Title
Bayesian Assurance and Sample Size Determination for Experimental Studies

Author
Pan, Jane

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Bayesian Assurance and Sample Size Determination for Experimental Studies

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Jane Aibo Pan

2022

ABSTRACT OF THE DISSERTATION

Bayesian Assurance and Sample Size Determination for Experimental Studies

by

Jane Aibo Pan
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2022
Professor Sudipto Banerjee, Chair

Determining the sample size to meet the inferential objectives of a study is of central importance in experimental design. There is an extensive collection of methods addressing this problem from diverse perspectives. The Bayesian paradigm, in particular, has attracted noticeable attention and includes different perspectives for sample size determination. While traditional Bayesian methods formulate sample size determination as a decision problem that optimizes a given utility functions (Lindley, 1997), practical experimental settings may require a more flexible approach based upon simulating analysis and design objectives (see, e.g., O'Hagan and Stevens, 2001). Building upon the latter approach, we devise a general Bayesian framework for simulation-based sample size determination using Bayesian assurance that can be easily implemented on modest computing architectures. We qualify the need for different priors for the design and analysis stage, working primarily in the context of conjugate Bayesian linear regression models, where we consider known and unknown variances. We also compare the assurance to a utility-based approach that involves the specification of objective functions to determine the rate of correct classification (Inoue, Berry, and Parmigiani, 2005). Throughout, we draw parallels with frequentist solutions, which arise

as special cases, and alternate Bayesian approaches with an emphasis on how the numerical results from existing methods arise as special cases in our framework.

We further extend our conjugate linear model's capabilities to encompass the multiple testing framework, where the assurance is now characterized by conditions placed on the Bayesian false discovery rate (FDR). Under this framework, we investigate the effects of multiple comparison adjustments on assurance and sample size determination. Adjustments include enforcing different assigned threshold values for the Bayesian FDR and conditions related to the credible interval condition, and varying the number of pairwise hypothesis tests being conducted. Of particular interest is observing how the number of pairwise tests being conducted affects the assurance under fixed constraints placed on the Bayesian FDR as defined in Müller et al., 2004. We assess how our proposed model performs in commonplace large-scale problems, specifically microarray data. Our methodology is implemented in a study of mammary cancer in the rat, where four distinct patterns of expression are provided. Future tasks involve assessing how our method performs when comparing more than two subgroups and enforcing objective ways of choosing optimal threshold values.

This dissertation captures the vast applicability of the two-stage framework, offering a robust Bayesian approach for sample size determination equipped for addressing a wide selection of problems taking place both within and outside clinical trial settings. There is broad potential for growth and development in the methods introduced, with numerous routes available for future exploration.

The dissertation of Jane Aibo Pan is approved.

Donatello Telesca

Jack Needleman

Catherine Crespi-Chun

Sudipto Banerjee, Committee Chair

University of California, Los Angeles

2022

To my loving family

TABLE OF CONTENTS

1 Introduction **1**

1.1 Literature Review and Dissertation Aims 1

1.2 Dissertation Structure 3

2 Background **6**

2.1 Conjugate Bayesian Linear Regression 6

2.2 Limitations for a single prior 8

2.3 Bayesian Assurance Using Design and Analysis Priors 9

2.3.1 Known Variance 9

2.3.2 Unknown Variance 11

3 Two-Stage Paradigm Applications **13**

3.1 Sample Size Determination in Cost-Effectiveness Setting 13

3.2 Design for Cost-Effectiveness Analysis 15

3.2.1 Simulation Results in the Known Variance Case 15

3.2.2 Simulation Results in Unknown Variance Case 16

3.3 Assurance Computation in the Longitudinal Setting 18

3.3.1 Longitudinal Example 20

3.4 Sample Size Determination Using Precision-Based Conditions 21

3.5 Sample Size Determination in a Beta-Binomial Setting 23

3.6 A Utility-Based Approach in the Bayesian Setting 25

3.6.1 Comparing the Rate of Correct Classification to Assurance 27

3.6.2 Relationship to Frequentist Setting 30
3.7 Discussion 32

4 bayesassurance R Package 33
4.1 Introduction 33
4.2 Closed-form Solution of Assurance 36
4.2.1 Special Case: Convergence with the Frequentist Setting 40
4.3 Simulation-Based Functions Using Conjugate Linear Models 44
4.3.1 Assurance Computation with Known Variance 44
4.3.2 Assurance Computation with Unknown Variance 49
4.3.3 Assurance Computation in the Longitudinal Setting 50
4.3.4 Assurance Computation for Unbalanced Study Designs 55
4.4 Visualization Features and Useful Tools 60
4.4.1 Overlapping Power and Assurance Plots 60
4.4.2 Design Matrix Generators 64
4.5 Discussion 70

5 Multiple Comparison Problems using Bayesian FDR Conditions 71
5.1 Introduction 71
5.2 Methodology 73
5.2.1 Pairwise Hypothesis Testing in Conjugate Linear Model Setting Using Design and Analysis Priors 74
5.2.2 Bayesian False Discovery Rate for Multiple Testing 77
5.2.3 Bayesian Assurance and Sample Size Determination for Multiple Testing 79
5.3 Simulation 81

5.3.1 Simulation Results 82
5.4 Case Application: Microarray Gene Expression 88
5.4.1 Case Application Methodology 88
5.4.2 Case Application Results 90
5.5 Discussion 92

6 Discussion 95
6.1 Summary and Takeaways 95
6.2 Limitations and Future Direction 97

A Appendix A: Algorithms 99
A.1 Algorithm 1 99
A.2 Algorithm 2 101
A.3 Algorithm 3 103
A.4 Algorithm 4 104
A.5 Algorithm 5 105
A.6 Algorithm 6 106
A.7 Algorithm 7 107

B Appendix B: Derivations and Specifications 108
B.1 Special Case Explanation in Section 2.3.1 108
B.2 Design Prior Specifications in O'Hagan and Stevens (2001) in Section 3.1 . . 109
B.3 Precision-Based Analysis Stage Objective in Section 3.4 110
B.4 Simulation Results under Precision-Based Conditions in Section 3.4 111
B.5 Convergence to Frequentist Setting in Section 3.4 112

B.6 Relation to Frequentist Setting in Beta-Binomial Setting in Section 3.5 . . . 113

B.7 Deriving Goal Function Threshold in Section 3.6 116

B.8 Utility Function Example from Reference Paper in Section 3.6 118

LIST OF FIGURES

3.1 Assurance curve based on results of algorithm corresponding to unknown variance. 17

3.2 Estimated assurance points for longitudinal example. 21

3.3 Assurance behavior in relation to rate of correct classification with a weakly assigned analysis prior ($n_a = 0$), a strongly assigned design prior ($n_d \to \infty$), fixed critical difference $\delta = 0.10$, variance $\sigma^2 = 1$, and utility $K = 1$. 28

3.4 Assurance behavior in relation to rate of correct classification for fixed critical difference $\delta = 0.10$, variance $\sigma^2 = 1$, and various precision settings within the analysis and design stages of the assurance framework. 29

3.5 Solid curve shows power and r^* values that yield equal sample sizes in the Bayesian and frequentist settings assuming the critical difference is $\delta = 0.1$. 31

4.1 Resulting assurance plot with specific points passed in marked in red. 39

4.2 Resulting power and assurance curve when weak analysis priors and strong design priors and enforced. 43

4.3 Estimated assurance points for longitudinal example. 54

4.4 Contour map of assurance values with varying sample sizes `n1`and `n2`. 58

4.5 Contour map of assurance values in cost-effectiveness application. 61

4.6 Power curve with exact and simulated assurance points for weak analysis prior and strong design prior. 62

4.7 Power curve with exact and simulated assurance points for weak analysis and design priors. 64

5.1 Assurance curves for various posterior credible interval thresholds, $t = 0.6, 0.7, 0.8$, and Bayesian FDR thresholds $\xi = 0.05, 0.10$. 83

5.2 Assurance curves for fixed FDR criteria that considers different posterior credible interval criteria and Bonferroni adjustments. 84

5.3 Assurance curves for fixed FDR criteria that considers different posterior credible interval criteria and Bonferroni adjustments. 85

5.4 Individual behavioral trends exhibited by estimated assurance (left) and Bayesian FDR (right) as the number of comparisons increases and the sample size per subgroup is varied. A 0.01 Bayesian FDR threshold is used for estimating the assurance values displayed on the left. Different colors signify different subgroup sample sizes, n. 90

5.5 Estimated Bayesian FDR as a function of sample size. Different colors signify different numbers of comparisons being conducted. 91

5.6 Assurance as a function of estimated Bayesian FDR. Different colors signify different numbers of comparisons being conducted. 92

B.1 Overlay of simulated results and frequentist results given a weak analysis prior such that $n_a \to 0$. 111

B.2 Overlay of simulated assurance results using posterior credible intervals and frequentist power results based on regular confidence intervals. 114

B.3 Utility curves using specifications provided by Inoue, Berry, and Parmigiani, 2005. 119

LIST OF TABLES

3.1 Recorded simulation results from Bayesian assurance algorithm with varying number of iterations. 16

3.2 Recorded results from the Bayesian assurance algorithm for unknown variance with varying number of iterations. 17

3.3 Recorded simulation results from Bayesian assurance algorithm with varying number of iterations. 18

3.4 Select set of overlapping sample sizes within the Bayesian and frequentist settings corresponding to varying values of r^* and β. 31

4.1 Overview of the functions available for use within the package. 35

5.1 Estimated assurance values under different probability thresholds (t) denoting the posterior probability of 0 not falling within the respective credible intervals, and fixed Bayesian FDR thresholds. Assurances increase with larger assigned thresholds. 82

5.2 Estimated assurance values under different probability thresholds (t)corresponding to the posterior probability of 0 not falling within the respective credible intervals, and whether or not the Bonferroni adjustment was enforced. The following assurance estimates are based on the analysis objective that the estimated Bayesian FDR falls below 0.05. Overall, the original results are not too different in comparison to Bonferroni-corrected estimates, but it is interesting to note that the original assurance estimates tend to be larger than the Bonferroni-corrected estimates when $n < 30$ and converge in behavior after, as can visually be seen in Figure 5.2. 86

5.3 Estimated assurance values under different probability thresholds (t) corresponding to the posterior probability of 0 not falling within the respective credible intervals, and whether or not the Bonferroni adjustment was enforced. The following assurance estimates are based on the analysis objective that the estimated Bayesian FDR falls below 0.10. Similar to the case when a restriction of $FDR < 0.05$ was implemented, the original assurance estimates tend to be larger than the Bonferroni-corrected estimates when $n < 30$ and converge in behavior after, as can visually be seen in Figure 5.3. 87

5.4 Estimated assurance values corresponding to different fixed Bayesian FDR threshold values and number of pairwise hypothesis tests. 93

ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser, Dr. Sudipto Banerjee, for his unwavering guidance and support during this long, arduous, but incredibly rewarding journey. Graduate school becomes significantly more bearable when your adviser brings out your potential and dedicates honest time and effort to help mold you into a better thinker, writer, and researcher. Thank you for believing in me even when I doubted myself. Your logic-driven advice and encouragement enabled me to look ahead and push forward. In addition, I'd also like to thank Dr. Kate Crespi, Dr. Donatello Telesca, and Dr. Jack Needleman, for taking the time out of their busy schedules to serve as my committee members.

There were several factors that drove me to pursue a PhD. I would like to thank the Meyerhoff Scholarship Program at UMBC for financially supporting my undergraduate career and providing me with unrivaled mentorship for my terminal degree in STEM. A special thanks to the Meyerhoff staff: Keith Harmon, Mitsue Wiggs, Sharon Johnson, Alicia Hall, and Michael Goodwyn for guiding both Pan siblings to their academic goals. My entire family is forever grateful for all you've done. I would also like to thank my Dad, Dr. Qiyuan Pan, and my brother, Dr. Zhigang Pan. You have been inspirations to me since my youth, and I often consider you both as my first mentors. Thank you Mom for being my biggest supporter. Our evening phone calls got me through some particularly rough times while living alone in LA. I would also like to thank Dr. Tom Belin for being the first UCLA faculty member to contact me directly. I distinctly recall our phone call while I was finishing up my final semester at college. I would like you to know that it was ultimately your warmth and sincerity that convinced me to join this program. You have been a great resource to me since then as my academic adviser, and I sincerely thank you for that.

I would like to thank all my peers (fellow cohort members, juniors, and alumni) for all the advice, laughs and memories. Thank you to our department's Student Affairs Officer, Roxy, for always looking out for me and managing all program-related logistics. Thank

you to my summer internship mentors: Mat and Matilde from FDA, Veronica, Brad, and Jianchang from Takeda, and Tony, Qing, and Junyi from Amgen, for granting me such amazing opportunities and helping shape my career goals. I've learned so much from you all. Thank you Fin for being my primary source of strength, motivation, and comfort. I am reminded everyday of how lucky I am to have you by my side. Thank you everyone. This journey wouldn't have been nearly as enjoyable without you all being a part of it.

VITA

2012 - 2016 Meyerhoff Scholarship Program, UMBC.

2016 B.S. (Mathematics) and B.S. (Statistics), UMBC.

2016 - 2017 Teaching Assistant, Department of Biostatistics, UCLA.

2017 - 2019 Statistical Analyst Intern, Doctor Evidence, LLC. Assisted with client consulting by helping answer medical and statistics-based questions through comprehensive literature searches.

2019 - 2021 Graduate Student Researcher, Semel Institute for Neuroscience and Human Behavior, UCLA. Provided assistance in analyzing youth-level sociodemographic disparity in special education and under-resourced youth with Autism Spectrum Disorder (ASD) within LAUSD.

2020 ORISE Summer Research Fellow, FDA. Identified drawbacks of the three-level hierarchical mixture model proposed by Scott and Donald Berry (2004). Configured an existing Bayesian hierarchical mixture model with additional exchangeability.

2021 SQS Summer Research Intern, Takeda Pharmaceuticals. Explored BART's performance in reltion to other well-established subgroup borrowing methods under varying scenarios of subgroup response and covariate heterogeneity.

2021 Design & Innovation Intern, Amgen Inc. Constructed R-based toolkit with emphasis on causal inference methods.

PUBLICATIONS

Accepted Manuscripts

Jane Pan et al. (2022). "Bayesian Additive Regression Trees (BART) with covariate adjusted borrowing in subgroup analyses". In: *Journal of Biopharmaceutical Statistics* 32 (4). URL: https://doi.org/10.1080/10543406.2022.2089160

Submitted for Review

Jane Pan and Sudipto Banerjee (2021a). "A Unifying Bayesian Approach for Sample Size Determination Using Design and Analysis Priors". In: *ArXiv e-prints*. arXiv: 2112.03509

Jane Pan and Sudipto Banerjee (2021b). "bayesassurance: An R package for calculating sample size and Bayesian assurance". In: *ArXiv e-prints*. arXiv: 2203.15154

In Preparation

Jane Pan and Sudipto Banerjee (n.d.). "Multiple Testing in the Bayesian Framework". In preparation

CHAPTER 1

Introduction

1.1 Literature Review and Dissertation Aims

Sample size determination (SSD) comprises a crucial aspect of statistical study designs. This dissertation considers a practical Bayesian approach and gleans insights from a flexible and analytically tractable framework. There is, by now, a substantial literature in classical and Bayesian settings. Classical sample size calculations have been treated in depth in texts such as Kraemer and Thiemann, 1987, Cohen, 1988, and Desu and Raghavarao, 1990, while extensions to linear and generalized linear models have been addressed in Self and Mauritsen, 1988, Self, Mauritsen, and O'Hara, 1992, Muller et al., 1992 and Liu and Liang, 1997. Bayesian settings have also received substantial attention towards sample size determination, including the use of distinct prior distributions for conducting Bayesian analysis and data generation (Brutti, Santis, and Gubbiotti, 2014; O'Hagan and Stevens, 2001; Sahu and Smith, 2006), the derivation of asymptotic estimations that provide closed-form expressions for Bayesian sample size (Clarke and Yuan, 2006), and specification of conditions that are characterized by historical data (Santis, 2007) and posterior quantiles (Santis, 2006). *The Statistician* (vol. 46, issue 2, 1997) includes a number of articles from different perspectives regarding Bayesian SSD (see, e.g., the articles by Adcock, 1997; Joseph, Berger, and Belisle, 1997; Lindley, 1997; Pham-Gia, 1997; Weiss, 2002). Within the Bayesian setting itself, there have been efforts to distinguish between a formal utility approach (Berger, 1985; Chaloner and Verdinelli, 1995; Inoue, Berry, and Parmigiani, 2005; Lindley, 1997; Müller and Parmigiani, 1995; Parmigiani, 2002; Raiffa and Schlaifer, 1961) and approaches that attempt to

determine sample size based upon some criterion of analysis or model performance (Gelfand and Wang, 2002; O'Hagan and Stevens, 2001; Rahme, Joseph, and Gyorkos, 2000). Other proposed solutions adopt a more tailored approach that are specific to given settings. For example, Ibrahim et al., 2012 specifically targets Bayesian meta-experimental design using survival regression models; Reyes and Ghosh, 2013 propose a framework based on Bayesian average errors capable of simultaneously controlling for Type I and Type II errors, while Joseph and Belisle, 1997a; Joseph, Berger, and Belisle, 1997; Joseph, Wolfson, and Berger, 1995b and Cao, Lee, and Alber, 2014 rely on lengths of posterior credible intervals to gauge their sample size estimates. Bayesian treatments specific to clinical trials can be found in Spiegelhalter, Freedman, and Parmar, 1993, Parmigiani, 2002, Berry, 2006, Berry et al., 2010, Lee and Zelen, 2000, and Lee and Chu, 2012.

Regardless of approach, all of the cited articles above are based around some well-defined objective that is desired in the analysis stage. The design of the study, therefore, should assure us that the analysis objective is met with a certain probability. In the Bayesian setting, we do not need to focus on specifying a null hypothesis. Instead, we assess the tenability of a hypothesis based upon the data we observe. A joint probability model is constructed for the parameters and the data using a prior distribution for the parameters and the likelihood function of the data conditional on the parameters. Inference proceeds from the posterior distribution of the parameters given the data. In the design stage we have not observed the data. Therefore, we formulate a data generating mechanism and, subsequently, consider the posterior distribution given the realized data to evaluate the tenability of a hypothesis. We then use the probability law associated with the data generating mechanism to assign a degree of assurance to our analysis objective.

This dissertation explores Bayesian assurance and subsequent SSD in the context of conjugate Bayesian linear regression. Of particular emphasis will be the data generating mechanism and motivation behind quantifying separate prior beliefs at the design and analysis stage of clinical trials (O'Hagan and Stevens, 2001). We will show how this framework

also produces classical sample size determination as a special case in terms of the numerical answers. The final stage of this dissertation further extends the model's capabilities to address multiple comparison problems, taking into account commonly cited multiplicity adjustments, including the Bonferroni correction and a Bayesian-based definition of the false discovery rate (Müller et al., 2004). The framework we develop here is built upon this simple idea that embodies the main takeaway of this dissertation: A clear analysis objective and proper sampling execution are all that is needed to provide us with the necessary sample size and corresponding assurance.

1.2 Dissertation Structure

The structure of this dissertation is organized around a series of individual manuscripts corresponding to projects that were worked on over the course of the graduate program. Projects are closely related and all depend on the two-stage paradigm we present early on in the dissertation.

In Chapter 2, we introduce the Bayesian assurance for SSD within a conjugate linear model framework. Here, we establish the general setup of the study design that we center our discussion around and specify the posterior distribution from which we draw inference from. We also discuss the limitations that occur from assigning a single prior to our model, motivating the need to assign two sets of priors to purposefully address two distinct study objectives. The two-stage design is then used as a template to formulate steps for estimating the assurance within cases of known and unknown variances.

In Chapter 3, we assess the performance of our two-stage model across different applications and hypothesis testing scenarios. We open the chapter with a cost-effectiveness application referenced in O'Hagan and Stevens, 2001, where we specify a hypothesis test that aligns with our linear model framework. We then conduct a simulation study to estimate the assurance values for a select set of sample sizes. Results are subsequently compared to those

reported by O'Hagan and Stevens, 2001 for evaluation. The remaining parts of this chapter investigate assurance and sample size from a utility-based approach that defines appropriate goal functions used to measure the rate of correct classification. In each case that is explored in this chapter, we make sure to tie our Bayesian-based methods back to the frequentist paradigm and identify scenarios where the two settings overlap in behaviors.

In Chapter 4, we present the **bayesassurance** R package, whose primary usage involves calculating the Bayesian assurance and sample size under the different settings outlined in Chapter 3. Building and launching the R package constitutes a large component of this dissertation as a lot of consideration was placed onto the feasibility of executing the functions and visual appeal of the outputs. We touch on key tools and functionalities of the package as well as provide some detailed examples that users can easily follow and reproduce on their own machines. The package is now available on CRAN and includes several detailed vignettes that we hope users will find helpful for navigating their way around the toolkit.

In Chapter 5, we address the multiple testing problem in the context of our conjugate linear model framework. We modify our model to account for multiple pairwise comparisons and investigate the effects of multiplicity adjustments (e.g. Bonferroni correction and a Bayesian-defined FDR) with respect to sample size and assurance. The Bayesian analogue of the FDR as defined by Müller et al., 2004 is of key interest for this project as it helps frame our updated study objectives and clarifies what the assurance is measuring in this particular setting. Sticking to the two-stage framework proposed in earlier chapters, a large portion of this chapter explains how the analysis and design stages are appropriately calibrated from the single hypothesis testing setting. We conclude the chapter with an assessment on how our proposed model performs in commonplace large-scale problems, particularly microarray data.

We conclude this dissertation with some key takeaways and final points of discussion in Chapter 6. Here, we view our work from a larger scale and elucidate on the universal importance and application of the topics discussed. We also discuss future goals and directions

that our project can move towards as Bayesian SSD is constantly evolving but will nonetheless continue playing an important role both within and beyond the scopes of clinical trial applications. Chapter 6.2 is an Appendix section that showcases derivations and supporting context for claims and statements included in the main text. A large part of the Appendix contains pseudoscripts of the algorithms that are mentioned throughout the dissertation that my projects heavily relied on. We encourage the reader to review these explanations in detail in the hopes of grasping a clearer, more profound understanding of the impact and universal applicability that our work portrays.

CHAPTER 2

Background

This chapter provides the structure of the framework, outlining key concepts that will be repeatedly used and referenced in later chapters. At its core, the model encompasses a conjugate Bayesian linear model framework, which is discussed in greater detail in Section 2.1. Within the model, we motivate the use of two stages characterized by distinctly defined priors that fulfill different study objectives. Advantages for the use of two priors are discussed in Section 2.2. Our two-stage model was largely influenced by the work of O'Hagan and Stevens, 2001, who showcases the two-stage design in the context of sample size determination in clinical trials. Our proposed work casts this approach into a clean and comprehensible linear model. We then elaborate upon two scenarios in Section 2.3: study designs with known variances and study designs with unknown variances. Changes that occur in the linear model and study objectives are communicated, and we explain how to address these two scenarios specifically.

2.1 Conjugate Bayesian Linear Regression

Consider a proposed study where a sample of size n is to be collected in the presence of p controlled explanatory variables, say $x_1, x_2, \ldots, x_p$, that will be known to the investigator for any unit i at the design stage. Let y_n denote the $n \times 1$ random vector of realizable values in the proposed sample. For analysis, the investigator will fit a hierarchical linear regression

model specifying the joint distribution of the parameters $\{\beta, \sigma^2\}$ and the data as

$$IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(y_n \,|\, X_n\beta, \sigma^2 V_n) \;, \tag{2.1}$$

where X_n is $n \times p$ with i-th row corresponding to $x_i^\top$, $\epsilon_n \sim N(0, \sigma^2 V_n)$, and V_n is a known $n \times n$ correlation matrix. We assume X_n and V_n are known for each sample size n from design and modeling considerations. Inference proceeds from the posterior distribution,

$$p(\beta, \sigma^2 \,|\, y_n) = \underbrace{IG(\sigma^2 \,|\, a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 \,|\, y_n)} \times \underbrace{N(\beta \,|\, M_n m_n, \sigma^2 M_n)}_{p(\beta \,|\, \sigma^2, y_n)} \;, \tag{2.2}$$

derived from (2.1), where $a_\sigma^* = a_\sigma + n/2$, $b_\sigma^* = b_\sigma + (1/2)\left\{\mu_\beta^\top V_\beta^{-1}\mu_\beta + y_n^\top V_n^{-1} y_n - m_n^\top M_n m_n\right\}$, $M_n^{-1} = V_\beta^{-1} + X_n^\top V_n^{-1} X_n$ and $m_n = V_\beta^{-1}\mu_\beta + X_n^\top V_n^{-1} y_n$. Sampling from (2.2) is achieved by first sampling $\sigma^2 \sim IG(a_\sigma^*, b_\sigma^*)$ and then sampling $\beta \sim N(M_n m_n, \sigma^2 M_n)$ for each sampled σ^2. See Gelman, Carlin, and Stern, 2013 for further details on Bayesian linear regression.

