for the development of various test statistics and their corresponding distributions [Fisher, 1925]. One of the most widely used test statistic is the t-statistic introduced by William Sealy Gosset (under the pseudonym “Student”) in 1908 [Student, 1908]. The t-statistic is particularly useful for comparing means when the population standard deviation is unknown and has found widespread application in fields ranging from psychology to economics. The t-statistics follows a t-distribution, which approaches the normal distribution as the sample size increases, making it a versatile tool for both small and large sample analyses.

Another fundamental test statistic is the chi-square statistic, developed by Karl Pearson in 1900 [Pearson, 1900]. The chi-square is used to assess the goodness-of-fit between observed data and expected frequencies under a hypothesized distribution. It has been extensively applied in categorical data analysis and has spawned numerous variations, including the likelihood ratio test and Pearson’s chi-square test. The F-statistic, introduced by Ronald Fisher in 1924, is central to the analysis of variance (ANOVA) and regression analysis [Fisher, 1924]. This test statistic allows for the comparison of variances between and within groups, enabling researchers to assess the significance of differences among multiple means simultaneously. The F-statistic follows an F-distribution and has been instrumental in fields such as experimental design and multivariate analysis. In recent years, there has been growing interest in developing robust test statistics that are less sensitive to violations of underlying assumptions. For example, Wilcox and
Keselman proposed a family of robust test statistics based on trimmed means and Winsorized variances, which perform well under conditions of non-normality and heteroscedasticity. These developments have expanded the toolkit available to researchers when dealing with real-world data that often deviates from ideal conditions [Wilcox and Keselman, 2003]. The advent of computational methods has also led to the development of resampling-based test statistics. Bootstrap methods, introduced by Efron in 1979, provide a powerful approach to estimating the sampling distribution of test statistics without relying on parametric assumptions [Efron, 1979]. Similarly, permutation tests offer a distribution-free alternative to traditional parametric tests, allowing for the construction of exact test statistics for a wide range of hypotheses [Ernst, 2004].

Just like P-values, the interpretation of test statistics has also been a subject of ongoing debate and research. The American Statistical Association’s statement on P-values marked a pivotal moment in statistical community’s approach to hypothesis testing. They highlighted the limitations of relying solely on test statistics and P-values for scientific inference [Wasserstein and Lazar, 2016]. The statement emphasized that P-values do not measure the probability that the studies hypothesis is true, nor do they measure the size of an effect or the importance of a result. This recognition has spurred a broader discussion about the limitations of null hypothesis significance testing (NHST) and the need for more comprehensive approaches to statistical inference. This has led to increased interest in alternative approaches, such as effect sizes and confidence intervals, to complement traditional hypothesis testing based on test statistics. The use of effect sizes, unlike P-values, quantify the magnitude of the difference between groups or the strength of a relationship between variables. Cohen popularized standardized effect sizes such as Cohen’s d for t-tests and r for correlations. These measures provide a more intuitive understanding of the practical significance of research findings [Cohen, 1988]. In 2013, Lakens further explored the calculation and interpretation of effect sizes in psychological research, emphasizing their importance in meta-analyses and power analyses. Confidence intervals have also received increased attention as a complement to or replacement for P-values [Lakens, 2013]. Cumming argued for a shift towards “the new statistics,” advocating for the routine reporting of
effect sizes and confidence intervals [Cumming, 2014]. Confidence intervals provide a range of plausible values for the population parameter, offering more information than a dichotomous reject/fail-to-reject decision based on a P-value. Moreover, they can be used to assess the precision of estimates and the practical significance of results. The concept of estimation thinking, as proposed by Cumming and Calin-Jageman in 2016, encourages researchers to focus on estimating the magnitude of effects rather than simply testing for their existence. This approach aligns with calls for more nuanced interpretations of research findings and recognition of uncertainty in scientific conclusions [Cumming and Calin-Jageman, 2016].

Another alternative gaining popularity and worth mentioning is the use of Bayesian methods. Kruschke and Liddell discussed the advantages of Bayesian estimation over NHST, highlighting its ability to quantify evidence in favor of both the null and alternative hypothesis [Kruschke and Liddell, 2018]. Bayesian approaches also allow for the incorporation of prior knowledge and the updating of beliefs based on new evidence, providing a more holistic framework for scientific inference. In conclusion, while test statistics and P-values continue to play a role in statistical Inference, there is a growing recognition of the need for more comprehensive approaches. The use of effect sizes, confidence intervals and Bayesian methods offer a richer toolbox for researchers to draw meaningful conclusions from their data. As the field continues to evolve, it is likely that we will see further integration of these approaches, leading to more nuanced and reliable scientific inferences.

Likelihood

There are different methods that can be used to compute the likelihood or frequentist P-value represented by $P(T \geq t_{obs} \mid H_0)$. We will discuss and utilize the empirical approach based on Monte Carlo procedures. The use of the Monte Carlo method for calculating empirical P-values have become increasingly prevalent in statistical analysis, particularly in fields dealing with complex datasets or non-standard distributions. This approach provides a more flexible alternative to traditional analytical methods, especially when theoretical distributions are difficult
to derive or when assumptions of parametric tests are violated. The fundamental principle of Monte Carlo methods in P-value calculation involves generating a large number of simulated datasets under the null hypothesis and comparing the test statistic from the observed data to the distribution of test statistics from the simulated datasets. This approach was pioneered by Barnard (1963), who introduced the concept of Monte Carlo tests as a way to obtain exact P-values for complex statistical problems [Barnard, 1963].

North, Curtis, and Sham made a significant contribution to this field in 2002 by highlighting a common misconception in the calculation of empirical P-values [North et al., 2002]. They proposed the formula:

\[
\frac{r + 1}{n + 1}
\]

as the correct method for calculating empirical P-values, where \( r \) is the number of simulated test statistics greater than or equal to the observed test statistic, and \( n \) is the total number of simulations. This formula offers an improvement over the more commonly used formula:

\[
\frac{r}{n}
\]

particularly when dealing with very small P-values. The importance of this correction becomes apparent when considering extreme cases.