The objective of the analysis is to ascertain if the data will favor $H : u^\top \beta > 0$, where u is a fixed $p \times 1$ vector. Decision on the tenability of H is often based on the $100(1-\alpha)\%$ posterior credible interval, $\left(u^\top M_n m_n - Z_{1-\alpha/2}\sigma\sqrt{u^\top M_n u},\; u^\top M_n m_n + Z_{1-\alpha/2}\sigma\sqrt{u^\top M_n u}\right)$, obtained from $p(\beta \,|\, \sigma, y_n)$. If y_n belongs to the set $\left\{y_n : u^\top M_m m_n > Z_{1-\alpha/2}\sigma\sqrt{u^\top M_n u}\right\}$, which we denote by $S_\alpha(n; y_n, \sigma, \mu_\beta, V_\beta, V_n)$, then the data favors H. This is equivalent to 0 being below the two-sided $100(1-\alpha)\%$ credible interval for $u^\top\beta$. Practical Bayesian designs will seek to assure the investigator that the above criterion will be achieved with a sufficiently high probability through the *Bayesian assurance*,

$$\delta(n; \sigma, u, \mu_\beta, V_\beta, V_n) = P_{y_n}(S_\alpha(n; y_n, \sigma, \mu_\beta, V_\beta, V_n)) \;. \tag{2.3}$$

Given the fixed values $\{\mu_\beta, V_\beta, \sigma, V_n\}$ and the vector u, the Bayesian assurance function evaluates the probability of rejecting the null hypothesis under the marginal probability distribution of the realized data corresponding to any n. Choice of sample size will be

determined by the smallest value of n that will ensure $\delta(n; \sigma, u, \mu_\beta, V_\beta, V_n) > \gamma$, where γ is the specified assurance.

2.2 Limitations for a single prior

Let us consider the special case when $X_n = 1_n$ so that β is a scalar with prior distribution $\beta \sim N(\beta_1, \sigma^2/n_0)$, where $n_0 > 0$ is a fixed precision parameter (sometimes referred to as "prior sample size"), $V_n = I_n$ and $H : \beta > \beta_0$. We decide in favor of H if the data lies in

$$S_\alpha(n; y, \sigma, \beta_0, \beta_1, n_0) = \left\{ \bar{y} : \bar{y} > \beta_0 - \frac{n_0}{n}(\beta_1 - \beta_0) - \sqrt{\left(1 + \frac{n_0}{n}\right)} \frac{\sigma}{\sqrt{n}} Z_\alpha \right\} ,$$

where the expression on the right reveals a convenient condition in terms of the sample mean. As $n_0 \to 0$, i.e., the prior becomes vague, $S_\alpha(n; y, \sigma, \beta_0, \beta_1, n_0)$ collapses to the critical region in classical inference for testing $H_0 : \beta = \beta_0$ against $H_a : \beta = \beta_1$. The Bayesian assurance function is

$$\delta(n; \sigma, \Delta, n_0) = \Phi\left(\sqrt{n_0} \left[\sqrt{1 + \frac{n_0}{n}} \left(\frac{\Delta}{\sigma} \right) + Z_\alpha \sqrt{\frac{1}{n}} \right] \right) , \tag{2.4}$$

where $\Delta = \beta_1 - \beta_0$. Given n_0, we will compute the sample size needed to detect a critical difference of Δ with probability $1 - \beta$ as $n = \arg\min\{n : \delta(\Delta, n) \geq 1 - \beta\}$. However, the limiting properties of the function in (2.4) are not without problems. When the prior is vague, i.e., $n_0 \to 0$, then $\lim_{n_0 \to 0} \delta(\Delta, n) = \Phi(0) = 0.5$, while in the case when the prior is precise, i.e., $n_0 \to \infty$ we obtain

$$\lim_{n_0 \to \infty} \delta(\Delta, n) = \begin{cases} 1 & \text{if } \Delta > 0 \\ 0 & \text{if } \Delta \leq 0 \end{cases} . \tag{2.5}$$

This is undesirable. Vague priors are customary in Bayesian analysis, but they propagate enough uncertainty that the marginal distribution of the data under the given model will force the assurance to be lower than 0.5. Regardless of how large a sample size we have,

we cannot assure the investigator with probability greater than 50% that H will be tenable. At the other extreme, where the prior is fully precise, it fully dominates the data (or the likelihood) and there is no information from the data that is used in the decision. Therefore, the assurance is a function of the prior only and we will always or never reject the null hypothesis depending upon whether $\Delta > 0$ or $\Delta < 0$. In order to resolve this issue, we work with two different sets of priors, one at the *design* stage and another at the *analysis stage*. Building upon O'Hagan and Stevens, 2001, we elucidate with the Bayesian linear regression model in the next section and offer a simulation-based framework for computing the Bayesian assurance curves.

2.3 Bayesian Assurance Using Design and Analysis Priors

We consider two scenarios based on the population variance σ^2 being known or not. Consider testing the tenability of $H : u^\top \beta > C$ given realized data from a study, where C is a known constant. Recall that $u^\top$ is a $1 \times p$ vector and β is an unknown $p \times 1$ vector of coefficients characterized in the linear regression setting, $y_n = X_n\beta + \epsilon_n, \epsilon_n \sim N(0, \sigma^2 V_n)$.

2.3.1 Known Variance

If σ^2 is known and fixed, then the posterior distribution of β is $p(\beta \,|\, \sigma^2, y_n) = N(\beta \,|\, M_n m_n, \sigma^2 M_n)$ as shown in (2.2). Hence,

$$\frac{u^\top \beta - u^\top M_n m_n}{\sigma\sqrt{u^\top M_n u}} \,\Big|\, \sigma^2, y_n \sim N(0,1)\ . \tag{2.6}$$

To evaluate the credibility of $H : u^\top \beta > C$, where u denotes a known $p \times 1$ vector and C is a known constant, we decide in favor of H if the observed data belongs in the region:

$$A_\alpha(u, \beta, C) = \left\{y_n : P\left(u^\top \beta \leq C | y_n\right) < \alpha\right\} = \left\{y_n : \Phi\left(\frac{C - u^\top M_n m_n}{\sigma\sqrt{u^\top M_n u}}\right) < \alpha\right\}.$$

Given the data y_n and the fixed parameters in the analysis priors, we can evaluate M_n and m_n and hence, for any given σ, C and α, ascertain if we have credibility for H or not.

In the design objective, we need to ask ourselves "What sample size is needed to assure us that the analysis objective is met $100\gamma\%$ of the time?" Therefore, we seek n so that

$$\delta(n) = P_{y_n}(A_\alpha(u, \beta, C)) = P_{y_n}\left\{y_n : \Phi\left(\frac{C - u^\top M_n m_n}{\sigma\sqrt{u^\top M_n u}}\right) < \alpha\right\} \geq \gamma \ , \tag{2.7}$$

where $\delta(n)$ is the Bayesian assurance. In order to evaluate (2.7), we will need the marginal distribution of y_n. In light of the paradox in (2.5), our belief about the population from which our sample will be taken is quantified using the design priors. Therefore, the "marginal" distribution of y_n under the design prior will be derived from

$$y_n = X_n\beta + e_n; \quad e_n \sim N(0, \sigma^2 V_n) \ ; \quad \beta = \mu_\beta^{(d)} + \omega; \quad \omega \sim N(0, \sigma^2 V_\beta^{(d)}) \ , \tag{2.8}$$

where $\beta \sim N(\mu_\beta^{(d)}, \sigma^2 V_\beta^{(d)})$ is the design prior on β. Substituting the β expression into the equation for y_n in (2.8) gives $y_n = X\mu_\beta^{(d)} + (X\omega + e_n)$ and, hence, $y_n \sim N\left(X\mu_\beta^{(d)}, \sigma^2 V_n^*\right)$, where $V_n^* = \left(XV_\beta^{(d)}X^\top + V_n\right)$. We now have a simulation strategy to estimate our Bayesian assurance. We fix sample size n and generate a sequence of J data sets $y_n^{(1)}, y_n^{(2)}, \ldots, y_n^{(J)}$, each of size n from $N\left(X\mu_\beta^{(d)}, \sigma^2 V_n^*\right)$. A Monte Carlo estimate of the Bayesian assurance is given as

$$\hat{\delta}(n) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\left(\left\{y_n^{(j)} : \Phi\left(\frac{C - u^\top M_n^{(j)} m_n^{(j)}}{\sigma\sqrt{u^\top M_n^{(j)} u}}\right) < \alpha\right\}\right) \ , \tag{2.9}$$

where $\mathbb{I}(\cdot)$ is the indicator function of the event in its argument, $M_n^{(j)}$ and $m_n^{(j)}$ are the values of M_n and m_n computed from dataset $y_n^{(j)}$. We repeat the steps needed to compute (2.9) for different values of n and obtain a plot of $\delta(n)$ against n. Our desired sample size is the smallest n for which $\hat{\delta}(n) \geq \gamma$, where we seek assurance of a $100\gamma\%$ chance of deciding in favor of H.

Algorithm 1 in Appendix A includes the pseudocode for computing Bayesian assurance in the known variance setting. A special case of the model can be considered where $X_n = 1_n$ is an $n \times 1$ vector of ones, β is a scalar, $V_n = I_n$ and we wish to evaluate the credibility of $H : \beta > \beta_0$. An explanation on how these specifications bring us back to the frequentist setting can be found in Section B.1 of Appendix B.

2.3.2 Unknown Variance

When σ^2 is unknown, the posterior distribution of interest is $p(\beta, \sigma^2 \,|\, y_n)$ as opposed to the original $p(\beta \,|\, \sigma^2, y_n)$ delineated in the known variance case. Since σ^2 is no longer fixed, it becomes challenging to define a closed form condition that is capable of evaluating the credibility of $H : u^\top \beta > C$. Hence, we do not obtain a condition similar to (2.6). However, our region of interest corresponding to our analysis objective still remains as $A_\alpha(u, \beta, C) = \left\{ y_n : P\left(u^\top \beta \leq C \,|\, y_n\right) < \alpha \right\}$ when deciding whether or not we are in favor of H. To implement this in a simulation setting, we rely on iterative sampling for both β and σ^2 to estimate the assurance. We specify analysis priors $\beta \,|\, \sigma^2 \sim N(\mu_\beta^{(a)}, \sigma^2 V_\beta^{(a)})$ and $\sigma^2 \sim IG(a^{(a)}, b^{(a)})$, where the superscripts (a) indicate analysis stage priors.

We had previously derived the posterior distribution of β in Section 2.3.1 expressed as $p(\beta \,|\, y_n, \sigma^2) = N(\beta \,|\, M_n m_n, \sigma^2 M_n)$, where $M_n = (V_\beta^{-1(a)} + X^\top V_n^{-1} X)^{-1}$ and $m_n = V_\beta^{-1(a)} \mu_\beta^{(a)} + X^\top V_n^{-1} y_n$. The posterior distribution of σ^2 is obtained by integrating out β from the joint posterior distribution of $\{\beta, \sigma^2\}$, which yields

$$\begin{aligned} p(\sigma^2 \,|\, y_n) &\propto IG(\sigma^2 \,|\, a^{(a)}, b^{(a)}) \times \int N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(y_n \,|\, X\beta, \sigma^2 V_n) d\beta \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a^{(a)} + \frac{n}{2} + 1} \exp\left\{ -\frac{1}{\sigma^2} \left(b^{(a)} + \frac{c^*}{2} \right) \right\} . \end{aligned} \tag{2.10}$$

Therefore, $p(\sigma^2 \,|\, y_n) = IG\left(\sigma^2 \,|\, a^*, b^*\right)$, where $a^* = a^{(a)} + \frac{n}{2}$ and $b^* = b^{(a)} + \frac{c^*}{2} = b^{(a)} + \frac{1}{2}\left\{ \mu_\beta^{\top(a)} V_\beta^{-1(a)} \mu_\beta^{(a)} + y_n^\top V_n^{-1} y_n - m_n^\top M_n m_n \right\}$.

Recall that the design stage objective aims to identify minimum sample size n needed to attain the assurance level specified by the investigator. Similar to Section 2.3.1 we will need the marginal distribution of y_n with priors placed on both β and σ^2. Derivation steps are almost identical to those outlined in (2.8) for the known σ^2 case. With the variance unknown, the marginal distribution of y_n under the design prior is derived from $y_n = X_n\beta + e_n,\ e_n \sim N(0, \sigma^2 V_n),\ \beta = \mu_\beta^{(d)} + \omega; \quad \omega \sim N(0, \sigma^2 V_\beta^{(d)})$, where $\beta \sim N(\mu_\beta^{(d)}, \sigma^2 V_\beta^{(d)})$ and $\sigma^2 \sim IG(a^{(d)}, b^{(d)})$. Substituting β into y_n gives us $y_n = X_n\mu_\beta^{(d)} + (X_n\omega + e_n)$ such that $X_n\omega + e_n \sim N(0, \sigma^2(V_n + X_n V_\beta^{(d)} X_n^\top))$. The marginal distribution of $p(y_n \,|\, \sigma^2)$ is therefore

$$y_n \,|\, \sigma^2 \sim N(X_n\mu_\beta^{(d)}, \sigma^2 V_n^*); \quad V_n^* = X_n V_\beta^{(d)} X_n^\top + V_n, \tag{2.11}$$

which specifies our data generation model for ascertaining sample size.

Each iteration comprises the design stage, where the data is generated, and an analysis stage where the data is analyzed to ascertain whether a decision favorable to the hypothesis has been made. In the design stage, we draw σ^2 from $IG(a_\sigma, b_\sigma^{(d)})$ and generate the data from our sampling distribution given in (2.11), $y_n \sim N(X\mu_\beta^{(d)}, \sigma^2(XV_\beta^{(d)}X^\top + V_n))$. For each such data set, $\{y_n, X_n\}$, we perform Bayesian inference for β and σ^2. Here, we draw J samples of β and σ^2 from their respective posterior distributions and compute the proportion of these J samples that satisfy $u^\top\beta_j > C$. If the proportion exceeds a certain threshold $1 - \alpha$, then the analysis objective is met for that dataset. The above steps for the design and analysis stage are repeated for R datasets and the proportion of the R datasets that meet the analysis objective, i.e., deciding in favor of H, correspond to the Bayesian assurance. Algorithm 2 in Appendix A includes the pseudocode for computing Bayesian assurance in the unknown variance setting.

CHAPTER 3

Two-Stage Paradigm Applications

This chapter utilizes our two-stage framework discussed in Chapter 2 for three existing sample size determination approaches. We show how these approaches emerge as special cases of our framework with an appropriate formulation of analysis and design stage objectives. Assurance curves are produced via simulation and pseudocodes of the algorithms can be found in Appendix A.

3.1 Sample Size Determination in Cost-Effectiveness Setting

The first application selects a sample size based on the cost-effectiveness of new treatments undergoing Phase 3 clinical trials (O'Hagan and Stevens, 2001). As outlined in Section 2.1, we construct the two-stage paradigm in the context of a conjugate linear model and generalize it to the case where the population variance σ^2 is unknown. We cast the example in O'Hagan and Stevens, 2001 within our framework to assess overall performance and our framework's ability to emulate results of the analysis in O'Hagan and Stevens, 2001.

Consider designing a randomized clinical trial where n_1 patients are administered Treatment 1 and n_2 patients are administered Treatment 2 under some suitable model and study objectives. Let c_{ij} and e_{ij} respectively denote the observed cost and observed efficacy outcomes corresponding to patient $j = 1, 2, \ldots, n_i$ receiving treatment i for $i = 1, 2$ treatments, where n_i is the number of patients in the i-th treatment group. The expectation of e_{ij} under treatment i is $E(e_{ij}) = \mu_i$, and the population mean cost under treatment i is $E(c_{ij}) = \gamma_i$.

The variances are taken as $Var(c_{ij}) = \tau_i^2$ and $Var(e_{ij}) = \sigma_i^2$. For simplicity, we assume equal sample sizes within the treatment groups so that $n = n_1 = n_2$. We also assume equal sample variances for the outcomes such that $\tau^2 = \tau_1^2 = \tau_2^2$ and $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

O'Hagan and Stevens, 2001 utilize the net monetary benefit measure,

$$\xi = K(\mu_2 - \mu_1) - (\gamma_2 - \gamma_1), \tag{3.1}$$

where $\gamma_2 - \gamma_1$ and $\mu_2 - \mu_1$ denote the true differences in costs and efficacy, respectively, between Treatment 1 and Treatment 2, and K represents the maximum price that a health care provider is willing to pay in order to obtain a unit increase in efficacy, also known as the threshold unit cost. The quantity ξ acts as a measure of cost-effectiveness. Since the net monetary benefit formula expressed in Equation (3.1) involves assessing the cost and efficacy components conveyed within each of the two treatment groups, we set $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2)^\top$, where μ_i and γ_i denote the efficacy and cost for treatments $i = 1, 2$. Next, we specify y_n as a $4n \times 1$ vector consisting of 2×1 vectors $y_{ij} = (c_{ij}, e_{ij})^\top$, $i = 1, 2$ and $j = 1, 2, \ldots, n$. Each individual observation is allotted one row in the linear model. The design matrix X_n is a $4n \times 4$ block diagonal with the $n \times 1$ vector of ones, 1_n, as the blocks. With $n = n_1 = n_2$, $\sigma^2 = \sigma_1^2 = \sigma_2^2$ and $\tau^2 = \tau_1^2 = \tau_2^2$, our variance matrix is $\sigma^2 \underset{4n\times 4n}{V_n} = \sigma^2 \begin{pmatrix} I_n & O & O & O \\ O & \frac{\tau^2}{\sigma^2} I_n & O & O \\ O & O & I_n & O \\ O & O & O & \frac{\tau^2}{\sigma^2} I_n \end{pmatrix}$, where σ^2 is factored out to comply with our conjugate linear model formulation expressed in (2.1).

In the analysis stage, we use the posterior distribution for β if σ^2 is fixed or for $\{\beta, \sigma^2\}$ if σ^2 needs to be estimated; recall Sections 2.3.1 and 2.3.2. The posterior distribution is needed only for the analysis stage, hence it is computed using the analysis priors. Since there is no data in the design stage, there is no posterior distribution. We use the design priors as specifications for the population from which the data are generated. That is, the

design priors yield the sampling distribution $y_n \mid \sigma^{2(d)}$.

Lastly, we use the design priors specified in O'Hagan and Stevens, 2001 (refer to Section B.2 of Appendix B) and set the posterior probability of deciding in favor of H to at least 0.975, which is equivalent to a Type I error of $\alpha = 0.025$ in frequentist two-sided hypothesis tests.

3.2 Design for Cost-Effectiveness Analysis

Consider designing a trial to evaluate the cost effectiveness of a new treatment compared to an original treatment. We seek the tenability of $H : \xi > 0$, where ξ is the net monetary benefit defined in Equation (3.1). Using the specifications provided in Section 3.1 we conduct simulations in the known and unknown σ^2 cases to emulate the results reported in O'Hagan and Stevens, 2001.

3.2.1 Simulation Results in the Known Variance Case

Table 3.1 presents estimated Bayesian assurance values corresponding to different values of K and sample sizes n. Displayed (K, n) pairs correspond to the specifications reported in O'Hagan and Stevens, 2001 that ensure a 0.70 assurance level. The "maxiter" variable denotes the number of datasets being simulated. All of the resulting assurance values in Table 3.1 for all combinations of K and n are close to 0.70. Looking by columns, we observe that the assurance values exhibit some deviations for all cases as we increase the number of datasets generated. No obvious trends of precision are observed in any of the four cases. Looking across the table rows, we observe that larger sample sizes tend to yield assurance values that are consistently closer to the 0.70 mark, which is to be expected. The first column, corresponding to the case with the largest sample size of $n = 1048$, consistently produced results that meet the assurance criteria of 0.70. These results show that sampling from the posterior provides results very similar to those reported by O'Hagan and Stevens,

Table 3.1: Recorded simulation results from Bayesian assurance algorithm with varying number of iterations.

	Outputs from Bayesian Assurance Algorithm			
maxiter	K = 5000 n = 1048	K = 7000 n = 541	K = 10000 n = 382	K = 20000 n = 285
250	0.708	0.676	0.688	0.716
500	0.701	0.714	0.676	0.698
1000	0.700	0.694	0.697	0.719

2001.

3.2.2 Simulation Results in Unknown Variance Case

We now consider the setting where σ^2 is unknown. This extends the analysis in O'Hagan and Stevens, 2001 who treated the cost-effectiveness problem with fixed variances. Table 3.2 presents the estimated Bayesian assurance values in this setting under the same (K, n) specifications used in Table 3.1. Table 3.2 does not align as closely as the assurance results we had obtained from implementing the fixed σ^2 simulation in Section 3.2.1.

Referring to Table 3.2, we let R denote the number of outer loop iterations. The primary purpose of the outer loop is to randomly draw design stage variances $\sigma^{2(d)}$ from the $IG(a^{(d)}, b^{(d)})$ distribution. Recall from Section 2.3.2 that $\sigma^{2(d)}$ is used for computing the variance of the marginal distribution from which we are drawing our sampled observations, $y \,|\, \sigma^{2(d)}$. The inner-loop iterations sample data using the marginal distribution of σ^2 from Equation (2.10). The number of iterations in the inner-loop is set to 750 for all cases. We notice that a majority of our trials report assurance values close to the 0.70 mark, particularly for the case in which we set the sample size to $n = 1048$. The trial that exhibited the greatest deviation was for threshold cost $K = 20000$ with corresponding sample size $n = 285$, which returned an assurance of 0.58. This is most likely attributed to using a smaller sample size to gauge the effect size.

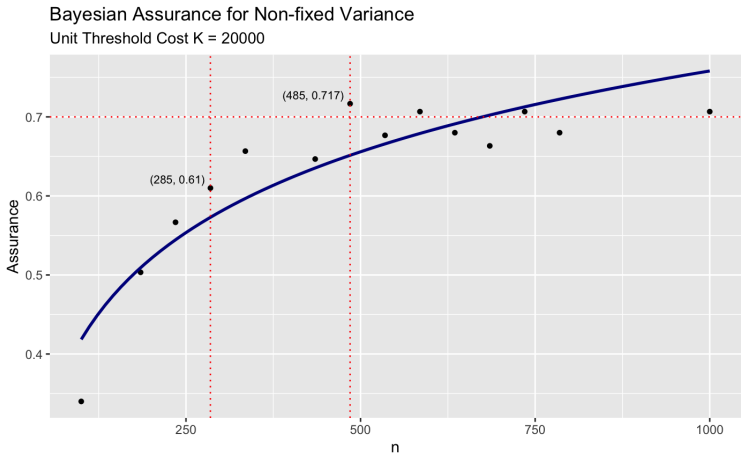


Figure 3.1: Assurance curve based on results of algorithm corresponding to unknown variance.

A visual depiction for this case can be seen in Figure 3.1. The dashed line on the left showcases the expected minimum sample size needed to achieve a 0.70 assurance whereas the dashed line on the right marks the point at which our algorithm actually achieves this desired threshold. The reality of the situation is that the problem setup gets changed quite a bit once we remove the assumption that σ^2 is known and fixed. If we look at the individual points marked on the plot, assurance values of 0.61 and 0.71 don't appear too different. If we were to solely account for the fact that these Monte Carlo estimates are subject to error given that the estimates are based on means and variances that were compositely sampled rather than being taken in as fixed assignments, our algorithm performs remarkably well, but there are still points to be wary about. The x-axis of the plot indicates that an assurance of

Table 3.2: Recorded results from the Bayesian assurance algorithm for unknown variance with varying number of iterations.

	Outputs from Bayesian Assurance Algorithm			
R	$K = 5000$ n = 1048	$K = 7000$ n = 541	$K = 10000$ n = 382	$K = 20000$ n = 285
100	0.698	0.718	0.72	0.601
150	0.702	0.713	0.72	0.579

0.70 (red dotted line) can only be ensured once we recruit a sample size of at least $n = 485$ per treatment group. This is substantially larger compared to the known σ^2 case, suggesting a need to recruit nearly twice as many participants as what was needed in Table 3.1. These results evince the pronounced impact of uncertainty in the design on the sample size needed to achieve a fixed level of Bayesian assurance.

3.3 Assurance Computation in the Longitudinal Setting

We demonstrate an additional concept that computes the assurance using longitudinal data. In this setting, n no longer refers to the number of subjects per study design group but rather the number of repeated measures reported for each subject assuming a balanced study design. Referring back to the linear regression model discussed in the general framework, we can construct a longitudinal model that utilizes this same linear regression form, where $y_n = X_n\beta + \epsilon_n$.