As pointed out by Phipson and Smyth, the commonly used formula can produce a P-value of zero when the observed test statistic is more extreme than all simulated statistics [Phipson and Smyth, 2010]. This is problematic as it implies absolute certainty in rejecting the null hypothesis, which is never truly justified in statistical inference. The proposed formula by North, et al. (2002) shown above avoids this issue by ensuring that the minimum possible P-value is

\[
\frac{1}{n + 1}
\]

In 1997, Davidson and Hinkley provided a comprehensive treatment of bootstrap
methods, which share many similarities with Monte Carlo methods for P-value calculation. They emphasized the importance of considering the discreteness of empirical distribution when interpreting results, especially for small numbers of simulations [Davison and Hinkley, 1997].

The choice of the number of simulations $N$ in Monte Carlo P-value calculation is crucial. Boos and Zhang (2000) investigated the relationship between the number of simulations and the accuracy of the Monte Carlo P-values. They found that while increasing $n$ generally improves accuracy, the gains diminish rapidly beyond a certain point. They suggested that 10,000 simulations are often sufficient for most practical purposes [Boos and Zhang, 2000].

In the context of genetic studies, Besag and Clifford introduced sequential Monte Carlo Tests, which can reduce computational burden by stopping the simulation process early when it’s clear that the null hypothesis cannot be rejected instead of running a fixed number of simulations for each test [Besag and Clifford, 1991]. In their approach, simulations are generated one at a time. After each simulation, a decision is made whether to continue or stop based on the current results. If the observed test statistic is less extreme than a certain number of simulated statistics, the process stops, concluding that there is not enough evidence to reject the null hypothesis. This sequential approach can lead to substantial computational savings and more importantly, Besag and Clifford showed that their method maintains proper control of Type I error rates ensuring that the probability of falsely rejecting the true null hypothesis remains at the specified significance level.

Gandy extended the work on Monte Carlo P-values by developing methods for constructing confidence intervals for these P-values [Gandy, 2009]. This development allows researchers to quantify the uncertainty in their empirical P-value estimates, providing a more nuanced interpretation of results. More recently, Zhou and Wright proposed a method for efficient computation of empirical P-values in genome wide studies [Zhou and Wright, 2015]. Their approach, based on importance sampling, allows for accurate estimation of very small P-values with fewer simulations than traditional Monte Carlo methods. The use of Monte Carlo methods for P-value calculation has also been extended to more complex statistical procedures.
For instance, Racine and MacKinnon developed a fast, wild-bootstrap method for inference in quantile regression models, demonstrating the versatility of simulation-based approaches in modern statistical practice [Racine and MacKinnon, 2007].

Monte Carlo methods for empirical P-value calculation have evolved significantly since their introduction with important contributions addressing issues of accuracy, efficiency, and interpretation. The correction proposed by North et al. (2002) represents a crucial refinement in this field, ensuring more reliable P-value estimates, especially for highly significant results. As computational power continues to increase and statistical methods become more complex, it is likely that Monte Carlo approaches will play an increasingly important role in statistical inference across various disciplines.

**Marginal Likelihood**

At the bottom of the Bayesian formula lies the normalizing constant commonly referred to as the marginal likelihood or the prior predictive probability of the data [Wikipedia contributors, 2024]. The concept of marginal probability in the context of hypothesis testing has also gained significant attention in statistical literature, particularly as computational methods have advanced. A marginal likelihood refers to the probability of observing a test statistic as extreme as or more extreme than the observed one, regardless of whether the null hypothesis is true or false. This concept plays a significant role in normalizing posterior distributions and facilitating model comparison. It is fundamental in Bayesian approaches to hypothesis testing and has important implications for the interpretation of statistical evidence [Berger and Delampady, 1987]. Mathematically, for a model with parameters $\theta$ and data $x$, the calculation of the marginal likelihood is given by

$$P(X) = \int P(X \mid \theta) \cdot P(\theta) d\theta$$
where $P(X \mid \theta)$ is the likelihood and $P(\theta)$ is the prior distribution. These calculations often involve complex integrals that are analytically intractable. This has led to various other methods that have been proposed and used in estimating marginal probabilities.

One such method is the Monte Carlo Integration introduced by Metropolis and Ulam in 1949. It involves generating random samples from the prior distribution and using the sample to approximate the integral [Metropolis and Ulam, 1949]. A significant improvement over Monte Carlo Integration is Importance Sampling detailed by Kass and Raftery in 1995. This method involves sampling from a proposal distribution $q(\theta)$ that is closer to the posterior distribution and then reweighing the samples to correct for the difference [Kass and Raftery, 1995]. This approach can significantly reduce the variance of the estimate, especially when the likelihood is concentrated in a small region of the parameter space. For more complex models and large datasets, Sequential Monte Carlo (SMC) methods, also known as particle filters, can be more efficient. These methods, developed by Gordon et al., update the estimate of the marginal likelihood as new data points are observed [Gordon et al., 1993]. SMC methods are particularly useful in time series analysis. The idea from Sequential Monte Carlo methods combined with that from Importance sampling generates a method known as Annealed Importance sampling introduced by Neal in 2001. This approach gradually transitions from the prior to the posterior distribution through a sequence of intermediate distributions, potentially offering more reliable estimates for more complex models [Neal, 2001]. In 2006, John Skilling introduced the Nested Sampling algorithm which provides both an estimate of the marginal likelihood and samples from the posterior distribution. This method works by evolving a set of “live points” through nested contours of the likelihood function, offering a particularly elegant solution for problems where the likelihood function is computationally expensive to evaluate [Skilling, 2006].

All the methods presented above each come with their own challenges. For example, the efficiency of basic Monte Carlo methods decreases as the dimension of the parameter space increases. Also, rare events of interest occur with very low probability making standard Monte Carlo methods inefficient. Recent research has focused on addressing these limitations like the
Hamiltonian Monte Carlo method presented in 2017 by Betancourt to tackle High-dimensional problems in Monte Carlo methods [Betancourt, 2017]. As stated earlier, because the marginal distribution of the data tends to be a complicated integral, it is often extremely difficult to compute. Although, different alternative approaches have been proposed over the years to tackle this issue, however, because the parameter of interest does not depend on the marginal probability, in some cases, the Bayes’ theorem is often reduced to

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

While this expression is theoretically easier to compute, it is nonetheless difficult to simulate given that simulation of values from the posterior distribution is a primary goal of Bayesian analysis [Lynch, 2004]. In the context of a simulation study, the method that serves us the best in calculating marginal probability is the Monte Carlo procedure described earlier. A common approach to point hypothesis testing using simulation studies is generating substantial datasets under both the null hypothesis ($H_0: \mu = 0$) and the alternative hypothesis ($H_1: \mu > 0$), then calculating the corresponding test statistic for each dataset. Both results are combined into a single vector and the proportion of test statistics that are as extreme as or more extreme than the observed test statistic $t$ is calculated. This proportion serves as an estimate for the marginal probability.

**Properties**

When calculating Bayesian P-values, we can expect the results to have certain properties and interpretations that differ from frequentist P-values. One key property to consider is how the obtained Bayesian P-values are interpreted. As explained earlier, Bayesian P-values quantify the degree to which the observed data supports the null hypothesis relative to the alternative hypothesis, taking into account prior beliefs. Bayesian P-values provide a measure of strength of evidence against the null hypothesis rather than a fixed threshold for accepting or rejecting
the null hypothesis. The influence that prior probabilities have on Bayesian P-values is also an important property to consider.

Prior probabilities reflect genuine prior beliefs or knowledge about the parameter of interest. They can be informative (based on previous studies or expert knowledge) or non-informative like Jeffreys [Jeffreys, 1961] or Uniform priors. Different prior probabilities can lead to different Bayesian P-values even for the same observed data. Therefore, Bayesian P-values are sensitive to the choice of prior probabilities. The impact of prior probabilities diminishes as the sample size increases (as the data becomes more informative). With a sufficient amount of data, the influence of prior probabilities and distributions diminishes, and the Bayesian P-values become more data driven. However, the rate of convergence and stability depends on the specific problem, the prior choices, and the characteristics of the data.

It is important to note that a small Bayesian P-value indicates weak evidence against the null hypothesis and suggests a higher posterior probability for the alternative hypothesis. Conversely, large Bayesian P-values indicate strong evidence supporting the null hypothesis and suggests a lower posterior probability for the alternative hypothesis. As the sample size increases, Bayesian P-values tend to converge to a stable value assuming the prior probabilities and distributions are appropriate. We will observe these in more detail in the sections to follow.
Chapter 3

Research Objective

The interpretation and application of P-values in statistical inference have been subjects of ongoing debate and controversy in the scientific community. Traditional frequentist P-values, while widely used, have been criticized for their potential misinterpretation, over-reliance on arbitrary thresholds, and failure to incorporate prior knowledge [Wasserstein and Lazar, 2016]. These limitations have led to calls for more nuanced approaches to statistical inference. One such approach that has had increased interest as a promising alternative is Bayesian methods, including, but not limited to, Bayesian P-values. Bayesian P-values offer a fundamentally different approach to hypothesis testing, allowing for several theoretical advantages, including the ability to incorporate prior knowledge and provide a continuous measure of evidence. However, despite their theoretical advantages, the practical implementation and interpretation of Bayesian P-values present several challenges that have not yet been fully addressed in current literature.

There have been several proposed methods for calculating Bayesian P-values including the posterior predictive approach introduced by Rubin in 1984 [Rubin, 1984] and further developed by Gelman et al. in 1996 [Gelman et al., 1996]. This method involves comparing the observed test statistic to the distribution of test statistics generated from the posterior predictive distribution. This is one of the earliest and most straightforward approaches to calculating Bayesian P-values. In later years, Box proposed another approach known as the prior predictive P-value, which compares the observed test statistic to the distribution of test statistics generated from the prior predictive distribution [Box, 1980]. This method has the advantage of not using a sample data twice, a criticism sometimes leveled at posterior predictive P-values. Another approach is the partial posterior predictive P-value introduced by Bayarri and Berger in 2000. This approach serves as a compromise between the prior and the posterior predictive approaches as it aims to address some of the criticisms of both methods while retaining their advantages [Bayarri and Berger, 2000]. In more recent years, Evans and Moshonov proposed the relative
surprise P-value which measures the change from prior to posterior probability of the observed statistic [Evans and Moshonov, 2006]. This approach aligns closely with the Bayesian principle of measuring the change in beliefs based on observed data. Despite these advancements, each of these methods has its limitations.

Posterior predictive P-values have been criticized for their potential conservatism and lack of uniformity under the null hypothesis. Prior predictive P-values, while avoiding double use of the data, can be highly sensitive to the choice of prior. Partial posterior predictive and relative surprise P-values, while addressing some of these issues, introduce additional computational complexities. Furthermore, most of these methods are designed for batch processing of data, where all observations are available at once. In many practical situations, however, data is collected sequentially, and researchers may wish to update their inferences as new data becomes available. This scenario calls for a more flexible approach to calculating Bayesian P-values that can accommodate sequential data collection and analysis.

This research aims to advance the field of Bayesian inference by developing an innovative approach to calculating Bayesian P-values. Our proposed method synthesizes and extends key concepts from existing work on sequential Bayesian analysis, addressing current limitations and enhancing computational efficiency. By building upon established framework of Bayesian P-values, we introduce a novel technique that effectively incorporates sequential data updating, thereby bridging the gap between theoretical Bayesian principles and practical implementation in dynamic data environments. This study not only contributes to the methodological advancement of Bayesian statistics but also offers researchers across various disciplines a more flexible and robust tool for hypothesis testing in sequential data analysis scenarios.

**Sequential Bayesian Analysis**

The evolution of sequential Bayesian P-values is rooted in the broader development of sequential analysis and Bayesian inference. Abraham Wald introduced the concept of sequential analysis in the 1940s, primarily in a frequentist context [Wald, 1945]. However, the Bayesian
approach to sequential problems began to take shape in the 1950s and 1960s, with pioneering work by Dennis Lindley and Leonard Jimmie Savage [Lindley, 1957], who explored the application of Bayesian principles to sequentially arriving data.