Consider a group of subjects in a balanced longitudinal study with the same number of repeated measures at equally-spaced time points. In the base case, where time is treated as a linear term, subjects can be characterized by

$$y_{ij} = \alpha_i + \beta_i t_{ij} + \epsilon_i,$$

Table 3.3: Recorded simulation results from Bayesian assurance algorithm with varying number of iterations.

	Outputs from Bayesian Assurance Algorithm			
maxiter	K = 5000 n = 1048	K = 7000 n = 541	K = 10000 n = 382	K = 20000 n = 285
250	0.708	0.676	0.688	0.716
500	0.701	0.714	0.676	0.698
1000	0.700	0.694	0.697	0.719

where y_{ij} denotes the j^{th} observation of subject i at time t_{ij}, α_i and β_i respectively denote the intercept and slope terms for subject i, and ϵ_i is an error term characterized by $\epsilon_i \sim N(0, \sigma_i^2)$.

In a simple case with two subjects, we can individually express the observations as

$$\begin{aligned}
y_{11} &= \alpha_1 + \beta_1 t_{11} + \epsilon_1 \\
&\vdots \\
y_{1n} &= \alpha_1 + \beta_1 t_{1n} + \epsilon_1 \\
y_{21} &= \alpha_2 + \beta_2 t_{21} + \epsilon_2 \\
&\vdots \\
y_{2n} &= \alpha_2 + \beta_2 t_{2n} + \epsilon_2,
\end{aligned}$$

assuming that each subject contains n observations. The model can also be expressed cohesively through matrices,

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \end{pmatrix}}_{y_n} = \underbrace{\begin{pmatrix} 1 & 0 & t_{11} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & t_{1n} & 0 \\ 0 & 1 & 0 & t_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & t_{2n} \end{pmatrix}}_{X_n} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_2 \end{pmatrix}}_{\epsilon_n} \tag{3.2}$$

bringing us back to the linear model structure. If higher degrees are to be considered for the time variable, such as the inclusion of a quadratic term, the model would be altered to include additional covariate terms that can accommodate to these changes. In the two-subject case, incorporating a quadratic term for the time variable in Equation (3.2) will

result in the following modified model:

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \end{pmatrix}}_{y_n} = \underbrace{\begin{pmatrix} 1 & 0 & t_{11} & 0 & t_{11}^2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & t_{1n} & 0 & t_{1n}^2 & 0 \\ 0 & 1 & 0 & t_{21} & 0 & t_{21}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & t_{2n} & 0 & t_{2n}^2 \end{pmatrix}}_{X_n} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \phi_1 \\ \phi_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_2 \end{pmatrix}}_{\epsilon_n}.$$

In general, for m subjects who each have n repeated measures, a one-unit increase in the degree of the time-based covariate will result in m additional columns being added to the design matrix X_n and m additional rows appended to the β vector.

3.3.1 Longitudinal Example

Assume there are two subjects and we want to test whether the growth rate of subject 1 is different in comparison to subject 2. This could have either positive or negative implications depending on the measurement scale. Figure 3.2 displays the estimated assurance points given the specifications.

Assigning an appropriate linear contrast lets us evaluate the tenability of an outcome. Let us consider the tenability of $u^\top \beta \neq C$, where $u = (1, -1, 1, -1)^\top$ and $C = 0$. The timepoints are arbitrarily chosen to be 0 through 120, which could be based on days, months, or years depending on the context of the problem. The number of repeated measurements per subject to be tested includes values 10 through 100 in increments of 5. This indicates that we are evaluating the assurance for 19 study designs in total. $n = 10$ divides the specified time interval into 10 evenly-spaced timepoints between 0 and 120.

For a more complicated study design comprised of more than two subjects that are divided into two treatment groups, consider testing if the mean growth rate is higher in the first

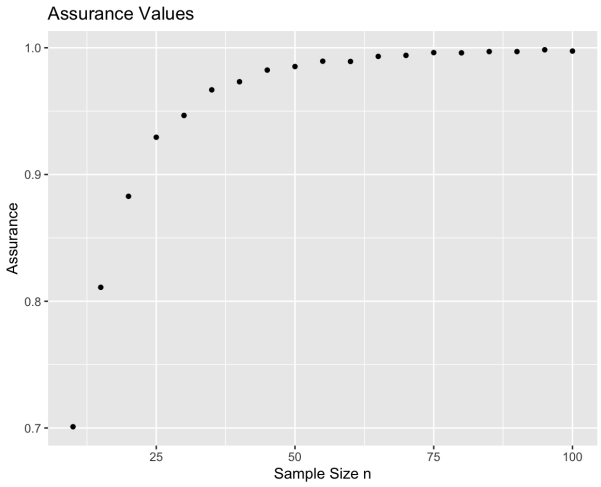


Figure 3.2: Estimated assurance points for longitudinal example.

treatment group than that of the second, e.g. if we have three subjects per treatment group, the linear contrast would be set as $u = (0, 0, 0, 0, 0, 0, 1/3, 1/3, 1/3, -1/3, -1/3, -1/3)^\top$.

3.4 Sample Size Determination Using Precision-Based Conditions

We now consider a few alternate Bayesian approaches for sample size determination and demonstrate how these methods can be embedded within the two-stage Bayesian framework. We also identify special cases that overlap with the frequentist setting.

Adcock, 1997 constructs rules based on a fixed precision level d. In the frequentist setting, if $X_i \sim N(\theta, \sigma^2)$ for $i = 1, ..., n$ observations and variance σ^2 is known, the precision can be calculated using $d = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$, where $z_{1-\alpha/2}$ is the critical value for the $100(1-\alpha/2)\%$ quartile of the standard normal distribution. Simple rearranging leads to following expression for sample size,

$$n = z^2_{1-\alpha/2}\frac{\sigma^2}{d^2} \ . \tag{3.3}$$

Given a random sample with mean $\bar{x}$, suppose the goal is to estimate population mean θ. The analysis objective decides whether or not the absolute difference between $\bar{x}$ and θ falls within a margin of error no greater than d. Given sample mean $\bar{x}$ and a pre-specified confidence

level α, the assurance is given as

$$\delta = P_{\bar{x}}\{\bar{x} : P(|\bar{x} - \theta| \leq d) \geq 1 - \alpha\} \ . \tag{3.4}$$

To formulate the problem in the Bayesian setting, suppose $x_1, \cdots, x_n$ is a random sample from $N(\theta, \sigma^2)$ and the sample mean is distributed as $\bar{x}|\theta, \sigma^2 \sim N(\theta, \sigma^2/n)$. We assign $\theta \sim N(\theta_0^{(a)}, \sigma^2/n_a)$ as the analysis prior, where n_a quantifies the amount of prior information we have for θ. Adhering to the notation in previous sections, subscript (a) denotes we are working within the analysis stage. Referring to (3.4), the analysis stage objective involves observing $|\bar{x} - \theta| \leq d$. If the analysis objective holds to a specified probability level, then the corresponding sample size of the data being passed through the condition is sufficient in fulfilling the desired precision level for the study. Additional steps can be taken to expand out the analysis objective given in Equation (3.4). These steps are outlined in Section B.3 of Appendix B, where the analysis stage objective is

$$\left\{\bar{x} : \Phi\left[\frac{\sqrt{n_a + n}}{\sigma}(\bar{x} + d - \lambda)\right] - \Phi\left[\frac{\sqrt{n_a + n}}{\sigma}(\bar{x} - d - \lambda)\right] \geq 1 - \alpha\right\}. \tag{3.5}$$

In the design stage, we need to construct a protocol for sampling data that will be used to evaluate the analysis objective. This is achieved by setting a separate design stage prior on θ such that $\theta \sim N(\theta_0^{(d)}, \sigma^2/n_d)$, where n_d quantifies our degree of belief towards the population from which the sample will be drawn. Given that $\bar{x}|\theta, \sigma^2 \sim N(\theta, \sigma^2/n)$, the marginal distribution of $\bar{x}$ can be computed using straightforward substitution based on $\bar{x} = \theta + \epsilon$; $\epsilon \sim N(0, \sigma^2/n)$ and $\theta = \theta_0^{(d)} + \omega$; $\quad \omega \sim N(0, \sigma^2/n_d)$. Substituting θ into the expression for $\bar{x}$ gives us $\bar{x} = \theta_0^{(d)} + (\omega + \epsilon)$; $(\omega + \epsilon) \sim N\Big(0, \frac{\sigma^2(n_d+n)}{n_d n}\Big) = N(0, \sigma^2/p)$, where $1/p = 1/n_d + 1/n$. The marginal of $\bar{x}$ is therefore $N(\bar{x}|\theta_0^{(d)}, \sigma^2/p)$, where we will be iteratively drawing our samples from to check if the sample means satisfy the condition derived in Equation (3.5). The Monte Carlo estimate of the assurance is therefore obtained

by evaluating the proportion of J samples that meet the analysis stage criteria,

$$\hat{\delta}(n) = \frac{1}{J}\sum_{j=1}^{J}\mathbb{I}\Bigg(\Bigg\{\bar{x}^{(j)} : \Phi\left[\frac{\sqrt{n_a+n}}{\sigma}(\bar{x}^{(j)}+d-\lambda^{(j)})\right] - \Phi\left[\frac{\sqrt{n_a+n}}{\sigma}(\bar{x}^{(j)}-d-\lambda^{(j)})\right] \geq 1-\alpha\Bigg\}\Bigg),$$

where the j-th sample mean is given by $\bar{x}^{(j)}$. Algorithm 3 in Appendix A provides the pseudocode for the above procedure and simulation results can be found in Section B.4 of Appendix B.

Notice the assurance in the precision-based setting is not linked to a hypothesis testing framework. Hence, we can not translate the above scenario to a frequentist-based paradigm that will enable direct comparisons between assurance and power. We can still demonstrate the relationship held between the Bayesian and frequentist settings through proper specifications of analysis and design stage precision parameters, n_a and n_d. Section B.5 in Appendix B discusses this in detail.

3.5 Sample Size Determination in a Beta-Binomial Setting

We revisit the hypothesis testing framework with proportions. Pham-Gia, 1997 outlines steps for determining exact sample sizes needed in estimating differences of two proportions in a Bayesian context. Let $p_i, i = 1, 2$ denote two independent proportions. In the frequentist setting, suppose the hypothesis test to undergo evaluation is $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$. As described in Pham-Gia, 1997, one method of approach is to check whether or not 0 is contained within the confidence interval bounds of the true difference in proportions given by $(\hat{p_1} - \hat{p_2}) \pm z_{1-\alpha/2}(SE(\hat{p_1})^2 + SE(\hat{p_2})^2)^{1/2}$, where $z_{1-\alpha/2}$ denotes the critical region, and $SE(\hat{p_i})$ denotes the standard error of p_i obtained by $SE(\hat{p_i}) = \sqrt{\frac{\hat{p_i}(1-\hat{p_i})}{n_i}}$. An interval without 0 contained within the bounds suggests there exists a significant difference between

the two proportions.

The Beta distribution is often used to represent outcomes tied to a family of probabilities. The Bayesian setting uses posterior credible intervals as an analog to the frequentist confidence interval approach. As outlined in Pham-Gia, 1997, two individual priors are assigned to p_1 and p_2 such that $p_i \sim \text{Beta}(\alpha_i, \beta_i)$ for $i = 1, 2$. In the case of binomial sampling, X is treated as a random variable taking on values $x = 0, 1, ..., n$ to denote the number of favorable outcomes out of n trials. The proportion of favorable outcomes is therefore $p = x/n$. Suppose a Beta prior is assigned to p such that $p \sim \text{Beta}(\alpha, \beta)$. The prior mean and variance are respectively $\mu_{prior} = \frac{\alpha}{\alpha+\beta}$ and $\sigma^2_{prior} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Conveniently, given that p is assigned a Beta prior, the posterior of p also takes on a Beta distribution with mean and variance

$$\begin{aligned} \mu_{posterior} &= \frac{\alpha + x}{\alpha + \beta + n} \\ \sigma^2_{posterior} &= \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}. \end{aligned} \tag{3.6}$$

Within the analysis stage, we assign two beta priors for p_1 and p_2 such that $p_i \sim \text{Beta}(\alpha_i, \beta_i)$, $i = 1, 2$. If we let $p_d = p_1 - p_2$ and p_{post} and $\text{var}(p)_{\text{post}}$ respectively denote the posterior mean and variance of p_d, it is straightforward to deduce that $p_{\text{post}} = \frac{\alpha_1+x_1}{\alpha_1+\beta_1+n_1} - \frac{\alpha_2+x_2}{\alpha_2+\beta_2+n_2}$ and $\text{var}(p)_{\text{post}} = \frac{(\alpha_1+x_1)(\beta_1+n_1-x_1)}{(\alpha_1+\beta_1+n_1)^2(\alpha_1+\beta_1+n_1+1)} + \frac{(\alpha_2+x_2)(\beta_2+n_2-x_2)}{(\alpha_2+\beta_2+n_2)^2(\alpha_2+\beta_2+n_2+1)}$ from Equation (3.6). Hence the resulting $100(1-\alpha)\%$ posterior credible interval equates to $p_{\text{post}} \pm z_{1-\alpha/2}\sqrt{\text{var}(p)_{\text{post}}}$, which, similar to the frequentist setting, would be used to check whether 0 is contained within the credible interval bands as part of our inference procedure. This translates to become our analysis objective, where we are interested in assessing if each iterated sample outputs a credible interval that does not contain 0. We can denote this region of interest as $A_\alpha\left(n_1, n_2; x_1, x_2, \alpha_1, \alpha_2, \beta_1, \beta_2\right)$ such that

$$A_\alpha\left(n_1, n_2; x_1, x_2, \alpha_1, \alpha_2, \beta_1, \beta_2\right) = \left\{ x_1, x_2 : 0 \notin \left(p_{\text{post}} \pm z_{1-\alpha/2}\sqrt{\text{var}(p)_{\text{post}}} \right) \right\}. \tag{3.7}$$

The assurance for assessing a significant difference in proportions is

$$\delta = P_{x_1,x_2}\left(A_\alpha\left(n_1, n_2; x_1, x_2, \alpha_1, \alpha_2, \beta_1, \beta_2\right)\right).$$

For the design stage, note that the simulated data in Beta-Binomial setting pertains to the frequency of positive outcomes, x_1 and x_2, observed among the two samples. These frequency counts are observed from samples of size n_1 and n_2 based on given probabilities, p_1 and p_2, that are passed in the analysis stage. Once p_1 and p_2 are assigned, x_1 and x_2 values are randomly generated from their corresponding binomial distributions, where $x_i \sim \text{Bin}(n_i, p_i), i = 1, 2$. The posterior credible intervals are subsequently computed to undergo assessment in the analysis stage. These steps are repeated iteratively starting from the generation of x_1 and x_2 values. The proportion of iterations with results that fall within the region of interest expressed in Equation (3.7) equates to the assurance. Algorithm 4 in Appendix A provides the pseudocode used to implement the simulation study. Section B.6 in Appendix B discusses parallel behaviors exhibited between Bayesian and frequentist settings using the above criteria.

3.6 A Utility-Based Approach in the Bayesian Setting

The utility based approach for Bayesian SSD (Lindley, 1997; Muller et al., 1992; Raiffa and Schlaifer, 1961) maximizes the expected utility function $\mathbb{E}_{(y,\theta)}[U(n, y, \theta, d)]$ as a function of n, where n is sample size, y_n is the realizable data, θ is an unknown parameter, and d is a decision to be taken based upon our inference for θ. The desired sample size is

$$n_* = \arg\max_n \mathbb{E}_{y_n \,|\, n}\left[\mathbb{E}_{\theta \,|\, y_n}\left[U(n, y_n, \theta, d)\right]\right], \tag{3.8}$$

where $\mathbb{E}_{\theta \,|\, y_n}[\cdot]$ and $\mathbb{E}_{y_n \,|\, n}[\cdot]$ are the expectations with respect to $p(\theta \,|\, y_n)$ and $p(y_n \,|\, n)$, respectively. Practical implementation, then requires specifying the utility function and the joint

model $p(y_n, \theta \,|\, n) = p(y_n \,|\, n) \times p(\theta \,|\, y_n)$.

Following Inoue, Berry, and Parmigiani, 2005, who connected Bayesian decision-theoretic SSD with classical SSD, we draw connections between Bayesian decision-theoretic SSD and Bayesian assurance. We extend Inoue, Berry, and Parmigiani, 2005 to the hierarchical linear model setting and, as before, consider the decision of favoring $H : u^\top \beta > C$. Let $d(y_n) \in \{0, 1\}$ be a binary decision function according to whether we decide in favor of H ($d(y_n) = 1$) or not ($d(y_n) = 0$). Inoue, Berry, and Parmigiani, 2005 considers the utility function $U(n, d(y_n), \beta)$ allocating K units for correctly deciding against H, 1 unit for correctly deciding in favor of H, and 0 for all other (incorrect) decisions. A general expression for expected utility is

$$G_B(n, v) = KP(H_0)P(\text{fail to reject } H_0 |\ H_0 \text{ is true}) + P(H_a)P(\text{reject } H_0 |\ H_a \text{ is true}), \quad (3.9)$$

where v denotes a vector of user-specified inputs in addition to sample size n. The value of $G_B(n, v)$ obtained once the inputs are processed is the overall r^* value.

Consider the linear hypothesis test $H_0 : u^\top \beta = c_0$ vs. $H_a : u^\top \beta = c_1$ under the linear model $y = X\beta + \epsilon$, where y is $n \times 1$, X is $n \times p$, β is $p \times 1$, u is $p \times 1$, and $\epsilon \sim N(0, \sigma^2 I_n)$, implying that c_0 and c_1 are scalars. Under this hypothesis testing framework, we assign appropriate cutoffs that enable the function to objectively determine the rate of correct classification in the Bayesian setting. First, we assign a prior on $u^\top \beta$ such that $P(H_0) = 1 - P(H_a) = \pi$ and assume that the null hypothesis is not rejected if the posterior probability of H_0 is at least $1/(1 + K)$, where K is the assigned utility associated with H_0 being correctly accepted. We also assume $u^\top \beta$ is estimable, implying there is some $z \in \mathbb{R}^n$ such that $u = X^\top z$. Section B.7 of Appendix B provides detailed derivation steps to arrive at the following expression denoting the probability of correctly accepting H_0:

$$P(\text{fail to reject } H_0 |\ H_0 \text{ is true}) = \Phi\left[\frac{\sigma\sqrt{z^\top z}}{\delta} \ln\left(\frac{K\pi}{1-\pi}\right) + \frac{\delta}{2\sigma\sqrt{z^\top z}}\right],$$

where Φ denotes the cumulative distribution function of the standard normal and $\delta = c_1 - c_0$. Similar steps can be carried out to obtain the probability of correctly rejecting H_0:

$$P(\text{reject } H_0 |\ H_a \text{ is true}) = 1 - \Phi\left[\frac{\sigma\sqrt{z^\top z}}{\delta}\ln\left(\frac{K\pi}{1-\pi}\right) - \frac{\delta}{2\sigma\sqrt{z^\top z}}\right] .$$

We now have all necessary components to obtain a full expression of the utility function based on the pre-specified cutoff, $P(H_0|y) \geq \frac{1}{1+K}$. Referring to Equation (3.9), we can substitute all fixed assignments and cutoff criteria to obtain

$$G_B(n, \mathbf{v}) = K\pi\Phi\left[\frac{\sigma\sqrt{z^\top z}}{\delta}\ln\left(\frac{K\pi}{1-\pi}\right) + \frac{\delta}{2\sigma\sqrt{z^\top z}}\right] + (1-\pi)\left(1 - \Phi\left[\frac{\sigma\sqrt{z^\top z}}{\delta}\ln\left(\frac{K\pi}{1-\pi}\right) - \frac{\delta}{2\sigma\sqrt{z^\top z}}\right]\right). \quad (3.10)$$

3.6.1 Comparing the Rate of Correct Classification to Assurance

The assurance and the rate of correct classification each produce curves that are visually similar when plotted against sample size. While the assurance reports the probability of satisfying a desired study objective, the rate of correct classification measures the probability of reaching an accurate conclusion from a decision-theoretic perspective. In what circumstances, then, do these two curves overlap and share the same probabilities?

Using the parameter assignments provided in Section B.8 of Appendix B, we study and compare the behaviors of the two metrics by plotting them simultaneously on the same graph. Figure 3.3 plots the rate of correct classification (x-axis) against the assurance (y-axis) in the case when the analysis stage prior is weakly assigned and the design stage prior is strongly assigned. Each point on the curve indicates a pair of matching assurance and rate of correct classification values that yield the same sample size. Note that the assurance values under these particular specifications also denote the approximate estimates of the frequentist power values. Furthermore, the critical difference is set to $\delta = 0.10$ and the variance is $\sigma^2 = 1$.

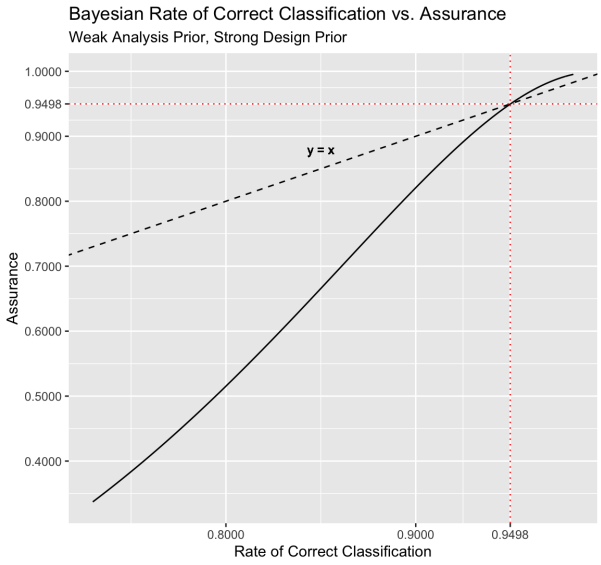


Figure 3.3: Assurance behavior in relation to rate of correct classification with a weakly assigned analysis prior ($n_a = 0$), a strongly assigned design prior ($n_d \to \infty$), fixed critical difference $\delta = 0.10$, variance $\sigma^2 = 1$, and utility $K = 1$.

A reference line with a slope of 1 is included to better visualize the convergence of the two probabilities, which occurs at approximately 0.95. Overall, we observe large disparities between the two probabilities, especially in the earlier parts of the graph as indicated by the larger gap presented between the curve and reference line. The sample sizes are not explicitly shown on the curve, though it can be inferred that larger sample sizes are tied to larger assurance and rate of correct classification values. For example, a sample size of 285 yields an assurance of approximately 0.52 and a rate of correct classification of approximately 0.8. A sample size of 1085 is needed to ensure an equally large r^* and assurance value of 0.95, as highlighted by the intersection of the red dashed lines.

Figure 3.4 treats the results displayed in Figure 3.3 as a default curve and compares it to those produced under various specifications of the precision parameters, n_a and n_d, where only one parameter is adjusted for each case. Dashed curves signify adjusted n_a values while dotted curves indicate modifications to n_d. Different colors denote different specifications of the precision parameters. The solid green curve in the center corresponds to the original

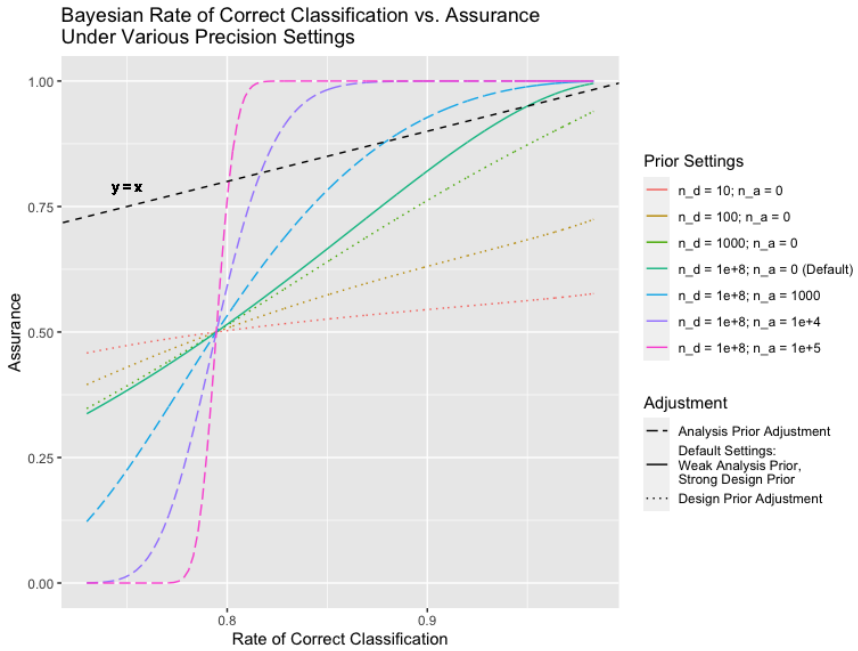


Figure 3.4: Assurance behavior in relation to rate of correct classification for fixed critical difference $\delta = 0.10$, variance $\sigma^2 = 1$, and various precision settings within the analysis and design stages of the assurance framework.

curve displayed in Figure 3.3.