The 1970s marked a significant shift towards Bayesian approaches in sequential decision-making and hypothesis testing. Adrian F.M. Smith’s 1975 paper on Bayesian sequential sampling [Smith, 1975] laid crucial groundwork for understanding how prior distributions could be updated with new data. This work set the stage for future developments in sequential Bayesian methods.

A major breakthrough came in the 1990s with the advent of Markov Chain Monte Carlo (MCMC) methods, revolutionizing Bayesian computation. The work of Alan Gelfand and Adrian F.M. Smith on Gibbs sampling [Gelfand and Smith, 1990] opened new avenues for implementing complex Bayesian models in sequential settings. This computational advancement was complemented by theoretical developments, notably the introduction of posterior predictive P-values by Meng in 1994 [Meng, 1994] and their further development by Gelman et al. in 1996 [Gelman et al., 1996]. These contributions provided a Bayesian alternative to frequentist P-values, forming the foundation for sequential Bayesian P-values.

The turn of the millennium saw continued advancements in sequential Bayesian methods. Researchers like Christian P. Robert [Robert, 2007] and Scott M. Lynch [Lynch, 2007] contributed significantly to the development of Bayesian posterior predictive checks, closely related to Bayesian P-values. During this period, the concept of sequential Bayesian updating in the context of P-values began to crystallize.

Sequential Bayesian P-values emerged as a novel approach to hypothesis testing, combining Bayesian inference principles with sequential analysis. This method involves using a portion of the data to update prior probabilities, with the remaining data used to calculate the Bayesian P-value [Berger and Pericchi, 1996]. The approach addresses longstanding issues in Bayesian inference, such as the Jeffreys-Lindley paradox [Lindley, 1957], which demonstrated how improper priors could lead to divergent results from frequentist P-values as sample sizes
increase. Kass and Raftery’s proposal in 1995 to use a “training sample” to convert improper priors into proper posterior distributions offered a solution to this paradox [Kass and Raftery, 1995], paving the way for more robust sequential Bayesian methods.

Building on this rich historical context and recent advancements, we now present our novel method for calculating sequential Bayesian P-values. Our approach, which we will demonstrate through a practical example, synthesizes key elements from existing work while introducing innovative techniques to enhance the efficiency and applicability of sequential Bayesian analysis.
Consider the following problem:

Let $X_1, \cdots, X_n$ be an i.i.d random sample from a normal distribution with unknown mean $\mu$ and variance 1. We wish to test the hypothesis $H_0: \mu = 0$ versus $H_1: \mu > 0$. Starting with a prior $\pi_0: P(H_0 = true)$, iteratively compute the posterior probabilities $\pi_1, \pi_2, \cdots, \pi_n$.

When analyzing multiple sequential samples, researchers often seek to determine whether the underlying distributions of these samples differ significantly from one another. To accomplish this, an appropriate test statistic must be selected and computed for each sample. In the context of the aforementioned problem, the null hypothesis posits that the difference between the means of the test statistics for any two consecutive samples (i.e., the $i^{th}$ and $(i + 1)^{th}$ samples) is equal to zero. Conversely, the alternative hypothesis suggests that this difference is greater than zero, indicating a potential shift in the distributions between the samples.

The primary objective of the question is to employ a Bayesian approach to calculating posterior P-values, which can then be used to assess the validity of the null hypothesis and determine if the sequential samples indeed follow approximately the same distribution. To tackle this problem, we will conduct a series of simulation studies using the R statistical software. These simulations will enable us to evaluate the performance and effectiveness of sequentially calculating Bayesian P-values under various scenarios.

Throughout the simulation process, we will leverage several of the methods and techniques discussed in the preceding sections to compute the posterior probabilities accurately. These methods may include, but are not limited to, specifying appropriate prior distributions, selecting suitable test statistics, and implementing efficient computational algorithms to handle the sequential nature of the data. By carefully designing and executing these simulation studies, we aim to gain valuable insights into the behavior and reliability of sequentially calculated Bayesian P-values when applied to the problem of comparing multiple sequential samples. The
results of these simulations will help inform researchers about the strengths and limitations of this approach and provide guidance on its practical application in real-world scenarios where assessing the similarity or difference between sequential distributions is of interest.

As mentioned earlier, the first step to generating Bayesian P-values is to determine an appropriate sample size that is large enough to provide a meaningful update to the prior distribution but small enough to avoid dominating the analysis. Moreno and Giron [Moreno and Girón, 2008] consider a sample data to be large when \( n = 100 \). Therefore, in our analysis, we will set \( n = 100 \) which should provide a balance between computational efficiency and the ability to update the prior distribution effectively. We will also provide results for \( n = 1000 \) samples to show the impact of sample size on Bayesian P-values. This range of sample sizes will allow us to observe how the Bayesian P-values behave under different data conditions and demonstrate the convergence properties as the sample size increases.

In our simulation study, we will utilize the \texttt{rnorm()} function available in R to generate random samples from a normal distribution \( N(\mu, 1) \). Here, our unknown parameter or statistic of interest \( \mu \) represents the mean of the sample data. We will also vary the value of the statistic \( \mu \) to represent different scenarios in order to investigate how it affects the resulting Bayesian P-values. The mean values we will use for comparison are:

- Null hypothesis \( (H_0) \) true : \( \mu = 0 \)
- Alternative hypothesis \( (H_1) \) with small effect : \( \mu = 0.1 \)
- Alternative hypothesis \( (H_1) \) with medium effect : \( \mu = 0.5 \)
- Alternative hypothesis \( (H_1) \) with large effect : \( \mu = 2 \)

For each scenario and sample size, we will generate \( N = 10,000 \) datasets and use the empirical Monte Carlo method to ensure stable estimates of the Bayesian P-values.

To analyze the difference between the observed data and the underlying parameter of interest \( \mu \), we will need to use a test statistic. Choosing a test statistic depends on the specific problem and the type of data being analyzed. Since we are dealing with a normally distributed
random sample in this case and testing hypothesis about the mean, the appropriate test statistic is the t-statistic. Nevertheless, for the purpose of research, we will introduce an alternative test statistic to compare the difference between the observed data and the underlying parameter of interest. The proposed t-statistic $T_i$ is calculated as follows:

$$T_i = \sum_{k=1}^{i} X_k, \quad k = 1, \ldots, n$$

(2)

The proposed t-statistic represents the sum of all $x_1, \ldots, x_i$ for each $T_i$. In the problem statement, $X_1, \ldots, X_n$ are assumed to be a random sample from a normal distribution $N(\mu, 1)$. The sum of normally distributed random variables follows a normal distribution as well. Specifically, if $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) random variables from $N(\mu, 1)$, then their sum $T_i$ follows a normal distribution with mean $i\mu$ and variance $i$.

Hence, the resulting distribution for $T_i \sim N(i\mu, i)$ where mean of the sum equals the sum of the means, and the variance of the sum equals the sum of the variances. The null hypothesis $H_0: \mu = 0$ and its alternative $H_0: \mu > 0$ concern the mean of the normal distribution. The proposed test statistic $T_i$, being the sum of the sample observations, is directly related to the sample mean. In fact, the sample mean can be expressed as

$$X_i = \frac{T_i}{i}$$

In the context of normal distributions with known variance, the sum of the sample observations is a sufficient statistic for the mean. This means that $T_i$ captures all the essential information about the parameter $\mu$ that can be extracted from the sample. Using a sufficient statistic ensures that no information is lost when computing the test statistic. The goal of sequential testing is to update the posterior probabilities as new data becomes available. This test statistic does that by summing $X_1, \ldots, X_i$ which accumulates the evidence from all the previous observations. The proposed test statistic is also easy to compute and update as new observations are added to the sample. This simplicity is advantageous when performing sequential testing and
updating the posterior probabilities iteratively.

Starting with mean \( \mu = 0 \) and sample size \( n = 100 \), we first need to initialize a the prior \( \pi_0: p(H_0 = \text{true}) \). To do this, we will use procedures from literature on FDR calculation. FDR stands for False Discovery Rate. In simple terms, FDR represents the tendency of an observed result to be false. When conducting multiple hypothesis tests simultaneously, the chance of making a Type I error (rejecting a null hypothesis when it is actually true) increases with the number of tests performed. FDR is designed to control or correct the expected proportion of false positive findings among all the rejected null hypothesis [Benjamini and Hochberg, 1995]. FDR is different from the Familywise Error Rate (FWER) which controls the probability of making at least one Type I error across all the hypothesis tests [Storey, 2002]. FDR is less conservative than FWER and is often controlled using methods such as the Bonferroni correction or the Benjamini-Hochberg correction [Dunn, 1961]. FDR corrections are statistical procedures that adjust the P-values or significance thresholds to control the FDR at a desired level. The most common FDR correction is the Benjamini-Hochberg procedure which works as follows:

1. Sort the P-values from the multiple hypothesis tests in ascending order.

2. Assign ranks to the sorted P-values (1 for the smallest, 2 for the second smallest, etc.).

3. Calculate the Benjamini-Hochberg critical value for each rank using the formula

\[
\frac{\text{rank} \times \text{desired FDR level}}{\text{total number of tests}}
\]

4. Compare each P-value to its corresponding Benjamini-Hochberg critical value.

5. Find the largest rank where the P-value is smaller than or equal to the critical value. Reject the null hypothesis for all tests with this rank or lower.

By controlling the FDR, researchers can be more confident that the significant findings they report are less likely to be false positives. A classical approach to controlling FDR using
software is by utilizing the \texttt{p.adjust()} function available in R. This base package function is nowadays often wrongly used to compute the estimated FDR. The function \texttt{p.adjust()} reports the adjusted P-values for FDR control not the estimated FDR. Since our aim is to estimate FDR, we will utilize the “FDRestimation” package provided in R by [Murray and Blume, 2022].

For $\mu = 0$, the estimated prior produced after FDR correction is 0.9998. Hence, we initialize our prior $\pi_0 = 0.9998$. The likelihood $P(T \geq t_{obs} \mid H_0)$ is estimated by generating $N = 10000$ datasets with sample size $i$ from the distribution $N(\mu = 0, \sigma^2 = 1)$. The test statistic $T_i$ is calculated using equation 2 and compared to the observed test statistic $t_{i,obs}$. Since this probability is conditioned on $H_0 : \mu = 0$ being true, we will set the mean to equal 0. To compare $T_i$ to $t_{i,obs}$, we will utilize a z-score. A z-score, also known as a standard score, is a statistical measure that quantifies how many standard deviations an individual data point is from the mean of a distribution [Fisher, 1925]. Using a z-score standardizes the test statistic making it easier to compare the observed values to the standard normal distribution $N(0, 1)$. We calculate the z-score for each test statistic and compare their differences. The z-score is used in the \texttt{pnorm()} function provided in base R to calculate the desired probability. The \texttt{pnorm()} function is used to calculate the cumulative distribution function (CDF) of a normal distribution. In other words, it returns the probability that a random variable following a normal distribution takes a value less than or equal to a given value. This probability can then be subtracted from 1 to obtain the desired probability $P(T_i \geq t_{i,obs} \mid H_0)$. The syntax for the \texttt{pnorm()} function as follows:

$$\texttt{pnorm}(q, \text{mean} = 0, \text{sd} = 1, \text{lower.tail} = \text{TRUE}, \cdots)$$

where $q$ is the value for which we want to calculate the probability (in our case is the z-score).

An alternative to taking the difference of the \texttt{pnorm()} value and 1 to obtain the desired probability is to set \texttt{lower.tail} = $F$ALSE which achieves the same result.