Among the set of curves displayed in Figure 3.4, intersections between the assurance and rate of correct classification are observed in cases where modifications are made to the analysis stage prior and the design stage prior is set to be precise ($n_d = \infty$). This is particularly true when n_a is altered to become more precise, where, as the value of n_a increases, the points of intersection between the assurance and rate of correct classification occur in earlier parts of the graph where the minimum required sample sizes are smaller. More specifically, in the case where $n_d = \infty$ and $n_d = 1e+5$, a sample size of 285 is associated with an assurance and rate of correct classification of 0.8. In the case where $n_d = \infty$ and $n_a = 1e+4$, a sample size of 325 is required to achieve an assurance and rate of correct classification of approximately 0.82. Overall, there is consistent disparity presented between the two sets of probabilities. As suggested in the reference curve, stronger convergence tends to occur when exact design priors and weak analysis priors are assigned to the study.

3.6.2 Relationship to Frequentist Setting

We can also construct an analogous utility function using a classical-based approach by modifying the cutoff condition to one that adheres to measures pertaining to the frequentist setting. We continue working in a hypothesis testing framework within a general linear model setting outlined in Section 3.6, given as $H_0 : u^\top \beta = c_0$ vs. $H_a : u^\top \beta = c_1$. As previously discussed, the Bayesian method formulates its utility function using the posterior probabilities of H_0 and H_a to determine appropriate sample sizes n_B. In the frequentist setting, we rely on power to characterize the utility function. To achieve a power of $1 - \beta$, the frequentist determines sample size n_F using

$$n_F = (z_\alpha + z_\beta)^2 \left(\frac{\sigma}{\delta}\right)^2, \tag{3.11}$$

where α and β are respectively Type I and Type II error rates and the critical difference is $\delta = c_1 - c_0$.

If we were working within the context of scalar sample means, we could directly showcase overlapping behaviors exhibited between the Bayesian and frequentist settings simply by substituting $n = n_F$ into the utility function, $G_B(n, \mathbf{v})$, given in Equation (3.9). The exact cutoffs are provided in Inoue, Berry, and Parmigiani, 2005. However, because our derived expression of the rate of correct classification in Equation (3.10) is not explicitly expressed in terms of n, we illustrate this relationship computationally through graphs. Note that n is included within the dimension of the design matrix X, where X is $n \times p$.

We identify corresponding pairs of power and r^* values that yield equal sample sizes in the two settings. Fixing the critical difference as $\delta = 0.1$, the significance level as $\alpha = 0.05$, and the variance as $\sigma^2 = 1$, we pass in a sequence of values for Type II error, β, ranging between 0 and 1 to obtain the corresponding set of sample sizes n such that $n = n_B = n_F$. We can subsequently determine the corresponding set of r^* values in the Bayesian setting that are tied to the same set of sample sizes under the same fixed parameter assignments.

Table 3.4: Select set of overlapping sample sizes within the Bayesian and frequentist settings corresponding to varying values of r^* and β.

Overlapping Sample Sizes

$n = n_F = n_B$	r^*	β
1577	0.9765	0.01
1368	0.9677	0.02
1244	0.9610	0.03
1153	0.9552	0.04
1083	0.9500	0.05
1024	0.9451	0.06
974	0.9406	0.07
931	0.9363	0.08
892	0.9322	0.09
857	0.9282	0.10

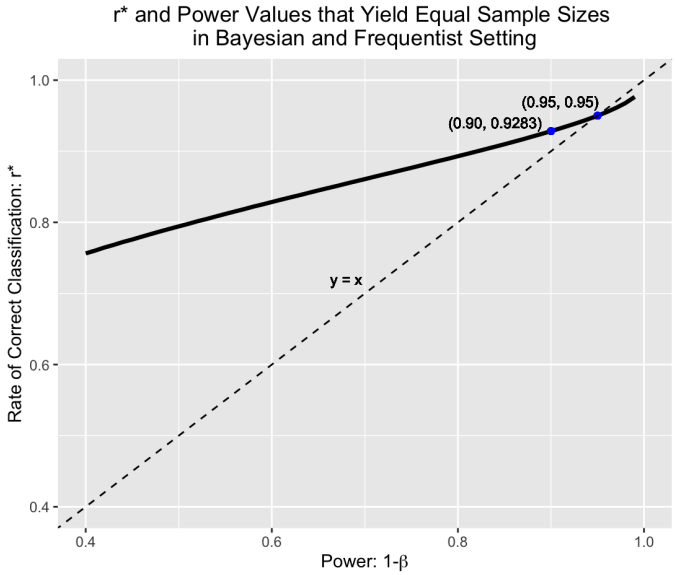


Figure 3.5: Solid curve shows power and r^* values that yield equal sample sizes in the Bayesian and frequentist settings assuming the critical difference is $\delta = 0.1$.

The results are plotted in Figure 3.5, with power on the x-axis and corresponding r^* values on the y-axis. Each point on the solid line indicates a $(1-\beta, r^*)$ point that yield equal sample sizes in the Bayesian and frequentist setting for the parameters specified above. We include a reference line with slope = 1 and observe larger discrepancies between the two values when

power is small. The gap becomes visually smaller as the power increases and the two lines coincide when both the frequentist power and rate of correct classification reach a value of 0.95, which corresponds to a sample size of 1083. This point is marked and labeled in the Figure 3.5. Table 3.4 reports a select set of r^* and β values yielding equal sample sizes in the two settings.

3.7 Discussion

This chapter presents a simulation-based Bayesian design framework for sample size calculations using assurance for deciding in favor of a hypothesis (analysis objective). It is convenient to describe this framework in two stages: (i) the design stage generates data from a population modeled using design priors; and (ii) the analysis stage performs customary Bayesian inference using analysis priors. The frequentist setting emerges as a a special case of the Bayesian framework with highly informative design priors and completely uninformative analysis priors. Our framework can be adapted and applied to a variety of clinical trial settings. Future directions of research and development can entail incorporating more complex analysis objectives into our framework. For example, the investigation of design and analysis priors in the use of Go/No Go settings (Pulkstenis, Patra, and Zhang, 2017), which refers to the point in time at which enough evidence is present to justify advancement to Phase 3 trials, will be relevant. Whether the method of choice involves looking at lengths of posterior credible intervals (Joseph, Berger, and Belisle, 1997) or determining cutoffs that minimize the weighted sum of Bayesian average errors (Reyes and Ghosh, 2013), such conditions are all capable of being integrated as part of our analysis stage objective within the generalized two-stage paradigm.

CHAPTER 4

bayesassurance R Package

In this chapter, we present a **bayesassurance** R package that computes the Bayesian assurance under various settings characterized by different assumptions and objectives. The package offers a constructive set of simulation-based functions suitable for addressing a wide range of clinical trial study design problems. We provide a detailed description of the underlying framework embedded within each of the power and assurance functions and demonstrate their usage through a series of worked-out examples. Through these examples, we hope to corroborate the advantages that come with using a two-stage generalized structure. We also illustrate scenarios where the Bayesian assurance and frequentist power overlap, allowing the user to address both Bayesian and classical inference problems provided that the parameters are properly defined. All assurance-related functions included in this R package rely on a two-stage Bayesian method described in Chapters 2 and 3 that assigns two distinct priors to evaluate the probability of observing a positive outcome, which in turn addresses subtle limitations that take place when using the standard single-prior approach.

4.1 Introduction

To date, there are a several existing R packages pertaining to Bayesian sample size calculation, each adopting different study design approaches. The **SampleSizeMeans** package produced by Lawrence Joseph and Patrick Belisle contains a series of functions used for determining appropriate sample sizes based on various Bayesian criteria for estimating means or differences between means within a normal setting (Joseph and Belisle, 2012). Criteria

considered include the Average Length Criterion, the Average coverage criterion, and the Modified Worst Outcome Criterion (Joseph and Belisle, 1997b; Joseph, Wolfson, and Berger, 1995a). A supplementary package, **SampleSizeProportions**, addresses study designs for estimation of binomial proportions using the same set of criteria listed (Joseph and Belisle, 2009). Our package, also addresses the application of criteria based within the normal and binomial setting, places strong emphasis on a two-stage framework primarily within the context of conjugate Bayesian linear regression models, where we consider the situation with known and unknown variances. The toolkit also offers the flexibility of considering unequal sample sizes for two samples as well as longitudinal study designs.

The **bayesassurance** package contains a collection of functions that can be divided into three categories based on design and usage. These include closed-form solutions, simulation-based solutions, and visualization and/or design purposes. All available functions are categorized and briefly described in Table 4.1. To avoid repetitiveness, we only touch on functions that fall under closed-form and simulation-based categories. The full manuscript that we intend on publishing later this year includes descriptions for all functions embedded in the package. This chapter is organized as follows. Section 4.2 introduces one of the more basic features included in the package that is linked to the fundamental closed-form solution of assurance. This section also introduces the notion of overlapping behaviors exhibited between the Bayesian and frequentist settings, a concept that is frequently brought up and discussed throughout the chapter. Section 4.3 explore simulation-based assurance methods, outlining the statistical framework associated with each method followed by examples worked out in R that users could replicate. Section 4.4 offers some useful graphical features and design matrix generators embedded within the assurance-based functions. In the following sections, we provide a detailed overview for each of the functions available grouped by category followed with worked out examples in R.

Function	Type	Description
`pwr_freq()`	closed-form solution	Returns the statistical power of the specified hypothesis test (either one or two-sided).
`assurance_nd_na()`	closed-form solution	Returns the exact Bayesian assurance for attaining a specified alternative.
`bayes_sim()`	simulation	Approximates the Bayesian assurance of attaining a specified condition for a balanced study design through Monte Carlo sampling.
`bayes_sim_unbalanced()`	simulation	Approximates the Bayesian assurance of attaining a specified condition for an unbalanced study design through Monte Carlo sampling.
`bayes_sim_unknownvar()`	simulation	Same as `bayes_sim` but assumes the variance is unknown.
`bayes_adcock()`	simulation	Determines the assurance of observing that the absolute difference between the true underlying population parameter and the sample estimate falls within a margin of error no greater than a fixed precision level, d.
`bayes_sim_betabin()`	simulation	Returns the Bayesian assurance corresponding to a hypothesis test for difference in two independent proportions.
`pwr_curve()`	visual tool	Constructs a plot with the power and assurance curves overlayed on top of each other for comparison.
`gen_Xn()`	design tool	Constructs design matrix using given sample size(s). Used for power and sample size analysis in the Bayesian setting.
`gen_Xn_longitudinal()`	design tool	Constructs design matrix using inputs that correspond to a balanced longitudinal study design.

Table 4.1: Overview of the functions available for use within the package.

4.2 Closed-form Solution of Assurance

This section goes over functions associated with the fundamental computations of assurance and power. As delineated in Chapter 2, the Bayesian assurance evaluates the tenability of attaining a specified outcome through the implementation of prior and posterior distributions. The `assurance_nd_na()` function computes the exact assurance using a closed-form solution. Suppose we seek to evaluate the tenability of $\theta > \theta_0$ given data from a Gaussian population with mean θ and known variance σ^2. Recalling our framework setup in Chapter 2, we assign two sets of priors for θ, one at the *design stage* and the other at the *analysis stage*. These two stages are the primary components that make up the skeleton of our generalized solution in the Bayesian setting and will be revisited in later sections. The analysis objective specifies the condition that needs to be satisfied. It defines a positive outcome, which serves as an overarching criteria that characterizes the study. In this setting, the analysis objective is to observe $P(\theta > \theta_0|\bar{y}) > 1-\alpha$. The design objective seeks a sample size that is needed to ensure that the analysis objective is met $100\delta\%$ of the time, where δ denotes the assurance.

To ensure the notation used in this section is clear, let $\theta \sim N\big(\theta_1, \frac{\sigma^2}{n_a}\big)$ be our analysis stage prior and $\theta \sim N\big(\theta_1, \frac{\sigma^2}{n_d}\big)$ be our design stage prior, where n_a and n_d are precision parameters that quantify the degree of belief towards parameter θ and the population from which we are drawing samples from to evaluate θ. Then given the likelihood $\bar{y} \sim N\big(\theta, \frac{\sigma^2}{n}\big)$, we can obtain the posterior distribution of θ by multiplying the analysis prior and likelihood:

$$N\left(\theta\middle|\theta_1, \frac{\sigma^2}{n_a}\right) \times N\left(\bar{y}\middle|\theta, \frac{\sigma^2}{n}\right) \propto N\left(\theta\middle|\frac{n_a}{n+n_a}\theta_1 + \frac{n}{n+n_a}\bar{y}, \frac{\sigma^2}{n+n_a}\right) .$$

This posterior distribution gives us $P(\theta > \theta_0|\bar{y})$ and the assurance is then defined as

$$\delta = P_{\bar{y}}\left\{\bar{y} : P(\theta > \theta_0|\bar{y}) > 1-\alpha\right\}.$$

The assurance expression can be expanded out further by using the marginal distribution of

$\bar{y}$, which is obtained by

$$\int N\left(\theta\middle|\theta_1, \frac{\sigma^2}{n_d}\right) \times N\left(\bar{y}\middle|\theta, \frac{\sigma^2}{n}\right) d\theta = N\left(\bar{y}\middle|\theta_1, \left(\frac{1}{n} + \frac{1}{n_d}\right)\sigma^2\right).$$

Since the assurance definition is conditioned on $\bar{y}$, we use this to standardize the assurance expression to obtain the following closed-form solution:

$$\delta(\Delta, n) = \Phi\left(\sqrt{\frac{nn_d}{n+n_d}}\left[\frac{n+n_a}{n}\frac{\Delta}{\sigma} + Z_\alpha\frac{n+n_a}{n}\right]\right). \tag{4.1}$$

The `assurance_nd_na()` function requires the following specified parameters:

1. `n`: sample size (either scalar or vector)
2. `n_a`: precision parameter within the analysis stage that quantifies the degree of belief carried towards parameter θ
3. `n_d`: precision parameter within the design stage that quantifies the degree of belief of the population from which we are generating samples from
4. `theta_0`: initial parameter value provided by the client
5. `theta_1`: prior mean of θ assigned in the analysis and design stage
6. `sigsq`: known variance
7. `alt`: specifies alternative test case, where `alt = "greater"` tests if $\theta_1 > \theta_0$,`alt = "less"` tests if $\theta_1 < \theta_0$, and `alt = "two.sided"` performs a two-sided test for $\theta_1 \neq \theta_0$. By default, `alt = "greater"`.
8. `alpha`: significance level

Consider the following code that loads in the **bayesassurance** package and assigns arbitrary parameters to `assurance_nd_na()` prior to executing the function.

```
R> library(bayesassurance)

R> n <- seq(100, 250, 10)
R> n_a <- 10
R> n_d <- 10
R> theta_0 <- 0.15
R> theta_1 <- 0.25
R> sigsq <- 0.30

R> out <- assurance_nd_na(n = n, n_a = n_a,  n_d = n_d, theta_0 = theta_0,
theta_1 = theta_1, sigsq = sigsq,  alt = "greater", alpha = 0.05)

R> head(out$assurance_table)
R> out$assurance_plot
```

```
     n Assurance
1  100 0.5228078
2  110 0.5324414
3  120 0.5408288
4  130 0.5482139
5  140 0.5547789
6  150 0.5606632
```

Running this block of code will return a table of assurance values and a graphical display of the assurance curve, shown in Figure 4.1. The first six rows of the table are reported in the outputs.

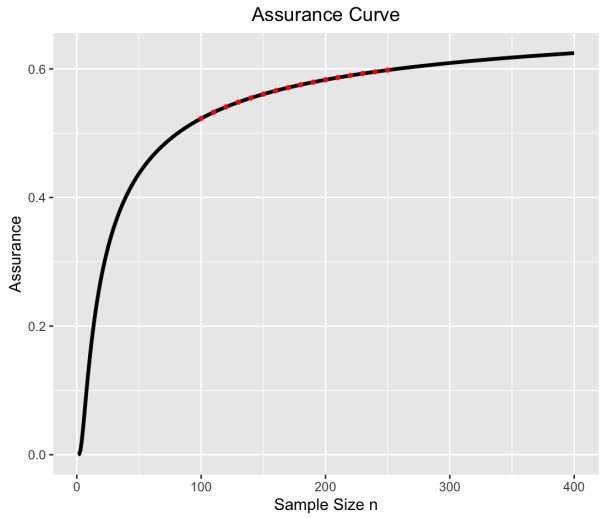


Figure 4.1: Resulting assurance plot with specific points passed in marked in red.

There are a few points worth noting from the above code block as they are pertinent to the vast majority of functions contained within **bayesassurance**. First, we are passing a vector of sample sizes for `n` and saving the results as an arbitrary variable `out`. The list of objects returned by the function is contingent on whether the user passes in a scalar or vector for `n`. If `n` is a scalar, this notifies the program that we only want to determine the assurance for one particular sample size. When this is the case, `out` will only report a single assurance value with no plot. On the other hand, if a vector of sample sizes is passed in to `n`, as is the case for the code sample above, this suggests the user wants to determine the assurance for an array of sample sizes, and the function will produce both a table and an assurance curve showcasing the results. As long as `n` holds a length of at least two, `assurance_nd_na` will create a graphical display of the assurance values for an array of sample sizes surrounding those values of `n` that were passed in, with specific points of interest labeled in red. Figure 4.1 shows the resulting assurance curve corresponding to the code segment above. The graph is created using **ggplot2**, an imported package that **bayesassurance** relies on. Simply typing `out$assurance_table` and `out$assurance_plot` will display the table and plot respectively in this particular set of examples.

4.2.1 Special Case: Convergence with the Frequentist Setting

Depending on how we define the parameters in `assurance_nd_na()`, we could demonstrate the direct relationship held between the Bayesian and classical settings, in which the frequentist power is no more than a special case of the generalized Bayesian solution. This can easily be seen when letting $n_d \to \infty$ and setting $n_a = 0$ in Equation (4.1), i.e. defining a weak analysis stage prior and a strong design stage prior, resulting in

$$\Phi\Big(\sqrt{n}\frac{\Delta}{\sigma} + Z_\alpha\Big), \tag{4.2}$$

which is equivalent to the frequentist power expression that takes the form

$$1 - \beta = P\Big(\bar{y} > \theta_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}\Big) = \Phi\Big(\sqrt{n}\frac{\Delta}{\sigma} + Z_\alpha\Big),$$

The following code chunk demonstrates this special case in R using the `assurance_nd_na()` function:

```
R> library(bayesassurance)

R> n <- seq(10, 250, 5)
R> n_a <- 1e-8
R> n_d <- 1e+8
R> theta_0 <- 0.15
R> theta_1 <- 0.25
R> sigsq <- 0.104

R> out <- assurance_nd_na(n = n, n_a = n_a,  n_d = n_d,
theta_0 = theta_0, theta_1 = theta_1, sigsq = sigsq,
```

```
alt = "greater", alpha = 0.05)

R> head(out$assurance_table)
R> out$assurance_plot
```

```
     n Assurance
1   10 0.2532578
2   15 0.3285602
3   20 0.3981637
4   25 0.4623880
5   30 0.5213579
6   35 0.5752063
```

The **bayesassurance** package includes a `pwr_freq()` function that determines the statistical power of a study design given a set of fixed parameter values that adhere to the closed-form solution of power and sample size. Continuing with the one-sided case, the solution is given by

$$1 - \beta = P\Big(\bar{y} > \theta_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}\Big) = \Phi\Big(\sqrt{n}\frac{\Delta}{\sigma} + Z_{\alpha}\Big), \tag{4.3}$$

where $\Delta = \theta_1 - \theta_0$ denotes the critical difference and Φ denotes the cumulative distribution function of the standard normal. Note this formula is equivalent to the special case of the assurance definition expressed in Equation (4.2).

To execute `pwr_freq()`, the following set of parameters need to be specified:

1. `n`: sample size (either scalar or vector)
2. `theta_0`: initial value specified in the null hypothesis; typically provided by the client

3. `theta_1`: alternative value to test against the initial value; serves as a threshold in determining whether the null is to be rejected or not

4. `alt`: specifies alternative test case, where `alt = "greater"` tests if $\theta_1 > \theta_0$; `alt = "less"` tests if $\theta_1 < \theta_0$; `alt = "two.sided"` performs a two-sided test for $\theta_1 \neq \theta_0$. By default, `alt = "greater"`.

5. `sigsq`: known variance

6. `alpha`: significance level of test

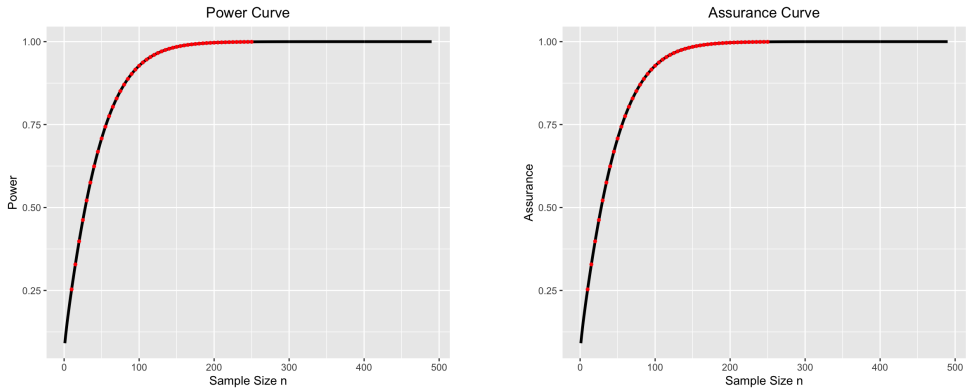
As a simple example, consider the following code chunk that directly runs `pwr_freq()` through specifying the above parameters and loading in **bayesassurance**:

```
R> library(bayesassurance)
R> pwr_freq(n = 20, theta_0 = 0.15, theta_1 = 0.35, sigsq = 0.30,
                alt = "greater", alpha = 0.05)
```

```
"Power: 0.495"
```

Running this simply returns the assurance printed as a statement since we are only evaluating one sample size, `n = 20`. Now consider the next segment of code.

```
R> library(bayesassurance)
R> n <- seq(10, 250, 5)
R> out <- pwr_freq(n = n, theta_0 = 0.15, theta_1 = 0.25, sigsq = 0.104,
alt = "greater", alpha = 0.05)
```



(a) Power curve (b) Assurance curve

Figure 4.2: Resulting power and assurance curve when weak analysis priors and strong design priors and enforced.

```
R> head(out$pwr_table)
R> out$pwr_plot
```

```
     n     Power
1   10 0.2532578
2   15 0.3285602
3   20 0.3981637
4   25 0.4623880
5   30 0.5213579
6   35 0.5752063
```

This above code produces the exact same results as the previous `assurance_nd_na()` example where we assign a weak analysis prior and a strong design prior to demonstrate the overlapping behaviors that occur between the two frameworks.

4.3 Simulation-Based Functions Using Conjugate Linear Models

Each simulation-based function is characterized by a well-defined objective that we seek to evaluate the tenability of. This takes place in the analysis stage. The functions take an iterative approach that alternates between generating a dataset in the design stage and evaluating whether or not the dataset satisfies the analysis stage criteria. The assurance equates to the proportion of datasets that meet the objective.

4.3.1 Assurance Computation with Known Variance

The simulation-based function, `bayes_sim()`, determines the assurance within the context of conjugate Bayesian linear regression models assuming known variance, σ^2. The execution of `bayes_sim()` is straightforward. An important attribute is that users are not required to provide their own design matrix, X_n, when executing `bayes_sim()`. The algorithm automatically accommodates for the case, `Xn = NULL`, using the built-in function, `gen_Xn()`, which constructs appropriate design matrices based on entered sample size(s). Section 4.4 discusses design matrix generators in greater detail.

Setting `Xn = NULL` facilitates calculation of assurances across a vector of sample sizes, where the function sequentially updates the design matrix for each unique sample size undergoing evaluation.

Implementing `bayes_sim()` requires defining the following set of parameters:

1. `n`: Sample size (either vector or scalar). If vector, each value corresponds to a separate study design.

2. `p`: Number of explanatory variables being considered. Also denotes the column dimension of design matrix `Xn`. If `Xn = NULL`, `p` must be specified for the function to assign a default design matrix for `Xn`.

3. `u`: a scalar or vector included in the expression to be evaluated, e.g. $u^\top \beta > C$, where

β is an unknown parameter that is to be estimated.

4. `C`: constant to be compared to

5. `Xn`: design matrix characterizing the observations given by the normal linear regression model $y_n = X_n\beta + \epsilon_n$, where $\epsilon_n \sim N(0, \sigma^2 V_n)$. See above description for details. Default `Xn` is an $np \times p$ matrix comprised of $n \times 1$ ones vectors that run across the diagonal of the matrix.

6. `Vbeta_d`: correlation matrix that characterizes prior information on β in the design stage, i.e. $\beta \sim N(\mu_\beta^{(d)}, \sigma^2 V_\beta^{(d)})$.

7. `Vbeta_a_inv`: inverse-correlation matrix that characterizes prior information on β in the analysis stage, i.e. $\beta \sim N(\mu_\beta^{(a)}, \sigma^2 V_\beta^{(a)})$. The inverse is passed in for computation efficiency, i.e. $V_\beta^{-1(a)}$.

8. `Vn`: an $n \times n$ correlation matrix for the marginal distribution of the sample data y_n. Takes on an identity matrix when set to `NULL`.

9. `sigsq`: a known and fixed constant preceding all correlation matrices `Vn`, `Vbeta_d` and `Vbeta_a_inv`.

10. `mu_beta_d`: design stage mean, $\mu_\beta^{(d)}$

11. `mu_beta_a`: analysis stage mean, $\mu_\beta^{(a)}$

12. `alpha`: specifies alternative test case, where `alt = "greater"` tests if $u^\top \beta > C$, `alt = "less"` tests if $u^\top \beta < C$, and `alt = "two.sided"` performs a two-sided test for $u^\top \beta \neq C$. By default, `alt = "greater"`.

13. `alpha`: significance level

14. `mc_iter`: number of MC samples evaluated under the analysis objective

4.3.1.1 Example 1: Scalar Parameter

The first example evaluates the tenability of $H : u^\top \beta > C$ in the case when β is a scalar. The following code segment assigns a set of arbitrary values for the parameters of `bayes_sim()` and saves the outputs as `assur_vals`. The first ten rows of the table is shown.

```
R> n <- seq(100, 300, 10)

R> assur_vals <- bayesassurance::bayes_sim(n, p = 1, u = 1,
      C = 0.15, Xn = NULL, Vbeta_d = 0, Vbeta_a_inv = 0,
      Vn = NULL, sigsq = 0.265, mu_beta_d = 0.25, mu_beta_a = 0,
      alt = "greater", alpha = 0.05, mc_iter = 5000)

R> head(assur_vals$assurance_table)
R> assur_vals$assurance_plot
```

	Observations per Group (n)	Assurance
1	100	0.6162
2	110	0.6612
3	120	0.6886
4	130	0.7148
5	140	0.7390
6	150	0.7746

We emphasize a few important points in this rudimentary example. Assigning a vector of values for `n` indicates we are interested in assessing the assurance for multiple study designs. Each unique value passed into `n` corresponds to a separate balanced study design

containing that particular sample size for each of the p groups undergoing assessment. In this example, setting `p = 1`, `u = 1` and `C = 0.15` implies we are evaluating the tenability of $H : \beta > 0.15$, where β is a scalar.Furthermore, `Vbeta_d` and `Vbeta_a_inv` are scalars to align with the dimension of β. A weak analysis prior (`Vbeta_a_inv = 0`) and a strong design prior (`Vbeta_d = 0`) are assigned to demonstrate the overlapping scenario taking place between the Bayesian and frequentist settings. Section 4.4 revisits this example when reviewing features that allow users to simultaneously visualize the Bayesian and frequentist settings in a single window. Finally, `Xn` and `Vn` are set to `NULL`, indicating they will each take on the default settings specified in the parameter descriptions above.

4.3.1.2 Example 2: Linear Contrasts

In this example, we revisit the cost-effectiveness application discussed in O'Hagan and Stevens, 2001 to demonstrate a real-world setting. The application considers a randomized clinical trial that compares the cost-effectiveness of two treatments. The cost-effectiveness is evaluated using a net monetary benefit measure expressed as

$$\xi = K(\mu_2 - \mu_1) - (\gamma_2 - \gamma_1),$$

where μ_1 and μ_2 respectively denote the efficacy of treatments 1 and 2, and γ_1 and γ_2 denote the costs. Hence, $\mu_2 - \mu_1$ and $\gamma_2 - \gamma_1$ correspond to the true differences in treatment efficacy and costs, respectively, between Treatments 1 and 2. The threshold unit cost, K, represents the maximum price that a health care provider is willing to pay for a unit increase in efficacy.

In this setting,we seek the tenability of $H : \xi > 0$, which if true, indicates that Treatment 2 is more cost-effective than Treatment 1.To comply with the conjugate linear model framework outlined in Equation (5.3), we set $u = (-K, 1, K, -1)^\top$, $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2)^\top$, and $C = 0$, giving us an equivalent form of $\xi > 0$ expressed as $u^\top \beta > 0$. All other inputs of this application were directly pulled from the paper. The following code sets up the inputs to be

passed into `bayes_sim()`.

```
R> n <- 285
R> p <- 4
R> K <- 20000 # threshold unit cost
R> C <- 0
R> u <- as.matrix(c(-K, 1, K, -1))
R> sigsq <- 4.04^2

## Assign mean parameters to analysis and design stage priors
R> mu_beta_d <- as.matrix(c(5, 6000, 6.5, 7200))
R> mu_beta_a <- as.matrix(rep(0, p))

## Assign correlation matrices (specified in paper)
## to analysis and design stage priors
R> Vbeta_a_inv <- matrix(rep(0, p^2), nrow = p, ncol = p)
R> Vbeta_d <- (1 / sigsq) * matrix(c(4, 0, 3, 0, 0, 10^7, 0,
    0, 3, 0, 4, 0, 0, 0, 0, 10^7), nrow = 4, ncol = 4)

R> tau1 <- tau2 <- 8700
R> sig <- sqrt(sigsq)
R> Vn <- matrix(0, nrow = n*p, ncol = n*p)
R> Vn[1:n, 1:n] <- diag(n)
R> Vn[(2*n - (n-1)):(2*n), (2*n - (n-1)):(2*n)] <- (tau1 / sig)^2 * diag(n)
R> Vn[(3*n - (n-1)):(3*n), (3*n - (n-1)):(3*n)] <- diag(n)
R> Vn[(4*n - (n-1)):(4*n), (4*n - (n-1)):(4*n)] <- (tau2 / sig)^2 * diag(n)
```

The inputs specified above should result in an assurance of approximately 0.70 according to O'Hagan and Stevens, 2001. The `bayes_sim()` returns a similar value, demonstrating that sampling from the posterior yields results similar to those reported in the paper.

```
R> library(bayesassurance)

R> assur_vals <- bayes_sim(n = 285, p = 4,
u = as.matrix(c(-K, 1, K, -1)), C = 0,
Xn = NULL, Vbeta_d = Vbeta_d,
Vbeta_a_inv = Vbeta_a_inv,
Vn = Vn, sigsq = 4.04^2,
mu_beta_d = as.matrix(c(5, 6000, 6.5, 7200)),
mu_beta_a = as.matrix(rep(0, p)),
alt = "greater", alpha = 0.05, mc_iter = 10000)

R> assur_vals


## [1] "Assurance: 0.722"
```

4.3.2 Assurance Computation with Unknown Variance

The `bayes_sim_unknownvar()` function operates similarly to `bayes_sim()` but is used when the variance, σ^2, is unknown, as previously described in Section 2.3.2. In the unknown variance setting, priors are assigned to both β and σ^2 in the analysis stage such that $\beta|\sigma^2 \sim N(\mu_\beta^{(a)}, \sigma^2 V_\beta^{(a)})$ and $\sigma^2 \sim IG(a^{(a)}, b^{(a)})$, where superscripts (a) indicate analysis priors. Determining the posterior distribution of σ^2 requires integrating out β from the joint

posterior distribution of $\{\beta, \sigma^2\}$, yielding

$$\begin{aligned} p(\sigma^2|y_n) &\propto IG(\sigma^2|a^{(a)}, b^{(a)}) \times \int N(\beta|\mu_\beta, \sigma^2 V_\beta) \times N(y_n|X\beta, \sigma^2 V_n) d\beta \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a^{(a)}+\frac{n}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left(b^{(a)} + \frac{c^*}{2}\right)\right\} . \end{aligned} \tag{4.4}$$

Hence, $p(\sigma^2|y_n) = IG\left(\sigma^2|a^*, b^*\right)$, where $a^* = a^{(a)} + \frac{n}{2}$ and $b^* = b^{(a)} + \frac{c^*}{2}$, where $c^* = b^{(a)} + \frac{1}{2}\Big(\mu_\beta^{\top(a)} V_\beta^{-1(a)} \mu_\beta^{(a)} + y_n^\top V_n^{-1} y_n - m_n^\top M_n m_n\Big)$.

Recall the design stage aims to identify a minimum sample size that is needed to attain the assurance level specified by the investigator. We will need the marginal distribution of y_n with priors placed on both β and σ^2. We denote these design priors as $\beta^{(d)}$ and $\sigma^{2(d)}$. With $\sigma^{2(d)}$ now treated as an unknown parameter, the marginal distribution of y_n, given $\sigma^{2(d)}$, under the design prior is derived from $y_n = X_n\beta^{(d)} + e_n$, $e_n \sim N(0, \sigma^{2(d)} V_n)$, $\beta^{(d)} = \mu_\beta^{(d)} + \omega; \quad \omega \sim N(0, \sigma^{2(d)} V_\beta^{(d)})$, where $\beta^{(d)} \sim N(\mu_\beta^{(d)}, \sigma^{2(d)} V_\beta^{(d)})$ and $\sigma^{2(d)} \sim IG(a^{(d)}, b^{(d)})$. Substituting $\beta^{(d)}$ into y_n gives us $y_n = X_n\mu_\beta^{(d)} + (X_n\omega + e_n)$ such that $X_n\omega + e_n \sim N(0, \sigma^{2(d)}(V_n + X_n V_\beta^{(d)} X_n^\top))$. The marginal distribution of $p(y_n|\sigma^{2(d)})$ is

$$y_n|\sigma^{2(d)} \sim N(X_n\mu_\beta^{(d)}, \sigma^{2(d)} V_n^*); \quad V_n^* = X_n V_\beta^{(d)} X_n^\top + V_n. \tag{4.5}$$

Equation (4.5) specifies our data generation model for ascertaining sample size.

4.3.3 Assurance Computation in the Longitudinal Setting

We demonstrate an additional feature embedded in the function tailored to longitudinal data, previously discussed in Section 3.3. In this setting, the variable n no longer refers to the number of subjects but rather the number of repeated measures reported for each subject assuming a balanced study design.

Consider a group of subjects in a balanced longitudinal study with the same number of repeated measures at equally-spaced time points. In the base case, in which time is treated

as a linear term, subjects can be characterized by

$$y_{ij} = \alpha_i + \beta_i t_{ij} + \epsilon_i,$$

where y_{ij} denotes the j^{th} observation of subject i at time t_{ij}, α_i and β_i respectively denote the intercept and slope terms for subject i, and ϵ_i is an error term characterized by $\epsilon_i \sim N(0, \sigma_i^2)$.

In a simple case with two subjects, we can individually express the observations as

$$\begin{aligned}
y_{11} &= \alpha_1 + \beta_1 t_{11} + \epsilon_1 \\
&\vdots \\
y_{1n} &= \alpha_1 + \beta_1 t_{1n} + \epsilon_1 \\
y_{21} &= \alpha_2 + \beta_2 t_{21} + \epsilon_2 \\
&\vdots \\
y_{2n} &= \alpha_2 + \beta_2 t_{2n} + \epsilon_2,
\end{aligned}$$

assuming that each subject contains n observations. The model can also be expressed cohesively using matrices,

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \end{pmatrix}}_{y_n} = \underbrace{\begin{pmatrix} 1 & 0 & t_{11} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & t_{1n} & 0 \\ 0 & 1 & 0 & t_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & t_{2n} \end{pmatrix}}_{X_n} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_2 \end{pmatrix}}_{\epsilon_n} \tag{4.6}$$

bringing us back to the linear model structure. If higher degrees are to be considered for the time variable, such as the inclusion of a quadratic term, the model would be altered

to include additional covariate terms that can accommodate for these changes. In the two-subject case, incorporating a quadratic term for the time variable in Equation (4.6) will result in the model being modified as follows:

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \end{pmatrix}}_{y_n} = \underbrace{\begin{pmatrix} 1 & 0 & t_{11} & 0 & t_{11}^2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & t_{1n} & 0 & t_{1n}^2 & 0 \\ 0 & 1 & 0 & t_{21} & 0 & t_{21}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & t_{2n} & 0 & t_{2n}^2 \end{pmatrix}}_{X_n} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \phi_1 \\ \phi_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_2 \end{pmatrix}}_{\epsilon_n}$$

In general, for m subjects who each have n repeated measures, a one-unit increase in the degree of the time-based covariate will result in m additional columns being added to the design matrix X_n and m additional rows added to the β vector.

When working in the longitudinal setting, additional parameters need to be specified in the `bayes_sim()` function. These include

1. `longitudinal`: logical that indicates the simulation will be based in a longitudinal setting. If `Xn = NULL`, the function will construct a design matrix using inputs that correspond to a balanced longitudinal study design.

2. `ids`: vector of unique subject ids

3. `from`: start time of repeated measures for each subject

4. `to`: end time of repeated measures for each subject

5. `num_repeated_measures`: desired length of the repeated measures sequence.This should be a non-negative number, will be rounded up otherwise if fractional.

6. `poly_degree`: degree of polynomial in longitudinal model, set to 1 by default.

By default, `longitudinal = FALSE` and `ids, from`, and `to` are set to `NULL` when working within the standard conjugate linear model. When `longitudinal = TRUE`, `n` takes on a different meaning as its value(s) correspond to the number of repeated measures for each subject rather than the total number of subjects in each group. When `longitudinal = TRUE` and `Xn = NULL`, `bayes_sim()` implicitly relies on a design matrix generator, `gen_Xn_longitudinal()`, that is specific to the longitudinal setting to construct appropriate design matrices. Section 4.4 discusses this in greater detail.

4.3.3.1 Example 3: Longitudinal Example

The following example uses similar parameter settings as the cost-effectiveness example we had previously discussed in Section 4.3.1.2, now with longitudinal specifications. We assume two subjects and want to test whether the growth rate of subject 1 is different in comparison to subject 2. This could have either positive or negative implications depending on the measurement scale. Figure 4.3 displays the estimated assurance points given the specifications.

Assigning an appropriate linear contrast lets us evaluate the tenability of an outcome. Let us consider the tenability of $u^\top \beta \neq C$ in this next example, where $u = (1, -1, 1, -1)^\top$ and $C = 0$. The timepoints are arbitrarily chosen to be 0 through 120- this could be days, months, or years depending on the context of the problem. The number of repeated measurements per subject to be tested includes values 10 through 100 in increments of 5. This indicates that we are evaluating the assurance for 19 study designs in total. $n = 10$ divides the specified time interval into 10 evenly-spaced timepoints between 0 and 120.

For a more complicated study design comprised of more than two subjects that are divided into two treatment groups, consider testing if the mean growth rate is higher in the first treatment group than that of the second, e.g. if we have three subjects per treatment group, the linear contrast would be set as $u = (0, 0, 0, 0, 0, 0, 1/3, 1/3, 1/3, -1/3, -1/3, -1/3)^\top$.

```
R> n <- seq(10, 100, 5)
```

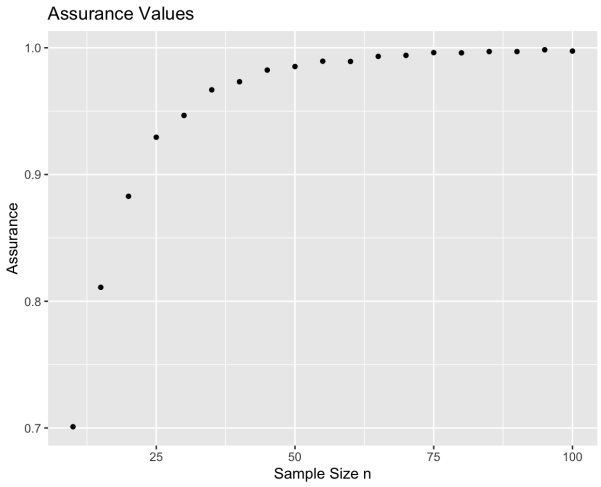


Figure 4.3: Estimated assurance points for longitudinal example.

```
R> ids <- c(1,2)
R> Vbeta_a_inv <- matrix(rep(0, 16), nrow = 4, ncol = 4)
R> sigsq = 100
R> Vbeta_d <- (1 / sigsq) * matrix(c(4, 0, 3, 0, 0, 6, 0, 0, 3,
0, 4, 0, 0, 0, 0, 6), nrow = 4, ncol = 4)

R> assur_out <- bayes_sim(n = n, p = NULL, u = c(1, -1, 1, -1),
C = 0, Xn = NULL, Vbeta_d = Vbeta_d,
Vbeta_a_inv = Vbeta_a_inv, Vn = NULL,
sigsq = 100, mu_beta_d = as.matrix(c(5, 6.5, 62, 84)),
mu_beta_a = as.matrix(rep(0, 4)),
mc_iter = 5000, alt = "two.sided", alpha = 0.05,
longitudinal = TRUE, ids = ids, from = 10, to = 120)

R> head(assur_out$assurance_table)
R> assur_out$assurance_plot

  Observations per Group (n) Assurance
```

1	10	0.6922
2	15	0.8056
3	20	0.8810
4	25	0.9244
5	30	0.9478
6	35	0.9626

4.3.4 Assurance Computation for Unbalanced Study Designs

The `bayes_sim_unbalanced()` function operates similarly to `bayes_sim()` in Section 4.3.1 but estimates the assurance of attaining $u^\top \beta > C$ specifically in an unbalanced design setting. Users provide two sets of sample sizes of equal length, whose corresponding pairs are considered for each study design case. The sample sizes need not be equal to one another, allowing for unbalanced designs. This is unlike `bayes_sim()` that strictly determines assurance for balanced cases, where users specify a single set of sample sizes whose individual entries correspond to a distinct trial comprised of an equal number of observations across all explanatory variables. The `bayes_sim_unbalanced()` function provides a higher degree of flexibility for designing unbalanced studies and offers a more advanced visualization feature. Users have the option of viewing assurance as a 3-D contour plot and assess how the assurance behaves across varying combinations of the two sets of sample sizes that run along the x and y axes.

The `bayes_sim_unbalanced()` function is similar to `bayes_sim()` in terms of parameter specifications with a few exceptions. Parameters unique to `bayes_sim_unbalanced()` are summarized below:

1. `n1`: first sample size (either vector or scalar).

2. `n2`: second sample size (either vector or scalar).

3. `repeats`: an integer value denoting the number of times `c(n1, n2)` is accounted for;

applicable for study designs that consider an even number of explanatory variables greater than two and whose sample sizes correspond to those specified in `n1` and `n2`. By default, `repeats = 1`. See Example 6 below.

4. `surface_plot`: logical parameter that indicates whether a contour plot is to be constructed. When set to `TRUE`, and `n1` and `n2` are vectors, a contour plot (i.e. heat map) showcasing assurances obtained for unique combinations of `n1` and `n2` is produced.

As in `bayes_sim`, it is recommended that users set `Xn = NULL` to facilitate the automatic construction of appropriate design matrices that best aligns with the conjugate linear model described in beginning of Chapter 2. Recall that every unique sample size (or sample size pair) passed in corresponds to a separate study that requires a separate design matrix. Should users choose to provide their own design matrix, it is advised that they evaluate the assurance for one study design at a time, in which a single design matrix is passed into `Xn` along with scalar values assigned for the sample size parameter(s). Saved outputs of the function include

1. `assurance_table`: table of sample size and corresponding assurance values
2. `contourplot`: contour map of assurance values if `surface.plot = TRUE`
3. `mc_samples`: number of Monte Carlo samples that were generated for evaluation

4.3.4.1 Example 4: Unbalanced Assurance Computation with Surface Plot

The following code provides a basic example of how `bayes_sim_unbalanced()` is executed. It is important to check that the parameters passed in are appropriate in dimensions, e.g. `mu_beta_a` and `mu_beta_d` should each contain the same length as that of `u`, and the length of `u` should be equal to the row and column dimensions of `Vbeta_d` and `Vbeta_a_inv`.

A table of assurance values is printed simply by calling `assur_out$assurance_table`, which contains the exact assurance values corresponding to each sample size pair. The

contour plot, shown in Figure 4.4, is displayed using `assur_out$contourplot`, and offers a visual depiction of how the assurance varies across unique combinations of `n1` and `n2`. Areas with lighter shades denote higher assurance levels. No discernible patterns or trends are observed based on the random behavior of the plot and the proximity of values reported in the table as the inputs were arbitrarily chosen with no context. The next example implements the function in a real-world setting that offers more sensible results.

```
R> library(bayesassurance)

R> n1 <- seq(20, 75, 5)
R> n2 <- seq(50, 160, 10)


R> assur_out <- bayes_sim_unbalanced(n1 = n1, n2 = n2,
repeats = 1, u = c(1, -1), C = 0, Xn = NULL,
Vbeta_d = matrix(c(50, 0, 0, 10),nrow = 2, ncol = 2),
Vbeta_a_inv = matrix(rep(0, 4), nrow = 2, ncol = 2),
Vn = NULL, sigsq = 100,  mu_beta_d = c(1.17, 1.25),
mu_beta_a = c(0, 0), alt = "two.sided", alpha = 0.05,
mc_iter = 5000, surface_plot = TRUE)

R> head(assur_out$assurance_table)
R> assur_out$contourplot


  n1  n2 Assurance
1 20  50    0.9504
2 25  60    0.9584
```

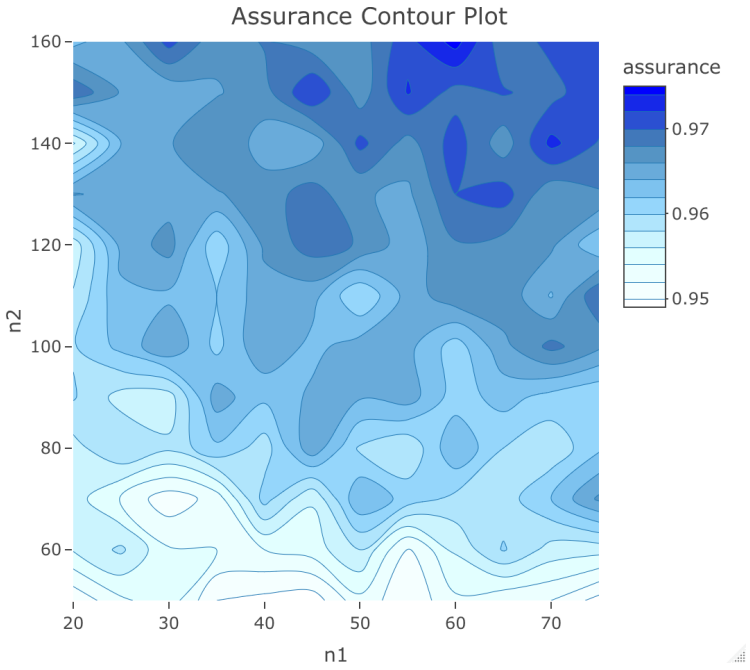


Figure 4.4: Contour map of assurance values with varying sample sizes `n1` and `n2`.

```
3 30  70    0.9508
4 35  80    0.9616
5 40  90    0.9624
6 45 100    0.9634
```

4.3.4.2 Example 5: Cost-effectiveness Application

We revisit the cost-effectiveness problem described in O'Hagan and Stevens, 2001. In addition to providing a 3-D graphical display of the assurance, this example also serves to demonstrate a setting where the `repeats` parameter becomes relevant.

Recall from Example 2 that two distinct sets of efficacy and cost measures are used to compare the cost-effectiveness of treatments 1 and 2. The efficacy and costs are denoted by μ_i and γ_i for $i = 1, 2$ treatments. Hence, the parameter we want to estimate contains four elements tied to the unknown efficacy and costs of treatments 1 and 2, i.e. $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2)^\top$.