Next, we estimate the evidence $P(T_i \geq t_{i,obs})$ using the same method as the one above. This probability is not conditioned on any fixed value for the parameter of interest $\mu$, therefore,
we will generate 10000 datasets with the distribution \( N(\mu = \mu, sd = 1) \). Since currently the value of \( \mu = 0 \), we are bound to obtain the same probabilities as the previously estimated likelihood. Hence, when \( H_0 \) is true, we can generally expect both the likelihood and marginal likelihood (evidence) to be about the same probabilities. At this point, we have derived all the values needed to compute the posterior \( \pi_1 \). This posterior is then used as the prior probability in the next sequence to find the posterior \( \pi_2 \). This process continues until the final posterior \( \pi_n \) is calculated. In simulation, we arrived at \( \pi_n = 0.12 \). This represents the probability that the null is true given \( T_i \geq t_{i,obs} \). Figure 1 below is a plot that shows the estimated posterior probabilities.

![pi over iterations](image)

Figure 1: Null hypothesis \( H_0 \) true: \( \mu = 0 \) and \( n = 100 \)

The R-code that produces the above plot and all others presented in this paper can be
found in appendix A. We will now run the process again setting $\mu = 0.1$ and $n = 100$. The obtained Bayesian P-value $\pi_n$ is 0.04 (see figure 2 below). It is important to note that running this simulation different times with a fixed $\mu$ value will produce different results. However, we can expect these results not to vary significantly. Ultimately, in theory, we can expect the mean values that are closer to 0 to have higher probabilities than those farther away from 0. As $\mu$ gets larger, we expect the estimated Bayesian P-value to approach 0.

![Figure 2: Alternative hypothesis $H_1$ with $n = 100$ and small effect: $\mu = 0.1$](image)

The prior initially estimated for this hypothesis test in figure 2 is 0.912. Remember that $\pi_0 = P(H_0)$ represents the probability that the null hypothesis (mean or difference in means) for the generated samples is 0. Hence, we can expect the priors to be closer to 1 when the mean
is closer to 0 and vice versa. Notice how the prior decreases as the value for \( \mu \) increases farther away from 0. We will now perform a hypothesis test with a medium effect by setting \( \mu = 0.5 \) (see figure 3 below).

![Graph of pi over iterations](image)

Figure 3: Alternative hypothesis \( H_1 \) with \( n = 100 \) and medium effect: \( \mu = 0.5 \)

The result from this test indicates that the probability of getting a mean of 0 given the test statistic \( T_i \) is greater than the observed statistic \( t_{i,\text{obs}} \) is 0.02 with an initial prior of 0.618. Comparing the medium effect (figure 3) to the small effect (figure 2), we can see the difference in how fast each plot converges to 0. Hence the bigger the effect size, the faster the Bayesian P-value approaches 0. We will perform another hypothesis test with a larger effect size than the one in figure 3 for comparison. Setting \( \mu = 2 \), the initial prior is computed to equal
0.097 with a P-value of 0.018 (see figure 4 below).

Figure 4: Alternative hypothesis $H_1$ with $n = 100$ and large effect: $\mu = 2$

As noted earlier, computational efficiency is a crucial consideration when calculating Bayesian P-values, especially with larger sample sizes or more complex models. As the sample size increases, the computational time for calculating the P-values can grow substantially, making it important to optimize the code and utilize efficient computational techniques.

The r-code provided in appendix A demonstrates an optimized approach by leveraging parallel computing techniques. Parallel computing allows for the simultaneous execution of multiple tasks on different CPU cores, thereby reducing the overall computation time. In the provided code, the likelihood and marginal likelihood calculations are performed in parallel,
distributing the workload across the specified number of CPU cores. This parallelization can significantly speed up the computation, particularly when dealing with large sample sizes or extensive simulations. It runs the values of the likelihood and marginal likelihood parallel to each other using the specified number of CPU cores.

To investigate the effects of varying sample sizes on the estimation of Bayesian P-values, we will employ the same means used in the previous analysis but in this case, we will set \( n = 1000 \). Seeing that the priors used in the earlier computations are informative, we anticipate that they will not exhibit substantial variation in response to changes in sample size [Berger and Sellke, 1987]. This assumption is based on the notion that informative priors, which incorporate prior knowledge or beliefs about the parameters of interest, are less likely to be heavily influenced by the size of the sample compared to uninformative or weakly informed priors.

By setting \( n = 1000 \) and \( \mu = 0 \), the provided r-code (Appendix A) generates a prior of 0.9999 and P-value of 0.06 (see figure 5 below)
The Law of Large Numbers states that as the number of independent trials or sample sizes increases, the average of the results will converge towards the expected value or the true mean of the population [Dekking et al., 2005]. Following this law, when estimating Bayesian P-values through Monte Carlo simulations, we expect the results for the posterior distribution to converge towards the true Bayesian P-value. The convergence of the Monte Carlo estimate to the true Bayesian P-value has important implications for the reliability and interpretability of the results. When the sample size is sufficiently large, the estimated Bayesian P-value can be considered a good representation of the true value, allowing researchers to make more confident inferences and decisions based on the results. With an insufficient sample size, the
chain may not have enough time to explore the entire posterior distribution, leading to biased or unreliable estimates [Gelman et al., 2013]. Notice how the estimated P-values in figure 5 slowly decreases as the number of iterations get larger compared to figure 1. It may take more than 1000 iterations to achieve the true Bayesian P-value. However, the computational complexity and resource constraints can make it challenging to run larger iterations in practice. Nevertheless, as the number of iterations increases, the Monte Carlo error, which is the variability due to the random sampling process, decreases [Kruschke, 2015]. Looking at figures 2, 3 and 4, we can see that our method achieves convergence early in the iteration and stays around those values as the iterations increase. The figure below shows the results of the hypothesis test for $n = 1000$ and a small effect size $\mu = 0.1$. 
Figure 6: Alternative hypothesis $H_1$ with $n = 1000$ and small effect: $\mu = 0.1$

We can see here that the estimated P-value quickly drops and remains steady over large numbers of $n$ at 0.06. For medium ($\mu = 0.5$) and large ($\mu = 2$) effect sizes this remains the case. The estimated p-values are 0.007 and 0.007 respectively (see figures 7 and 8 below).
Figure 7: Alternative hypothesis $H_1$ with $n = 1000$ and medium effect: $\mu = 0.5$
Figure 8: Alternative hypothesis $H_1$ with $n = 1000$ and large effect: $\mu = 2$

Table 1 below displays the estimated Bayesian P-values for each test performed to aid in easier visual comparison. When $\mu = 0$, we notice that an increase in the sample size from 100 to 1000 reduces the estimated P-value drops by half. This means that for $\mu = 0$, it takes over a sample size of 100 to achieve the true Bayesian P-value. This is about the same for $\mu = 0.5$ and $\mu = 2$. 
Table 1: Simulated Bayesian P-values

<table>
<thead>
<tr>
<th></th>
<th>$\mu = 0$</th>
<th>$\mu = 0.1$</th>
<th>$\mu = 0.5$</th>
<th>$\mu = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td>0.12</td>
<td>0.04</td>
<td>0.02</td>
<td>0.018</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.06</td>
<td>0.06</td>
<td>0.007</td>
<td>0.007</td>
</tr>
</tbody>
</table>

When $\mu = 0.1$, we observe a slight increase in the estimate. To address potential outliers in P-value estimation, we applied the Monte Carlo averaging method discussed further in chapter 5. By implementing this method and increasing the sample size to $n = 1000$, we obtained an average Bayesian P-value of 0.04. This result suggests that our initial estimate at $n = 100$ closely approximates the true Bayesian P-value. The consistency between these estimates across different sample sizes reinforces the robustness of our approach and indicates that a sample size ($n = 100$) can be sufficient enough in estimating the true P-value given $\mu = 0.1$. 
Chapter 5
Advantages and Limitations

So far, we have explored the Bayesian approach to sequentially computing P-values. We have also compared the benefits of Bayesian methods over frequentist methods for calculating P-values. In this section, we will delve deeper into the advantages and limitations of the sequential Bayesian approach in computing P-values, comparing it with frequentist methods and other Bayesian methods.

It is important to note that the scope of this research, in alignment with a significant body of literature in the field, focuses primarily on the examination of one-sided hypothesis testing. The Bayesian framework offers considerable flexibility and potential for extension to more complex hypothesis structures beyond one-sided tests. This adaptability is a significant advantage of the Bayesian approach, particularly when addressing multifaceted research questions or when the directionality of effects is uncertain. Unlike frequentist methods, which often require specific tailoring for different types of hypotheses, Bayesian methods provide a unified probabilistic framework that can naturally accommodate two-sided hypotheses, interval hypotheses, or even multiple competing hypotheses simultaneously [Kruschke, 2015]. This extensibility is rooted in the fundamental Bayesian principle of updating prior beliefs with observed data, which remains consistent regardless of the hypothesis structure [Gelman et al., 2013]. Consequently, researchers can more easily transition from simple to complex hypotheses within the same methodological framework, potentially leading to more nuanced and comprehensive analyses of scientific phenomena [Wagenmakers et al., 2010]. The ability to handle diverse hypothesis structures not only enhances the versatility of statistical analyses but also aligns more closely with the often complex nature of research questions in many scientific disciplines.

The standard Bayesian approach to calculating P-values typically involves specifying a prior distribution for the parameters of interest, combining it with the likelihood function based on the observed data, and then computing the posterior distribution. Here, the prior distribution is
specified upfront and remains fixed throughout the analysis. In contrast, the sequential Bayesian approach uses Monte Carlo methods by updating the prior and corresponding P-values iteratively as new data becomes available. Instead of specifying a single prior distribution and computing the posterior distribution once, the sequential approach starts with an initial prior distribution and updates it sequentially with each new observation or batch of observations. The P-value is then calculated based on the updated posterior distribution at each step. The sequential Bayesian approach may be more sensitive to the choice of the initial prior distribution as it serves as the starting point for the sequential updating process. In contrast, the standard Bayesian approach relies on the specified prior distribution throughout the analysis.

When it comes to proportional computational efficiency, the sequential Bayesian approach can be more efficient than the standard approach especially when dealing with large datasets or streaming data [Wagenmakers et al., 2018]. By updating the P-values incrementally, the sequential approach avoids the need to recompute the posterior distribution from scratch each time new data arrives. The interpretability of both methods contains similar Bayesian perspectives expressing the probability of observing an effect at least as extreme as the one in the observed data given the null hypothesis. However, the sequential approach allows for a more dynamic and evolving interpretation when the P-values are updated with each new observation. The sequential approach is also particularly useful in scenarios where data is collected over time or when the sample size is not fixed in advance [Berger and Berry, 1988]. It is well-suited for online learning, real time monitoring, and adaptive decision making.

The standard Bayesian approach is more commonly used in situations where the entire dataset is available upfront, and a one-time analysis is sufficient. The choice between both methods depends on the specific research question, the nature of the data, and the computational resources available. Researchers should carefully consider the assumptions, limitations, and interpretations associated with each approach and select the one that aligns best with their research objectives and the characteristics of the data at hand.

When computing Bayesian P-values through simulation studies, we may run into
the issue of outlier effects. Outliers can occur in the simulated test statistics or the resulting P-values themselves and are mostly prevalent for small sample sizes. These outliers can affect the interpretation and reliability of the results. During simulation process, random samples are generated from the assumed null distribution to calculate the test statistics. If the random number generator produces extreme values or if the assumed null distribution has heavy tails, some of the simulated test statistics may be unusually large or small compared to most of the simulated values. These outliers can inflate the variance of the simulated test statistics and shift the distribution, affecting the calculated Bayesian P-values. In simulation, outliers can lead to overly conservative or overly liberal P-values. To mitigate the impact of outliers, several strategies like sensitivity analysis can be employed.

Conducting sensitivity analyses involves varying the prior probabilities, the assumed null distribution, or the simulation parameters. This can help assess the robustness of the Bayesian P-values to different assumptions and identify potential outliers. A more simplistic approach involves running the simulation multiple times with the same parameters and taking the average of the results. This can help mitigate the impact of outliers to some extent. This approach is known as “Monte Carlo averaging”. In the book “Introducing Markov chain Monte Carlo”, Gilks et al. discusses the use of averaging over multiple simulations to obtain a more suitable and reliable estimate in MCMC calculations [Gilks et al., 1996].