It was previously assumed that the treatments contain an equal number of observations, suggesting that the sample sizes across each of the four explanatory variables are also equal. Using `bayes_sim_unbalanced()` offers the added flexibility of constructing an unbalanced study design between treatments 1 and 2. Since the two treatments each contain two components to be measured, we use the `repeats` parameter to indicate that we want two sets of sample sizes, `c(n1, n2)`, passed in, i.e. `c(n1, n2, n1, n2)`. It then becomes clear that our study design consists of n_1 observations for the efficacy and cost of treatment 1, and n_2 observations for those of treatment 2. Figure 4.5 displays a contour plot with a noticeable increasing trend of assurance values across larger sets of sample sizes.

```
R> library(bayesassurance)
R> n1 <- c(4, 5, 15, 25, 30, 100, 200)
R> n2 <- c(8, 10, 20, 40, 50, 200, 250)

R> mu_beta_d <- as.matrix(c(5, 6000, 6.5, 7200))
R> mu_beta_a <- as.matrix(rep(0, 4))
R> K = 20000 # threshold unit cost
R> C <- 0
R> u <- as.matrix(c(-K, 1, K, -1))
R> sigsq <- 4.04^2
R> Vbeta_a_inv <- matrix(rep(0, 16), nrow = 4, ncol = 4)
R> Vbeta_d <- (1 / sigsq) * matrix(c(4, 0, 3, 0, 0, 10^7, 0, 0,
3, 0, 4, 0, 0, 0, 0, 10^7),nrow = 4, ncol = 4)

R> assur_out <- bayes_sim_unbalanced(n1 = n1, n2 = n2, repeats = 2,
                u = as.matrix(c(-K, 1, K, -1)), C = 0, Xn = NULL,
                Vbeta_d = Vbeta_d, Vbeta_a_inv = Vbeta_a_inv,
                Vn = NULL, sigsq = 4.04^2,
```

```
                mu_beta_d = as.matrix(c(5, 6000, 6.5, 7200)),
                mu_beta_a = as.matrix(rep(0, 4)),
                alt = "greater", alpha = 0.05, mc_iter = 5000,
                surface_plot = TRUE)

R> assur_out$assurance_table
R> assur_out$contourplot

   n1  n2 Assurance
1   4   8    0.1614
2   5  10    0.1724
3  15  20    0.3162
4  25  40    0.3942
5  30  50    0.4440
6 100 200    0.6184
7 200 250    0.7022
```

4.4 Visualization Features and Useful Tools

4.4.1 Overlapping Power and Assurance Plots

To facilitate the understanding of the relationship held between Bayesian and frequentist settings, the `pwr_curves()` function produces a single plot with the power curve and assurance points overlayed on top of one another. Recall the primary difference held between `pwr_freq()` and `assurance_nd_na()` is the need to specify additional precision parameters, n_a and n_d, in `assurance_nd_na()`. Knowing that power and sample size analysis in the frequentist setting is essentially a special case to the Bayesian assurance with precision parameters tailored to weak analysis priors and strong design priors, the `pwr_curves()`

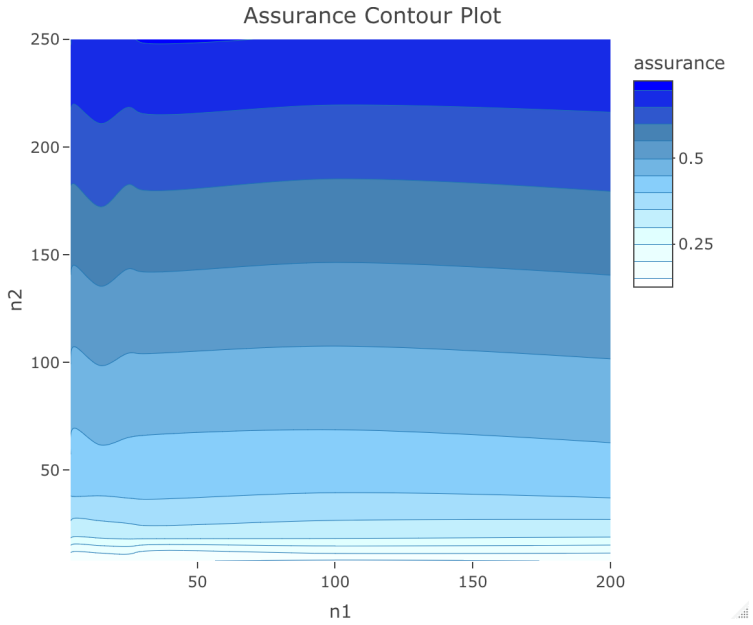


Figure 4.5: Contour map of assurance values in cost-effectiveness application.

function serves as a visualization tool in seeing how varying precision levels' affect assurance values and how these assurance values compare to those of strictly weak analysis and strong design priors, i.e. power values.

The `pwr_curves()` function takes the combined set of parameters presented in `pwr_freq()` and `assurance_nd_na()`, which includes `n, n_a, n_d, theta_0, theta_1, sigsq,` and `alpha`. For further customization, users have the option to include a third set of points in their plot along with the power and assurance curves. These additional points would correspond to the simulated assurance results obtained using `bayes_sim()`. Optional parameters to implement this include

1. `bayes_sim`: logical that indicates whether the user wishes to include simulated assurance results obtained from `bayes_sim()`. Default setting is `FALSE`.

2. `mc_iter`: specifies the number of MC samples to evaluate given `bayes_sim = TRUE`.

The following code segment runs the `pwr_curves()` function using a weak analysis stage prior (`n_a` is set to be small) and a strong design stage prior (`n_d` is set to be large). Implementing

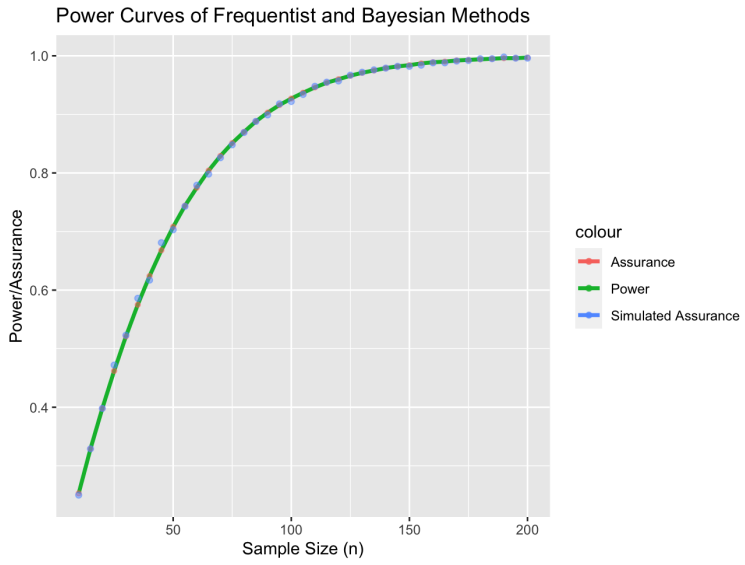


Figure 4.6: Power curve with exact and simulated assurance points for weak analysis prior and strong design prior.

this produces a plot where the assurance points lay perfectly on top of the power curve as shown in Figure 4.6. The simulated assurance points obtained from `bayes_sim()` are also plotted as we set `bayes_sim = TRUE`. These points are highlighted in blue, which lie very close in proximity to those of the exact assurance points highlighted in red. We can also view individual tables of the three sets of points by directly calling them from the saved outputs, e.g. `out$power_table` shows the individual frequentist power values for each sample size. The output we provide shows the first ten rows.

```
R> library(bayesassurance)
```

```
R> out <- pwr_curve(n = seq(10, 200, 10), n_a = 1e-8, n_d = 1e+8,
sigsq = 0.104, theta_0 = 0.15,theta_1 = 0.25, alt = "greater", alpha = 0.05,
bayes_sim = TRUE, mc_iter = 5000)
```

```
R> head(out$power_table)
R> head(out$assurance_table)
```

```
R> out\$plot
```

```
     n     Power
1   10 0.2532578
2   20 0.3981637
3   30 0.5213579
4   40 0.6241155
5   50 0.7080824
6   60 0.7754956
```

```
     n Assurance
1   10 0.2532578
2   20 0.3981637
3   30 0.5213579
4   40 0.6241155
5   50 0.7080824
6   60 0.7754956
```

The next code segment considers the scenario in which both analysis and design stage priors are weak (`n_a` and `n_d` are set to be small). This special case shows how the assurance behaves when vague priors are assigned. Substituting 0 in for both n_a and n_d in Equation (4.1) results in a constant assurance of $\Phi(0) = 0.5$ regardless of the sample size and critical difference. Figure 4.7 illustrates these results, where we have the regular power curve and the flat set of assurance points at 0.5 for both exact and simulated cases. Note that some of these points appear purple due to the overlaps that occur between the exact and simulated assurance values.

```
R> library(bayesassurance)
```

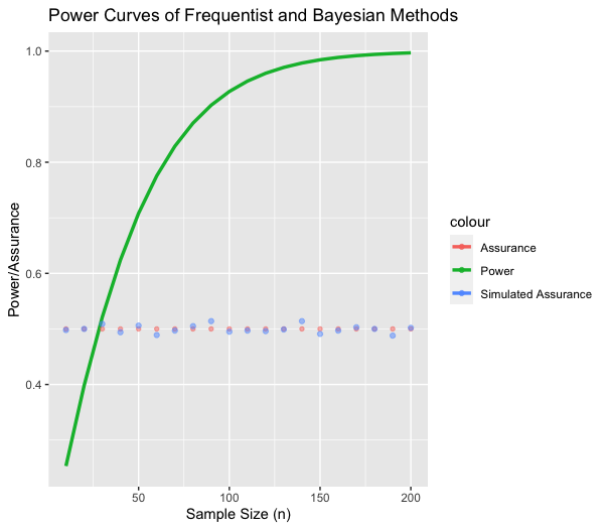


Figure 4.7: Power curve with exact and simulated assurance points for weak analysis and design priors.

```
R> pwr_curve(n = seq(10, 200, 10), n_a = 1e-8, n_d = 1e-8,
sigsq = 0.104, theta_0 = 0.15,theta_1 = 0.25, alt = "greater", alpha = 0.05,
bayes_sim = TRUE, mc_iter = 5000)
```

4.4.2 Design Matrix Generators

The next sections go over design matrix generators that run in the background of selected functions within the **bayesassurance** package when the `Xn` parameter is set to `NULL`. We include these functions in case users wish to see how design matrices are constructed under this particular setting.

4.4.2.1 Standard Design Matrix Generator

The standard design matrix generator, `gen_Xn()`, is relevant to a majority of the simulation-based assurance functions discussed throughout the paper. It is mentioned in Section 4.3

that the assurance function under known variance,`bayes_sim()`, does not require users to specify their own design matrix X_n. Users have the option of setting `Xn = NULL`, which prompts the function to construct a default design matrix using `gen_Xn()` that complies with the general linear model $y_n = X_n\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 V_n)$. The function is automatically administered in the background while `bayes_sim()` is actively in use.

When directly executed in an R script or console, the `gen_Xn()` function takes in a single parameter, `n`, which can either be a scalar or vector. The length of `n` corresponds to the number of groups being assessed in the study design as well as the column dimension of the design matrix, denoted as `p`. Therefore, in general, the resulting design matrix is of dimension $n \times p$. If a scalar value is specified for `n`, the resulting design matrix carries a dimension of $n \times 1$.

In the following example, we pass in a vector of length $p = 4$, which outputs a design matrix of column dimension 4. Each column is comprised of ones vectors with lengths that align with the sample sizes passed in for `n`. The row dimension is therefore the sum of all the entries in `n`. In this case, since the values 1, 3, 5, and 8 are being passed in to `n`, the design matrix to be constructed carries a row dimension of $1 + 3 + 5 + 8 = 17$ and a column dimension of 4.

```
R> n <- c(1,3,5,8)
R> gen_Xn(n = n)

      [,1] [,2] [,3] [,4]
 [1,]    1    0    0    0
 [2,]    0    1    0    0
 [3,]    0    1    0    0
 [4,]    0    1    0    0
 [5,]    0    0    1    0
 [6,]    0    0    1    0
```

```
 [7,]    0    0    1    0
 [8,]    0    0    1    0
 [9,]    0    0    1    0
[10,]    0    0    0    1
[11,]    0    0    0    1
[12,]    0    0    0    1
[13,]    0    0    0    1
[14,]    0    0    0    1
[15,]    0    0    0    1
[16,]    0    0    0    1
[17,]    0    0    0    1
```

The `bayes_sim()` function and its related family of functions generate design matrices using `gen_Xn()` in a very particular way. Each unique value contained in `n` that is passed into `bayes_sim()` corresponds to a distinct study design and thus requires a distinct design matrix. The `gen_Xn()` function interprets each i^{th} component of `n` as a separate balanced study design comprised of n_i participants within each of the p groups, where `p` is a parameter specified in `bayes_sim()`. For example, if we let `Xn = NULL` and pass in `n <- 2`, `p <- 4` for `bayes_sim()`, `gen_Xn()` will process the vector `n <- c(2, 2, 2, 2)` in the background.

Hence, we'd obtain an 8×4 matrix of the form

$$X_n = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

4.4.2.2 Design Matrix Generator in Longitudinal Setting

Section 3.3 describes how the linear model is extended to incorporate time-based covariates within the context of a longitudinal setting. For this special case, a separate function is used to generate design matrices that are appropriate for this setting. The `genXn_longitudinal()` constructs its design matrices differently than `gen_Xn()` and therefore requires a different set of parameter specifications. When the `longitudinal` parameter is set to `TRUE` in `bayes_sim()`, the user is required to specify the following set of parameters, which are directly passed into `genXn_longitudinal()`:

1. `ids`: vector of unique subject ids, usually of length 2 for study design purposes
2. `from`: start time of repeated measures for each subject
3. `to`: end time of repeated measures for each subject
4. `num_repeated_measures`: desired length of the repeated measures sequence. Should be a non-negative number, will be rounded up if fractional.
5. `poly_degree`: degree of polynomial in longitudinal model, set to 1 by default.

Referring back to the model that was constructed for the case involving two subjects, we observe in Equation (4.6) that the design matrix contains vectors of ones within the first half of its column dimension and lists the timepoints for each subject in the second half. Constructing this design matrix requires several components. The user needs to specify subject IDs that are capable of uniquely identifying each individual in the study. Next, the user needs to specify the start and end time as well as the number of repeated measures reported for each subject. The number of repeated measures denotes the number of evenly-spaced timepoints that take place in between the start and end time. Since we are assuming a balanced longitudinal study design, each subject considers the same set of timepoints. Finally, if the user wishes to consider time covariates of higher degrees, such as a quadratic or cubic function, this can be altered using the `poly_degree` parameter, which takes on a default assignment of 1.

In the following code, we pass in a vector of subject IDs and specify the start and end timepoints along with the desired length of the sequence. The resulting design matrix contains vectors of ones with lengths that correspond to the number of repeated measures for each unique subject.

```
R> ids <- c(1,2,3,4)
R> gen_Xn_longitudinal(ids, from = 1, to = 10, num_repeated_measures = 4)

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    0    0    0    1    0    0    0
[2,]    1    0    0    0    4    0    0    0
[3,]    1    0    0    0    7    0    0    0
[4,]    1    0    0    0   10    0    0    0
[5,]    0    1    0    0    0    1    0    0
[6,]    0    1    0    0    0    4    0    0
[7,]    0    1    0    0    0    7    0    0
```

```
 [8,]    0    1    0    0    0   10    0    0
 [9,]    0    0    1    0    0    0    1    0
[10,]    0    0    1    0    0    0    4    0
[11,]    0    0    1    0    0    0    7    0
[12,]    0    0    1    0    0    0   10    0
[13,]    0    0    0    1    0    0    0    1
[14,]    0    0    0    1    0    0    0    4
[15,]    0    0    0    1    0    0    0    7
[16,]    0    0    0    1    0    0    0   10
```

The next code block modifies the previous example to incorporate a quadratic term. Notice there are four additional columns being aggregated to the design matrix. These four columns are obtained from squaring the four columns that precede this set of columns.

```
R> ids <- c(1,2,3,4)
R> gen_Xn_longitudinal(ids, from = 1, to = 10, num_repeated_measures = 4,
poly_degree = 2)

      1 2 3 4  1  2  3  4   1   2   3   4
 [1,] 1 0 0 0  1  0  0  0   1   0   0   0
 [2,] 1 0 0 0  4  0  0  0  16   0   0   0
 [3,] 1 0 0 0  7  0  0  0  49   0   0   0
 [4,] 1 0 0 0 10  0  0  0 100   0   0   0
 [5,] 0 1 0 0  0  1  0  0   0   1   0   0
 [6,] 0 1 0 0  0  4  0  0   0  16   0   0
 [7,] 0 1 0 0  0  7  0  0   0  49   0   0
 [8,] 0 1 0 0  0 10  0  0   0 100   0   0
 [9,] 0 0 1 0  0  0  1  0   0   0   1   0
[10,] 0 0 1 0  0  0  4  0   0   0  16   0
```

```
[11,] 0 0 1 0  0  0  7  0   0   0  49   0
[12,] 0 0 1 0  0  0 10  0   0   0 100   0
[13,] 0 0 0 1  0  0  0  1   0   0   0   1
[14,] 0 0 0 1  0  0  0  4   0   0   0  16
[15,] 0 0 0 1  0  0  0  7   0   0   0  49
[16,] 0 0 0 1  0  0  0 10   0   0   0 100
```

4.5 Discussion

This article introduced **bayesassurance**, a new R package that determines the Bayesian assurance for various conditions using a two-stage framework. The goal of this package is to provide a convenient and user-friendly implementation accessible to a wide range of data analysts, and showcase the generalized aspect of the Bayesian assurance in relation to classical approaches to power and sample size analysis. We hope we have provided organized, well-documented open-source code that can be used to address a wide selection of clinical trial study designs and demonstrate the feasibility of applying Bayesian methods to such problems.

CHAPTER 5

Multiple Comparison Problems using Bayesian FDR Conditions

In this chapter, we investigate the effects of multiple comparison adjustments relative to sample size and assurance. Of particular interest is observing how the number of pairwise tests being conducted affects the assurance under fixed constraints placed on the Bayesian FDR as defined in Müller et al., 2004. For analysis, we study the influences on assurance exhibited by factors such as the number of pairwise comparisons, sample size, and pre-select Bayes FDR threshold values. We assess how our proposed model performs in commonplace large-scale problems, specifically microarray data. Our methodology is hence applied in a study of mammary cancer in the rat, where four distinct patterns of expression are provided (Shepel et al., 1998).

5.1 Introduction

The false discovery rate (FDR) is a powerful metric used to control for false positives that arise in the multiple testing setting. First introduced by Benjamani and Hochberg, 1995 and thus colloquially referred to as the Benjamini-Hochberg procedure, the method has since gained widespread attention and has experienced major developments to address a wider selection of problems.

Modifications extending upon the fundamental concepts of the FDR have been proposed within the last several decades, with substantial contributions made in the Bayesian setting.

Wacholder et al., 2004 demonstrates how the FDR is controlled using a p-value based statistic known as the false positive report probability (FPRP), an approach further investigated by Whittemore, 2007. Storey, 2003 introduces a modified version of the FDR known as the positive false discovery rate (pFDR) and presents a Bayesian analogue to the p-value dubbed as the q-value. A counterpart of the FDR, the false nondiscovery rate (FNR), was introduced by Genovese and Wasserman, 2002 and also includes a method capable of minimizing both the FDR and FNR through proper specification of threshold values. Efron, 2008 and Wen, 2018 draw connections between the Bayesian and frequentist approaches to the FDR. Other related and noteworthy contributions include Efron et al., 2001, Genovese and Wasserman, 2002, Storey and Tibshirani, 2003, Genovese and Wasserman, 2004, Scott and Berger, 2006, among many others.

Multiple comparison problems are commonly applied in genomics when learning about differential gene expression for an immense selection of genes in microarray studies. The FDR is regarded as a practical tool for managing such large-scale data, exhibiting substantial gains in power compared to methods that control for family-wise error rates (Benjamani and Hochberg, 1995) and possessing a higher tendency to identify more true positive associations (Xu, Ciampi, and Greenwood, 2014). There is by now, an ample collection of multiple testing methods specific to microarray analysis, many of which rely on the construction of Bayesian hierarchical models, discussed by Lee et al., 2000, Newton et al., 2001, Kendziorski et al., 2003, Gottardo et al., 2005, and Gelman, Hill, and Yajima, 2012, just to name a few.

Falling in a similar bracket, our proposed method takes on a decision-theoretic approach that casts the multiple testing structure into a conjugate Bayesian linear model framework. The method is characterized by a two-stage method that specifies distinct priors within the design and analysis stages of the study. Details of this method can be found in Pan and Banerjee, 2021a. We adopt a Bayesian interpretation of the FDR from Müller et al., 2004 and establish distinct cases to explore the effects of multiple comparison adjustments on sample size and assurance, the Bayesian-equivalent of statistical power. This is achieved both

through simulation and a real-world application involving microarray data provided by Shepel et al., 1998. Common microarray software uses the FDR to guide gene selection. Sample size determination, in particular, the number of arrays needed to satisfy pre-specified conditions, is a key element to consider in genomics. Several papers explore sample size determination in the context of microarray applications (Pan, Lin, and Le, 2002; Lee and Whitmore, 2002; Zien et al., 2002; Bickel, 2003; Vickerstaff et al., 2019; Tseng and Shao, 2012; Efron, 2007). Müller et al., 2004 chooses the optimal number of microarray replications through loss functions that control for false-positive and false-negative decisions, and Mukherjee et al., 2004 estimates dataset size requirements using empirical learning curves. Our primary objective lies in investigating behaviors exhibited by the assurance and sample size through adjusting criteria constructed under FDR-based constraints using the conjugate linear model.

In Section 5.2, we frame the multiple hypothesis testing problem in the conjugate linear model setting and describe how the Bayesian FDR is incorporated into the estimation of sample size and assurance. Section 5.3 outlines the simulation study with emphasis on defining the design and analysis stage objectives formulated specifically in the context of multiple testing. This is followed by results reported in Section 5.3.1. Section 5.4 applies our method on a real-world application using microarray data that assess differential gene expression for rats with varying susceptibilities to breast cancer. We conclude with a few takeaways and points of discussion in Section 5.5.

5.2 Methodology

This section describes an adjusted linear model that enables conducting multiple comparisons. We tie this back to the two-stage framework and define appropriate design and analysis stage objectives and priors, setting the foundation for estimating the assurance.

5.2.1 Pairwise Hypothesis Testing in Conjugate Linear Model Setting Using Design and Analysis Priors

We construct a linear model taking the form of a one-way ANOVA that includes a collection of distinct groups relative to a single factor. Consider a study design with $i = 1, \cdots, n$ observations in each of the $j = 1, \cdots, J$ groups. Each observation can be characterized by a statistical model expressed as

$$y_{ij} = \mu_j + \epsilon_{ij}; \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \tag{5.1}$$

where y_{ij} denotes the i^{th} observation in the j^{th} group, μ_j denotes the mean in group j, and ϵ_{ij} denotes the error term of the i^{th} observation in the j^{th} group, which are independent and identically normally distributed with mean 0 and known variance σ^2. We can explicitly express (5.1) in matrix form for the full set of observations such that

$$\underbrace{\begin{pmatrix} y_{11} \\ \vdots \\ y_{n1} \\ y_{12} \\ \vdots \\ y_{n2} \\ \vdots \\ y_{1J} \\ \vdots \\ y_{nJ} \end{pmatrix}}_{y} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & \vdots \\ 0 & 1 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_J \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n1} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{n2} \\ \vdots \\ \epsilon_{1J} \\ \vdots \\ \epsilon_{nJ} \end{pmatrix}}_{\epsilon},$$

where the vector of observations Y is of $nJ \times 1$ dimension, design matrix X is of dimension $nJ \times J$, the vector of parameters β is $J \times 1$, and the vector of error terms ϵ is $nJ \times 1$. Hence,

we have the following model:

$$y = X\beta + \epsilon; \quad \epsilon \sim N(0, \sigma^2 V_y), \tag{5.2}$$

where V_y is a known $nJ \times nJ$ correlation matrix.

We want to conduct multiple independent hypothesis tests that compare unique combinations of mean pairs across J subgroups. The linear model structure in (5.1) suggests there are $\binom{J}{2} = \frac{J!}{2!(J-2)!}$ unique pairwise comparisons to be made. In this section, we focus on evaluating a single comparison before shifting into the multiple testing framework. In essence, each hypothesis test can be assessed using a two-stage method based within the conjugate Bayesian linear regression framework (Pan and Banerjee, 2021a). Adhering to the same two-stage structure used in previous chapters, we assign separate priors in the design and analysis stages of the study to fulfill different purposes as well as address limitations that can result from assigning a single prior as described in Chapter 2. The analysis stage aims to construct critical regions for assessing the hypothesis tests while the design stage targets sample size determination.