Several other articles have also mentioned or suggested the use of averaging over replications in Markov Chain Monte Carlo Methods. However, they do not specifically focus on averaging over replications for computing Bayesian P-values. While this is the case, the general principle of using multiple simulations and averaging the results to mitigate the impact of outliers and obtain a more stable estimate is well-established in Bayesian literature. Although averaging over replications is a useful strategy, it may not completely eliminate the influence of extreme outliers. Using other strategies such as sensitivity analysis in conjunction with the Monte Carlo averaging method is the most common approach to tackling outliers.
Chapter 6

Conclusion

In this paper, we have explored the sequential Bayesian approach to calculating P-values and compared its performance with traditional frequentist methods. Through simulation studies using the R statistical software, we have demonstrated the advantages of the sequential Bayesian approach in providing a more nuanced and informative assessment of the strength of evidence against the null hypothesis.

The sequential Bayesian approach offers several key benefits over frequentist P-values. By incorporating prior information and updating the P-values as new data becomes available, this approach allows for a more dynamic and adaptive hypothesis testing framework. The ability to specify informative priors provides a way to incorporate expert knowledge and previous research findings into the analysis, leading to more accurate and reliable results.

Our simulation studies have shown that the sequential Bayesian approach converges to stable P-values as the sample size increases, providing a consistent measure of evidence against the null hypothesis. We have also observed that the choice of prior probabilities and the effect size play a crucial role in determining the behavior of the Bayesian P-values. As the effect size moves further away from the null value, the Bayesian P-values tend to approach zero more quickly, indicating stronger evidence against the null hypothesis.

Furthermore, we have highlighted the computational advantages of the sequential Bayesian approach, particularly when dealing with large sample sizes or complex models. By leveraging parallel computing techniques and efficient algorithms, the proposed method can handle computationally intensive tasks and provide faster results compared to traditional approaches. However, it is important to acknowledge the limitations of the sequential Bayesian approach. The choice of prior probabilities, specification of the test statistic, and possible outliers in generated data can have a significant impact on the results. Researchers must carefully consider these choices and conduct sensitivity analyses to assess the robustness of their findings.
Additionally, the interpretation of Bayesian P-values may require a shift in thinking compared to frequentist P-values, as they represent the probability of the null hypothesis being true given the observed data and prior information.

Despite these limitations, the sequential Bayesian approach offers a promising alternative to frequentist methods for hypothesis testing. By providing a more comprehensive and flexible framework for assessing the strength of evidence, this approach has the potential to improve the reliability and reproducibility of scientific findings. Future research could explore the application of the sequential Bayesian approach to a wider range of statistical problems and models. Extensions to handle more complex data structures, such as hierarchical or correlated data, could further enhance the utility of this method. Additionally, the development of user-friendly software packages and tutorials could facilitate the adoption of the sequential Bayesian approach by researchers across various fields.

In conclusion, the sequential Bayesian approach to calculating P-values presents a powerful and informative alternative to traditional frequentist methods. By incorporating prior information, updating P-values sequentially, and providing a more nuanced assessment of the strength of evidence, this approach has the potential to revolutionize hypothesis testing and contribute to more robust and reliable scientific discoveries.
References


[Pearson, 1900] 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.


Appendix A

R Code

# Clear environment
rm(list=ls())

library(FDRestimation)
library(parallel)

# Create a cluster with the desired number of cores
num.cores <- detectCores() - 1
cl <- makeCluster(num.cores)

clusterEvalQ(cl, library(parallel))

n <- 100
mu <- 0
N <- 10000

# Pre-compute standard normals for reuse
std.normals <- matrix(rnorm(n * N), nrow = N, ncol = n)

# Function to generate observations and test statistics
generate_data <- function(i) {

}
Xi <- rnorm(i, mean = mu, sd = 1)
Ti <- rnorm(i, mean = i*mu, sd = sqrt(i))
list(Xi = Xi, Ti = Ti)

# Function to compute empirical S and likelihood
compute_statistics <- function(i) {
  data <- generate_data(i)
  Xi <- data$Xi
  Ti <- data$Ti

  t_obs <- sum(Xi)
  mean_Xi <- mean(Xi)
  var_Xi <- var(Xi)

  results <- mclapply(1:N, function(j) {
    Til <- Ti + std_normals[j, 1:i]  # Reuse pre-computed standard normals
    t <- (sum(Til) - t_obs - (mean(Til) - mean_Xi)) / sqrt(var(Til) + var_Xi)
    t0 <- (sum(Til) - t_obs) / sqrt(var(Til) + var_Xi)
    c(pnorm(t, lower.tail = FALSE), pnorm(t0, lower.tail = FALSE))
  })

  results_matrix <- do.call(rbind, results)
  c(sum(results_matrix[,1]), sum(results_matrix[,2]))
}
clusterExport(cl, c("std_normals", "generate_data", "mu", "N"))

# Compute si and pi in parallel
results <- parLapply(cl, 2:n, compute_statistics)
results_matrix <- do.call(rbind, results)

si <- (results_matrix[,1] + 1) / (N + 1)
p <- (results_matrix[,2] + 1) / (N + 1)

# Compute pi
pi <- numeric(n)

init_prior <- parLapply(cl, 1:N, function(i) {
    test = pnorm(abs(mean(rnorm(n, mu, 1))), lower.tail = F)
    return(test)
})

# Stop the cluster
stopCluster(cl)

# clear unused memory
gc()

prior <- suppressWarnings(mean(p.fdr(unlist(init_prior))$fdrs))
print(prior)
for (i in 1:n) {
    pi[i] <- (prior * p[i]) / si[i]
    prior <- pi[i]
}

plot(pi, type="l", main="pi over iterations", xlab="Iteration", ylab="pi")

# Get the last iteration
pi_n = pi[n-1]