Consider evaluating the tenability of $H : u_{jj'}^\top \beta > 0$ for all $j \neq j'$, where $u_{jj'} = (0 \cdots 0, \underbrace{1}_{j}, \cdots, \underbrace{-1}_{j'} \cdots 0)^\top$ and $\beta = (\mu_1, \cdots, \mu_j, \cdots, \mu_{j'}, \cdots, \mu_J)^\top$. Referring to the linear regression model expressed in (5.2), we specify an analysis prior on β such that $\beta \sim N\left(\mu_\beta^{(a)}, \tau^2 V_\beta^{(a)}\right)$, where τ^2 is a known scalar, V_β is a known correlation matrix, and superscripts (a) denote analysis stage parameters. The analysis stage prior quantifies uncertainty of the parameter estimate, β. Inference is drawn from the posterior distribution of β such that $p(\beta|y) \propto N(\beta|\mu_\beta^{(a)}, \tau^2 V_\beta^{(a)}) \times N(y|X\beta, \sigma^2 V_y)$, which simplifies to

$$p(\beta|y) \propto N(\beta|Mm, M), \tag{5.3}$$

where $M^{-1} = \sigma^2 V_\beta^{-1(a)} + \tau^2 X^\top V_y^{-1} X$ and $m = \sigma^2 V_\beta^{-1(a)} \mu_\beta^{(a)} + \tau^2 X^\top V_y^{-1} y$. We evaluate the

tenability of H using the $100(1-\alpha)\%$ posterior credible interval,

$$post_{CI} = \left(u_{jj'}^{\top} Mm - Z_{1-\alpha/2}\sqrt{u_{jj'}^{\top} M u_{jj'}}, \quad u_{jj'}^{\top} Mm + Z_{1-\alpha/2}\sqrt{u_{jj'}^{\top} M u_{jj'}} \right), \tag{5.4}$$

in which the realized data y will favor H if y belongs to the set

$$S_{\alpha}(n; y, \sigma^2, \tau^2, \mu_{\beta}^{(a)}, V_{\beta}^{(a)}, V_y) = \left\{ y : u_{jj'}^{\top} Mm > Z_{1-\alpha/2}\sqrt{u_{jj'}^{\top} M u_{jj'}} \right\}. \tag{5.5}$$

This is equivalent to 0 falling below the $100(1-\alpha)\%$ posterior credible interval for $u_{jj'}^{\top}\beta$ in (5.4).

In the design stage, we formulate a data generating mechanism to sample data and evaluate the tenability of $H : u_{jj'}^{\top}\beta > 0$ given realized data y. Evaluating the credibility of H requires specifying the marginal distribution of y. This can be derived by placing a design prior on β such that $\beta \sim N(\mu_{\beta}^{(d)}, \sigma^2 V_{\beta}^{(d)})$, where superscripts (d) denote design stage parameters. The design stage prior reflects our belief about the population from which our realized data is taken. It follows that the marginal distribution of y under the design prior can be derived from

$$y = X\beta + \epsilon; \quad \epsilon \sim N\left(0, \sigma^2 V_y\right) \; ; \quad \beta = \mu_{\beta}^{(d)} + \omega; \quad \omega \sim N\left(0, \sigma^2 V_{\beta}^{(d)}\right) \; .$$

Substituting the equation for β into the equation for y equates to $y = X\mu_{\beta}^{(d)} + (X\omega + \epsilon)$, leading to $y \sim N(X\mu_{\beta}^{(d)}, \sigma^2(XV_{\beta}^{(d)}X^{\top} + V_y))$. In the single testing case, practical Bayesian designs will seek to assure the investigator that the condition in (5.5) will be achieved with a sufficiently high probability such that

$$P_y\left(S_{\alpha}(n; y, \sigma^2, \tau^2, \mu_{\beta}^{(a)}, V_{\beta}^{(a)}, V_y) \right) > \gamma, \tag{5.6}$$

where n denotes the sample size and γ is a pre-specified threshold value. For assessing a single pairwise comparison, the expression in (5.6) evaluates the probability of meeting our specified

objective under the marginal probability distribution of the realized data y corresponding to any sample size n. Choice of sample size will be determined by the smallest value of n that will ensure (5.6) is achieved. We will later understand the importance of this condition and see how the design and analysis objectives are modified to align with the goals of multiple testing.

5.2.2 Bayesian False Discovery Rate for Multiple Testing

Conducting multiple comparisons naturally leads to an increase in the likelihood of making false inferences and wrong conclusions. To motivate the problem in the context of linear regression models as expressed in (5.2), consider a study with $J = 5$ subgroups, indicating there are $k = \frac{5!}{2!(5-2)!} = 10$ statistical tests to be performed. If each two-sided test is independently evaluated at a significance level of $\alpha = 0.05$, the chance of identifying at least one false positive across $k = 10$ tests increases to $1 - (1 - 0.05)^{10} \approx 0.40$. This prompts the need to formulate statistical procedures that not only fulfill specified study objectives but also account for the increased potential of error as more tests are being conducted. Addressing the multiplicity issue has been a widely discussed topic in classical statistical inference, with commonly cited approaches that include adjusting for familywise error rates (FWER) using methods such as the Bonferroni correction (whose conservative properties are explored and tested in various applications described in Bland and Altman, 1995, Armstrong, 2014, and VanderWheele and Mathur, 2019) and Tukey's test (Tukey, 1949), or controlling for Type I error rates under the False Discovery Rate (FDR) metric and its variations (e.g. Wacholder et al., 2004, Storey, 2003, Whittemore, 2007), in which the expected proportion of discoveries that are false is controlled at a fixed threshold. Similar to multiple testing in the classical framework, we consider the multiplicity issue in the Bayesian setting as well. As such, we construct a modified version of the multiple testing approach in the context of conjugate linear regression models that enables us to effectively control for the FDR in the Bayesian setting.

Recall our desire to seek the tenability of $H : u_{jj'}^{\top}\beta > 0$ in Section 5.2.1. For evaluating multiple pairwise tests, our approach is to formulate the multiple comparison problem within the linear model framework that controls for the Bayesian FDR. Let $d_{jj'}$ denote an indicator for the ascertainment of H based on pre-defined decision rules, where $d_{jj'} = 1$ characterizes a "discovery". Let $r_{jj'}$ denote an indicator that represents the true result, where $r_{jj'} = 1$ indicates that H is really true. The FDR is therefore

$$FDR = \frac{\sum_{(j,j')}(1 - r_{jj'})d_{jj'}}{\sum_{(j,j')} d_{jj'}}, \tag{5.7}$$

interpreted as the proportion of false discoveries. Noting that $r_{jj'}$ is the only unknown quantity, let $v_{jj'} = P\left(r_{jj'} = 1|y\right) = P\left(u_{jj'}^{\top}\beta > 0|y\right)$, the posterior probability of H. We use this definition to construct cutoff-based decision rules that will be used for evaluating each comparison. For a pre-specified threshold value $1 - \alpha$, we let

$$d_{jj'} = \begin{cases} 1, & \text{if} \quad v_{jj'} > 1 - \alpha \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

This indicates we will ultimately decide in favor of H if $v_{jj'}$ exceeds the probability threshold $1 - \alpha$, where α is typically set to a small value, e.g. 0.05. We can then re-express Equation (5.7) with respect to this newly defined decision rule,

$$FDR = \frac{\sum_{(j,j')} \mathbb{I}(u_{jj'}^{\top}\beta \leq 0)\mathbb{I}(v_{jj'} > 1 - \alpha)}{\sum_{(j,j')} \mathbb{I}(v_{jj'} > 1 - \alpha)} \tag{5.9}$$

where $\mathbb{I}(\cdot)$ is an indicator function. Adopting the Bayesian analogue of FDR from Müller et al., 2004, the estimated FDR is given as the posterior expectation of the FDR,

$$\overline{FDR} = E\left(FDR|y\right) = \frac{\sum_{(j,j')}\left(1 - v_{jj'}\right)\mathbb{I}\left(v_{jj'} > 1 - \alpha\right)}{\sum_{(j,j')}\mathbb{I}\left(v_{jj'} > 1 - \alpha\right)}. \tag{5.10}$$

The Bayesian FDR will serve a major role for sample size determination within the multiple

testing framework.

5.2.3 Bayesian Assurance and Sample Size Determination for Multiple Testing

Recall the assurance is a Bayesian counterpart of statistical power that evaluates the probability of attaining a specified study objective given the observed data. For standalone linear hypothesis tests described in Section 5.2.1, the objective is observing that 0 falls below the posterior credible interval expressed in Equation (5.4). In this context, the assurance aims to measure the probability of fulfilling this credible interval-based criterion, which we denoted as region $S_\alpha(n; y, \sigma^2, \tau^2, \mu_\beta^{(a)}, V_\beta^{(a)}, V_y)$ in (5.5). Following this notation, the Bayesian assurance for a single hypothesis test is therefore

$$\delta_S = P_y\left(S_\alpha(n; y, \sigma^2, \tau^2, \mu_\beta^{(a)}, V_\beta^{(a)}, V_y)\right). \tag{5.11}$$

In the multiple testing setting, the objective for constructing the assurance needs to be framed with respect to the Bayesian FDR. We develop an analysis plan that evaluates the tenability of the FDR-based objective, which involves implementing a data generating mechanism that samples realized data to estimate the assurance. We adopt the two-stage framework discussed in Section 5.2.1, where separate priors are assigned to address each component of the analysis.

Suppose we conduct pairwise tests for J distinct subgroups. This suggests that we need to individually assess $k = \frac{J!}{2!(J-2)!}$ comparisons, each being pulled from the model in (5.1) through proper specification of vectors u and β. In the analysis stage, the multiple testing framework follows the same procedure as the single hypothesis test case outlined in Section 5.2.1. We assign an analysis prior on β such that $\beta \sim N\left(\mu_\beta^{(a)}, \tau^2 V_\beta^{(a)}\right)$ and conclude that H is attained if the posterior probability of H is met at a sufficiently high probability, specifically, $v_{jj'} > 1 - \alpha$, as outlined in our decision rule in Equation (5.8). Once a conclusion is reached for each of the k comparisons, the Bayesian FDR is subsequently determined using

Equation (5.10). Recall that for the multiple testing framework, we are ultimately interested in controlling for the FDR. It becomes clear then, that our analysis objective in the multiple testing setting is to observe that the Bayesian FDR falls below a pre-specified threshold value, in which

$$\overline{FDR} = \frac{\sum_{(j,j')} (1 - v_{jj'}) \, I \left(v_{jj'} > 1 - \alpha \right)}{\sum_{(j,j')} I \left(v_{jj'} > 1 - \alpha \right)} < \xi, \tag{5.12}$$

where ξ is a threshold Bayesian FDR value.

The design stage in the multiple testing framework starts with the same design strategy applied in the single hypothesis testing case. We assign a design prior for β such that $\beta \sim N\left(\mu_\beta^{(d)}, \sigma^2 V_\beta^{(d)}\right)$, and, as previously derived in Section 5.2.1, generate data from the marginal distribution of y given as $y \sim N\left(X\mu_\beta^{(d)}, \sigma^2\left(XV_\beta^{(d)}X^\top + V_y\right)\right)$. The sampled datasets assess the credible interval condition in (5.5) for each individual hypothesis test before evaluating the overall multiple testing analysis objective in (5.12). At this stage, the above steps only account for evaluating the tenability of (5.12) for a single set of hypothesis tests. To estimate the assurance in the multiple testing setting, the above protocol needs to be executed multiple times for different design stage priors, where each design stage prior produces distinct sets of iterative-drawn data to assess the hypothesis tests. By doing so, we can determine the probability of attaining the analysis objective expressed in (5.12). Hence, the Bayesian assurance in the context of multiple testing is

$$\delta_M = P_y\left(y : \overline{FDR} < \xi\right) = P_y\left\{ y : \frac{\sum_{((j,j')} (1 - v_{jj'}) \, I \left(v_{jj'} > 1 - \alpha \right)}{\sum_{(j,j')} I \left(v_{jj'} > 1 - \alpha \right)} < \xi \right\}. \tag{5.13}$$

The Bayesian assurance function evaluates the probability of attaining the analysis objective under the marginal probability distribution of the realized data corresponding to any sample size n. Choice of sample size will be determined by the smallest value of n that will ensure $\delta_M > \gamma$, where γ is the specified assurance.

5.3 Simulation

We develop a simulation using design and analysis stage priors discussed in Section 5.2. The simulation study consists of two components. The first component involves a function that conducts the primary simulation steps, including evaluating all pairwise hypothesis tests and estimating the Bayesian FDR. Algorithm 5 provides a pseudo-script of the credible interval-based assessments, in which R sets of data are successively drawn to undergo inference. Monte Carlo estimates of $v_{jj'}$'s are determined as the proportion of R datasets that satisfy the credible interval condition derived in (5.5):

$$v_{jj'} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\Big(\{y_r : u_{jj'}^{\top} M^{(r)} m^{(r)} > Z_{1-\alpha/2}\sqrt{u_{jj'}^{\top} M^{(r)} u_{jj'}}\Big),$$

where $\mathbb{I}(\cdot)$ is an indicator function, $M^{(r)}$ and $m^{(r)}$ are the values of M and m computed from dataset y_r. This is repeated for k pairwise comparisons, resulting in k distinct $v_{jj'}$'s characterizing the posterior probabilities of satisfying Equation (5.5). Each $v_{jj'}$ is compared to a pre-specified parameter value to dictate the test result. The result of each comparison is saved as a binary variable, `reject.ind`, with 1 denoting that we decide in favor of $H : u_{jj'}^{\top}\beta > 0$ and 0 suggesting otherwise. Once this is done for all pairwise tests, the expected Bayesian FDR is subsequently determined using Equation (5.10). A pseudo-script of this procedure is provided in Algorithm 6, where `u.dat` is a dataframe containing rows of linear contrasts, $u_{jj'}$, corresponding to each hypothesis test that is to be passed through Algorithm 5 for assessment.

The second component estimates the Bayesian assurance, measuring the probability that the analysis objective is met as expressed in (5.12). Multiple design stage prior means, $\mu_{\beta}^{(d)}$, are passed through Algorithms 5 and 6, where realized data are sampled and evaluated under the analysis objective. That is to say, if we assign q individually assigned $\mu_{\beta}^{(d)}$'s, we end up with q distinct Monte Carlo estimates of the Bayesian FDR corresponding to each design

Table 5.1: Estimated assurance values under different probability thresholds (t) denoting the posterior probability of 0 not falling within the respective credible intervals, and fixed Bayesian FDR thresholds. Assurances increase with larger assigned thresholds.

Assurance Estimates

	$t = 0.6$		$t = 0.7$		$t = 0.8$	
n	$FDR < 0.05$	$FDR < 0.10$	$FDR < 0.05$	$FDR < 0.10$	$FDR < 0.05$	$FDR < 0.10$
10	0.26	0.59	0.45	0.75	0.61	0.90
12	0.32	0.54	0.45	0.87	0.77	0.95
14	0.35	0.76	0.49	0.87	0.85	0.97
16	0.47	0.83	0.69	0.93	0.79	0.97
18	0.59	0.84	0.69	0.96	0.91	1.00
20	0.58	0.93	0.76	0.93	0.92	0.98
22	0.70	0.94	0.82	0.97	0.90	0.98
24	0.70	0.95	0.87	0.98	0.94	0.99
26	0.78	0.93	0.90	0.98	0.98	1.00
28	0.77	0.97	0.88	0.97	0.97	1.00
30	0.78	0.95	0.88	0.97	0.98	0.99
32	0.76	0.96	0.92	0.98	0.96	0.99
34	0.75	0.97	0.89	0.99	0.97	1.00
36	0.82	0.97	0.92	0.99	0.97	1.00
38	0.83	0.97	0.95	0.98	0.98	1.00
40	0.83	0.98	0.92	1.00	0.99	1.00

stage prior. The assurance can then be estimated as the proportion of q sets of design stage priors that meet the FDR-based analysis objective,

$$\delta_M = \frac{1}{q}\sum_{i=1}^{q} \mathbb{I}\Big\{\mu_{\beta}^{(d)}{}_i : \overline{FDR}_i < \xi\Big\}. \tag{5.14}$$

Algorithm 7 provides a pseudo-script of the assurance estimation component, where `mu.dat` is a dataframe containing different design stage prior means in each row.

5.3.1 Simulation Results

We produce assurance curves for different criteria settings characterized by different specifications for the posterior probability cutoff in Equation (5.8) and for the Bayesian FDR

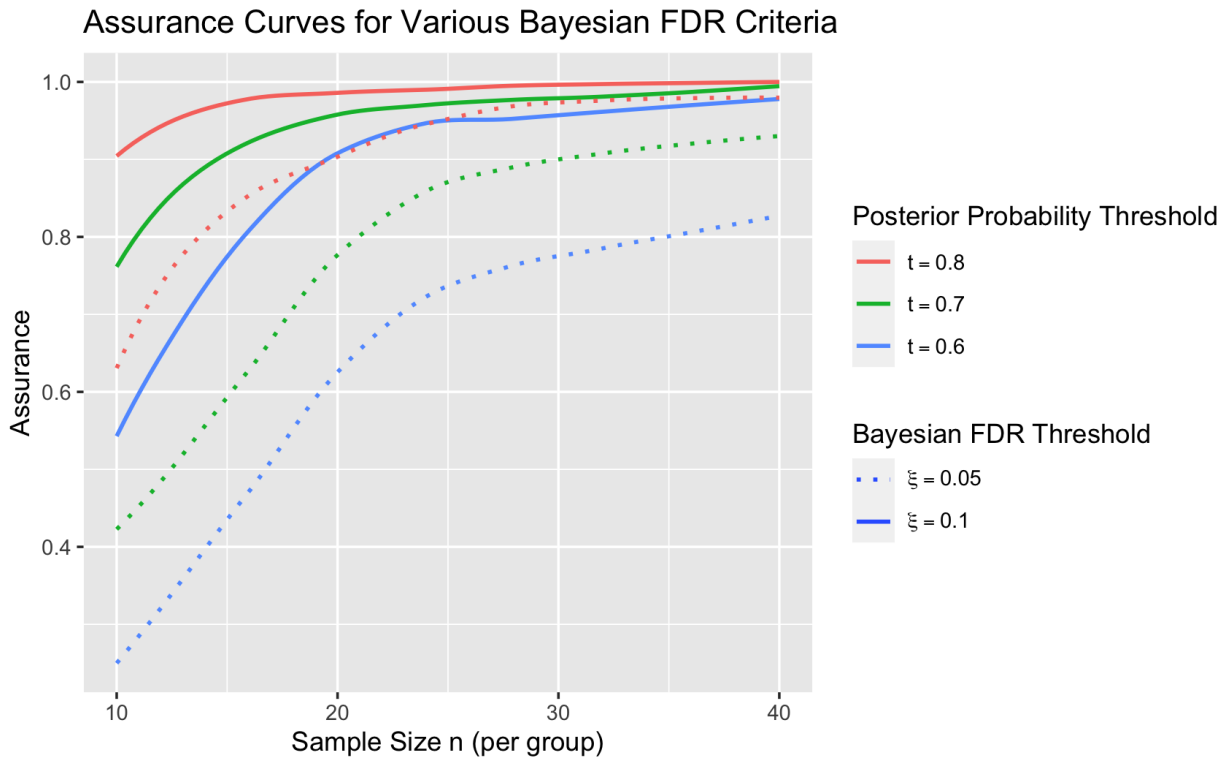


Figure 5.1: Assurance curves for various posterior credible interval thresholds, $t = 0.6, 0.7, 0.8$, and Bayesian FDR thresholds $\xi = 0.05, 0.10$.

condition in Equation (5.12). For conciseness, let t denote the $1-\alpha$ threshold value we had originally specified as the posterior probability cutoff in Equation (5.8) and let ξ continue denoting the Bayesian FDR threshold value. These specifications are used for ascertaining H for distinct comparisons and for estimating the overall assurance.

We assume $J = 5$ groups and specify V_y as a $5n \times 5n$ identity matrix, $\mu_\beta^{(a)}$ as a 5×1 zero vector, $V_\beta^{-1(a)}$ as a 5×5 zero matrix, $V_\beta^{(d)}$ as a 5×5 identity matrix, and X as a $5n \times 5$ design matrix adhering to the structure referenced in Equation (5.1). To estimate the assurance, we generate $q = 100$ design stage prior means from the normal distribution such that ${\mu_\beta^{(d)}}_i \sim N\left((0,0,0,0,0)^\top, I_5\right)$ for $i = 1, \cdots, 100$, where I_5 denotes a 5×5 identity matrix. Each ${\mu_\beta^{(d)}}_i$ is used to sample realized data over $R = 500$ iterations, as outlined in Algorithm 5. We also assign variances $\sigma^2 = \tau^2 = 10$ and set $\alpha = 0.05$.

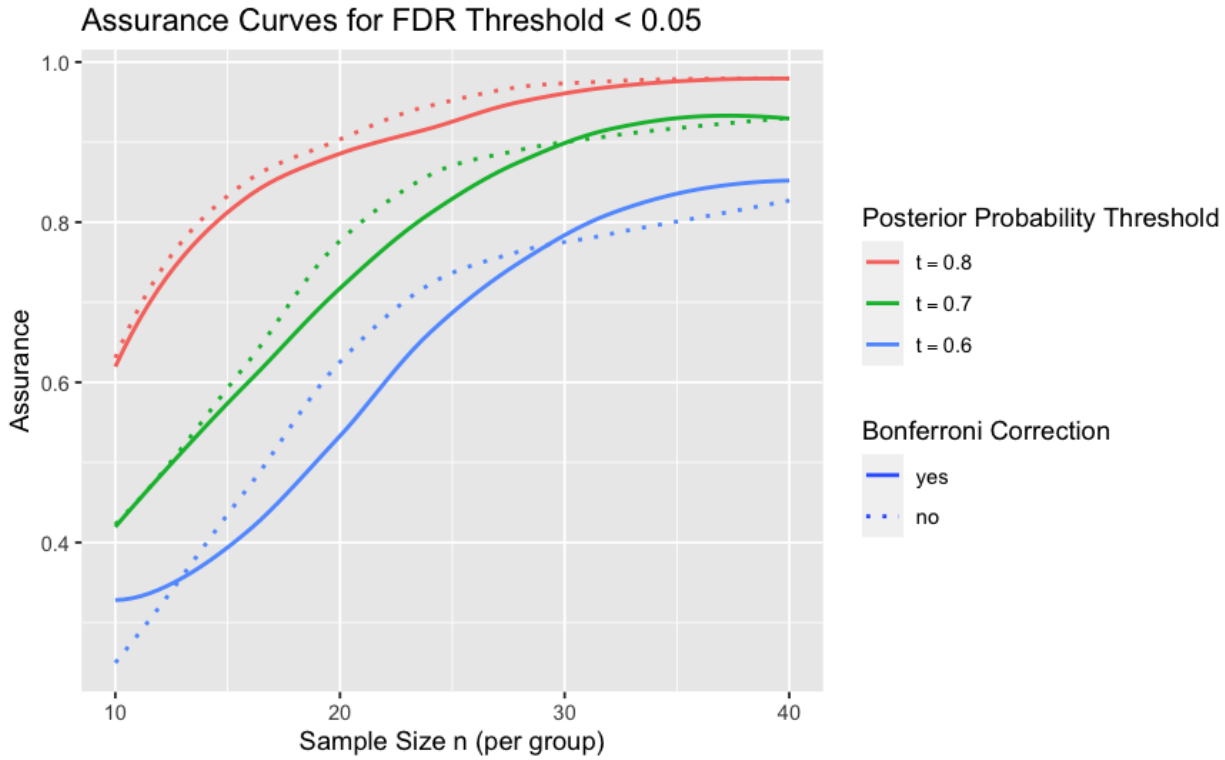


Figure 5.2: Assurance curves for fixed FDR criteria that considers different posterior credible interval criteria and Bonferroni adjustments.

Figure 5.1 displays assurance curves created under different cutoffs assigned for threshold t and ξ. Specifically, t takes one of three values, 0.6, 0.7, or 0.8, and ξ is set to either 0.05 or 0.1, for a total of six cases. Different colors signify different cutoffs for t and different line types correspond to different cutoffs for ξ. Table 5.1 contains specific estimates obtained from separate simulations with corresponding specifications for the distinct cases.

Larger sets of assurance values are observed for higher cutoffs. Higher cutoffs specified for t are associated with stricter minimum posterior probability requirements that verify the tenability of H, resulting in a higher degree of assurance. Additionally, higher cutoffs specified for ξ apply more lenient restrictions on the maximum permitted error rate. For these reasons, it should come as no surprise that the case with the highest cutoffs ($t = 0.8$ and $\xi = 0.1$) outputs the largest assurance values. The Bayesian FDR condition appears to play a larger role in the overall assurance estimations as suggested by the disparity between

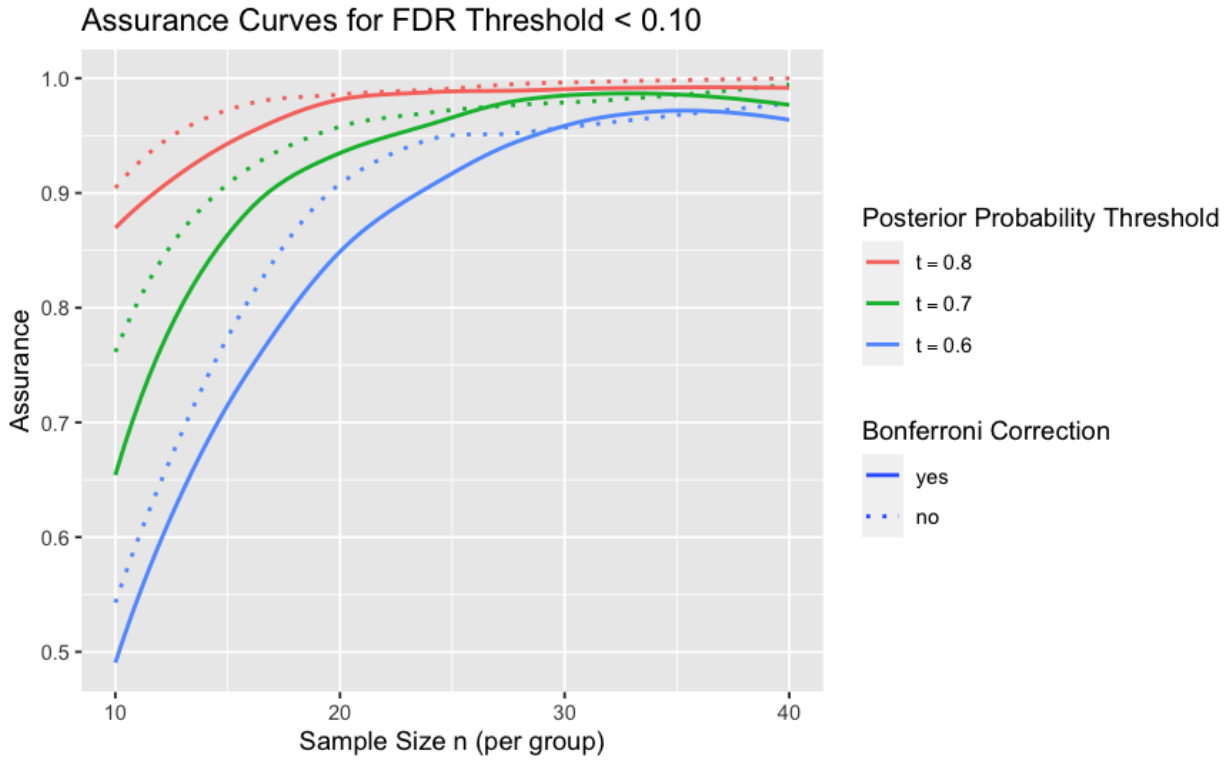


Figure 5.3: Assurance curves for fixed FDR criteria that considers different posterior credible interval criteria and Bonferroni adjustments.

the two sets of curves. Higher assurance curves are associated with a cutoff of $\xi = 0.10$ while lower assurance curves are for $\xi = 0.05$. We also observe an overlap between assurance curves for cases involving $t = 0.6; \xi = 0.1$ and $t = 0.8; \xi = 0.05$.

Figures 5.2 and 5.3 provide visual insight on how the assurance behaves when estimated using Bonferroni corrections for fixed Bayesian FDR cutoffs, i.e. ξ. Solid lines indicate use of the Bonferroni correction and dotted lines denote the non-adjusted cases. Figures 5.2 and 5.3 display look at fixed cutoffs of $\xi = 0.05$ and $\xi = 0.1$ respectively, and Tables 5.2 and 5.3 report the specific estimated values. Both figures show minimal difference in behavior between Bonferroni-adjusted assurance curves and regular assurance curves. Slightly larger assurance values tend to result from applying Bonferroni adjustments in comparison to respective assurance curves that are produced using the same cutoff for t.

Table 5.2: Estimated assurance values under different probability thresholds (t)corresponding to the posterior probability of 0 not falling within the respective credible intervals, and whether or not the Bonferroni adjustment was enforced. The following assurance estimates are based on the analysis objective that the estimated Bayesian FDR falls below 0.05. Overall, the original results are not too different in comparison to Bonferroni-corrected estimates, but it is interesting to note that the original assurance estimates tend to be larger than the Bonferroni-corrected estimates when $n < 30$ and converge in behavior after, as can visually be seen in Figure 5.2.

Assurance Estimates for $FDR < 0.05$

	$t = 0.6$		$t = 0.7$		$t = 0.8$	
n	Original	Bonferroni-Corrected	Original	Bonferroni-Corrected	Original	Bonferroni-Corrected
10	0.26	0.31	0.45	0.41	0.61	0.62
12	0.32	0.38	0.45	0.50	0.77	0.71
14	0.35	0.37	0.49	0.56	0.85	0.81
16	0.47	0.42	0.69	0.57	0.79	0.82
18	0.59	0.42	0.69	0.65	0.91	0.91
20	0.58	0.56	0.76	0.70	0.92	0.83
22	0.70	0.59	0.82	0.82	0.90	0.91
24	0.70	0.66	0.87	0.80	0.94	0.94
26	0.78	0.73	0.90	0.83	0.98	0.95
28	0.77	0.73	0.88	0.85	0.97	0.92
30	0.78	0.78	0.88	0.92	0.98	0.94
32	0.76	0.81	0.92	0.93	0.96	0.99
34	0.75	0.85	0.89	0.95	0.97	0.99
36	0.82	0.84	0.92	0.94	0.97	0.98
38	0.83	0.87	0.95	0.91	0.98	0.98
40	0.83	0.83	0.92	0.93	0.99	0.97

Table 5.3: Estimated assurance values under different probability thresholds (t) corresponding to the posterior probability of 0 not falling within the respective credible intervals, and whether or not the Bonferroni adjustment was enforced. The following assurance estimates are based on the analysis objective that the estimated Bayesian FDR falls below 0.10. Similar to the case when a restriction of $FDR < 0.05$ was implemented, the original assurance estimates tend to be larger than the Bonferroni-corrected estimates when $n < 30$ and converge in behavior after, as can visually be seen in Figure 5.3.

Assurance Estimates for $FDR < 0.10$

	$t = 0.6$		$t = 0.7$		$t = 0.8$	
n	Original	Bonferroni-Corrected	Original	Bonferroni-Corrected	Original	Bonferroni-Corrected
10	0.59	0.49	0.75	0.66	0.9	0.88
12	0.54	0.61	0.87	0.75	0.95	0.88
14	0.76	0.65	0.87	0.83	0.97	0.94
16	0.83	0.75	0.93	0.92	0.97	0.95
18	0.84	0.78	0.96	0.91	1.00	0.97
20	0.93	0.9	0.93	0.93	0.98	0.98
22	0.94	0.88	0.97	0.94	0.98	0.99
24	0.95	0.86	0.98	0.97	0.99	1.00
26	0.93	0.93	0.98	0.97	1.00	0.98
28	0.97	0.97	0.97	0.97	1.00	0.98
30	0.95	0.96	0.97	0.99	0.99	0.99
32	0.96	0.98	0.98	1.00	0.99	1.00
34	0.97	0.96	0.99	0.98	1.00	1.00
36	0.97	0.96	0.99	0.98	1.00	0.99
38	0.97	0.97	0.98	0.98	1.00	0.99
40	0.98	0.97	1.00	0.98	1.00	0.99

5.4 Case Application: Microarray Gene Expression

We implement our methodology in a real-world application using microarray data reported by Shepel et al., 1998, who studied different genetic crosses of rats in an effort to identify potential breast cancer susceptibility genes. The study considers crosses between four distinct inbred lines, including two parental strains with differing susceptibility levels to breast cancer (carcinoma-resistant Copenhagen (COP) rats and carcinoma-sensitive Wistar-Furth (WF) rats) and two offspring congenial lines derived from the parental rat strains. For the purpose of enforcing pairwise comparisons through our linear model, we focus solely on the two offspring congenial lines, denoted as CI and CII. Intensity measurements are obtained for 26,379 genes recorded on 5 CI chips and 2 CII chips. A portion of the data containing 5000 genes can be accessed directly in R by calling `data(gould)` in the `EBarrays` package. For each gene, we are interested in making inference about whether there is differential expression between the two offspring congenial lines.

5.4.1 Case Application Methodology

The overarching goal is to observe how the Bayesian FDR and assurance behave individually and simultaneously as we vary the number of hypothesis tests being assessed. Let $g = 1, \cdots, G$ denote index values for the individual cases, where case g is characterized by comparing g unique gene pairs between congenial lines CI and CII. For clarity, we are comparing intensity measurements of the same respective gene reported in each of the CI and CII chips. A set of G data subsets containing appropriate numbers of unique gene types are pulled from the original dataset. Hence, the g^{th} data subset contains intensity measurements for g specific gene types from the two congenial lines, for a total of $2g$ data entries.

For our analysis, we consider distinct cases involving g gene pairs such that $g = 1, \cdots, 20$. We start by randomly selecting 20 unique gene IDs from the pool of 5000 available gene types in the `EBarrays` R package. Next, we randomly select once more from the set of 20

gene types to construct our first data subset containing two gene expressions corresponding to that particular gene, one for each congenial line. We denote this data subset as d_1. Treating d_1 as our starting point, the remaining data subsets, $d_2, \cdots, d_{20}$, are constructed by cumulatively adding on one gene selected at random without replacement from the set of remaining genes. Hence, each of the smaller datasets are subsets of the larger ones in order to preserve previous results, allowing us to analyze the effects of a cumulatively increasing set of pairwise comparisons.

We refer to the linear model structure in (5.1) to specify our model parameters. Since the microarray application contains subgroups characterized by both the congenial line and gene type, we explicitly define our vector of parameters as $\beta = (\mu_{11}, \cdots, \mu_{1g}, \mu_{21}, \cdots, \mu_{2g})^\top$ to clearly indicate the mean intensities of each gene within the two subgroups. Hence, our case application contains $J = 2g$ subgroups, and we specify V_y as a $nJ \times nJ$ identity matrix, $\mu_\beta^{(a)}$ as a $J \times 1$ zero vector, $V_\beta^{-1(a)}$ as a $J \times J$ zero matrix, and X as a $nJ \times J$ design matrix adhering to the structure referenced in Equation (5.1). The design stage prior parameters are directly determined from the dataset. Specifically, $\mu_\beta^{(d)}$ contains elements corresponding to the mean intensity measurements of each gene, where the mean is taken across 5 CI chips and 2 CII chips. Assuming independence among the set of genes, the diagonal elements of $V_\beta^{(d)}$ correspond to the variances of the intensity measurements of each gene. To estimate the assurance, we generate $q = 100$ design stage prior means from the normal distribution such that ${\mu_\beta^{(d)}}_i \sim N\left(\mu_{\beta^{(d)}}, V_{\beta^{(d)}}\right)$ for $i = 1, \cdots, 100$. Each ${\mu_\beta^{(d)}}_i$ is used to sample realized data over $R = 500$ iterations, similar to our simulation design in Section 5.3. We also assign variances $\sigma^2 = \tau^2 = 1$ and set $\alpha = 0.05$. With the parameters fully specified, we then proceed to compare the intensity measurements for each gene type between the two congenial lines, e.g. we want to assess the tenability of $H : \mu_{11} \neq \mu_{21}$, $\cdots$, $H : \mu_{1g} \neq \mu_{2g}$.

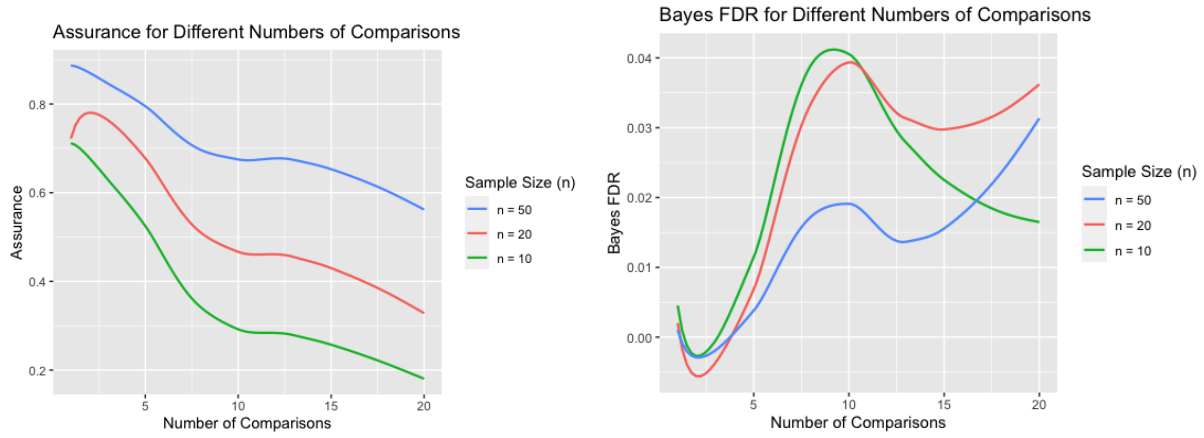


Figure 5.4: Individual behavioral trends exhibited by estimated assurance (left) and Bayesian FDR (right) as the number of comparisons increases and the sample size per subgroup is varied. A 0.01 Bayesian FDR threshold is used for estimating the assurance values displayed on the left. Different colors signify different subgroup sample sizes, n.

5.4.2 Case Application Results

We start by examining how the Bayesian FDR and assurance behave separately in relation to sample size (per subgroup) and number of comparisons. Figure 5.4 provides a side-by-side panel of the estimated assurance and estimated Bayesian FDR as we vary the number of hypothesis tests for a select set of sample sizes. As expected, the assurance showcases a decreasing behavior as the number of conducted hypothesis tests increases, with higher overall assurances exhibited for larger sample sizes. The Bayesian FDR plot is harder to gain insight from as there is no apparent relationship drawn between the Bayesian FDR and the number of comparisons as well as no clear distinction between the different sample sizes enforced. To gain a different perspective, we create Figure 5.5, which displays the estimated Bayesian FDR relative to sample size n for cases that are characterized by the number of pairwise comparisons k, specifically $k = 5, 10$, and 20. We are essentially switching what is originally reported on the x-axis and legend in Figure 5.4. In this modified figure, we observe a generally decreasing trend in Bayesian FDR estimates for all three cases as the sample size n increases, with sharp drops occurring for $n < 10$. For $k = 5$ comparisons, we observe a smooth, downward-sloping curve. This is in contrast to the minor fluctuations seen for

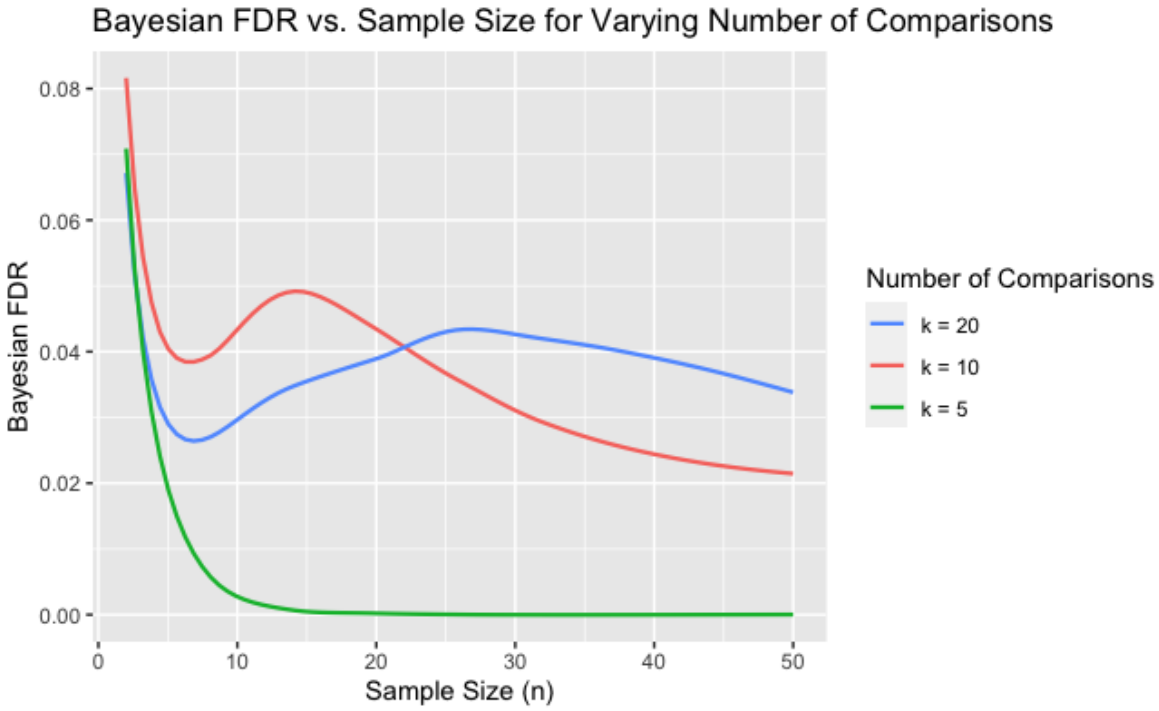


Figure 5.5: Estimated Bayesian FDR as a function of sample size. Different colors signify different numbers of comparisons being conducted.

larger values of k before steadily leveling off, with $k = 20$ leveling off at a later point for larger values of n. This observation could attribute to a larger number of comparisons being assessed, which may result in a slower rate of adjustment prior to experiencing converging behavior. As n becomes larger, providing us with more consistent estimates, we notice a switch in direction in the estimated Bayesian FDR's for cases with $k = 10$ and $k = 20$ comparisons. The case containing $k = 20$ comparisons eventually returns consistently larger estimates of the Bayesian FDR in comparison to the $k = 10$ case, which is to be expected for a higher number of hypothesis tests.

We next examine how the assurance and estimated Bayesian FDR behave simultaneously in relation to one another. Referring to Figure 5.6, the results exhibit monotonically increasing curves that resemble closely to those of regular assurance curves taken across sample sizes n. In general, it makes intuitive sense that the assurance increases as we loosen the restriction

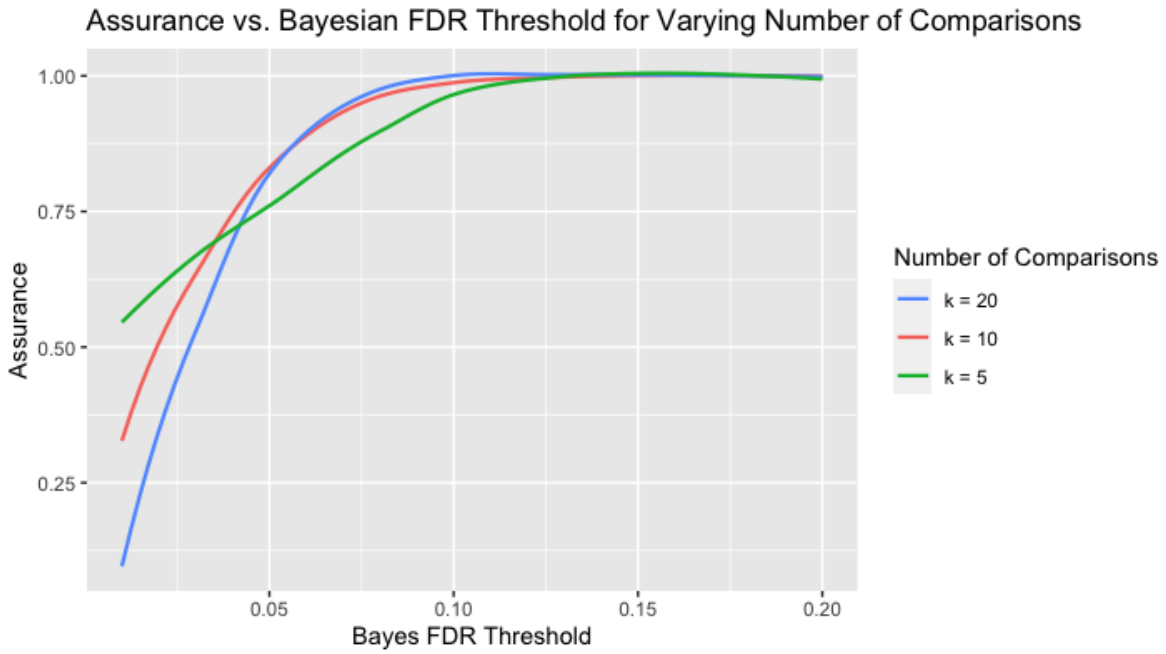


Figure 5.6: Assurance as a function of estimated Bayesian FDR. Different colors signify different numbers of comparisons being conducted.

of the fixed Bayes FDR threshold that needs to be fulfilled. A higher Bayes FDR threshold suggests a higher tolerance for false discovery rates. Table 5.4 reports the estimated Bayesian assurance corresponding to the sequence of Bayes FDR threshold specifications under three cases characterized by number of comparisons. The relationship held between the assurance and Bayes FDR threshold does not appear to have any heavy influences from increasing the number of pairwise comparisons being conducted. The three curves approximately converge in behavior past the 0.05 FDR threshold mark.

5.5 Discussion

This chapter constructs an extended version of the two-stage conjugate Bayesian linear model proposed by Pan and Banerjee, 2021a that provides the added capability of addressing multiple comparison problems using an ANOVA-based approach. By formulating the problem in this way, we gain more control over managing multiplicity concerns through the application

Table 5.4: Estimated assurance values corresponding to different fixed Bayesian FDR threshold values and number of pairwise hypothesis tests.

Assurance Estimates for Various Bayesian FDR Threshold Values

	Number of Pairwise Comparisons		
Bayesian FDR Threshold	$k = 5$	$k = 10$	$k = 20$
0.01	0.54	0.34	0.13
0.02	0.63	0.47	0.24
0.03	0.67	0.61	0.48
0.04	0.72	0.73	0.71
0.05	0.76	0.83	0.83
0.06	0.79	0.90	0.89
0.07	0.83	0.93	0.94
0.08	0.89	0.97	0.99
0.09	0.95	0.98	1.00
0.10	0.97	0.99	1.00
0.11	0.99	1.00	1.00
0.12	0.99	0.98	1.00
0.13	1.00	1.00	1.00
0.14	0.99	1.00	1.00
0.15	1.00	1.00	1.00
0.16	1.00	1.00	1.00
0.17	1.00	1.00	1.00
0.18	1.00	1.00	1.00
0.19	1.00	1.00	1.00
0.20	1.00	1.00	1.00

of our two-stage design. In the analysis stage, we can set restrictions on metrics such as the FDR to directly incorporate the possibility of detecting false discoveries into our model. Under this specified analysis objective, we can then design a data generating mechanism that lets us evaluate this tenability of the specified condition. This in turn, provides us with an estimated assurance value, providing the probability of ensuring that the FDR does not exceed a pre-specified threshold. Hence, our model framework offers the added advantage of producing a study design that simultaneously takes in account both multiplicity control and the Bayesian assurance, providing useful guidance measures to facilitate the process of designing a study involving multiple comparisons.

Our model has demonstrated high feasibility and performance both in simulation studies and real-world data. These example applications provided realistic insights on how the Bayesian FDR and assurance are affected individually as well as to one another in the context of sample size determination. Our model has already demonstrated high favorability in fulfilling a wide selection of research objectives due to its generalized structure. The added capability of addressing multiple comparison problems further enhances its flexibility and universal appeal in the realm of assurance and sample size determination. We eventually want to extend our conjugate linear model's capability to conduct tests that considers more than two subgroups, taking inspiration from Kendziorski et al., 2003, who utilizes Empirical Bayes calculations to make inference about the pattern of differential expression among all four inbred lines reported by Shepel et al., 1998.

CHAPTER 6

Discussion

6.1 Summary and Takeaways

This dissertation addresses the power and sample size analysis problem within a Bayesian setting. Our work revolves around a two-stage framework that distributes the uncertainty scale onto two distinct components, providing the researcher with greater control over precision-based specifications as well as the population from which the data is bring drawn from. We identify several major takeaways to be gained from this dissertation, which entails offering a universal framework that lets users determine an optimal sample size ideal for their own set of conditions, loosening the restrictions of needing to know the population variance in advance, formulating a generalized approach that avoids relying on closed-form solutions, constructing an R package with functions that are in direct alignment to the applications mentioned in this thesis, and introducing extensions to our model that account for multiple testing.

Sample size determination constitutes a major component of study design as the overarching objective of nearly all empirical studies is to make inference about a population from a sample with the intention of saving on cost and time. The question researchers seek to answer is how that optimal sample size is obtained, and under what measures (e.g. target variances, minimum statistical power, credible/confidence interval based specifications) should be considered to determine that optimal sample size, as different studies and experiments contain different set study objectives. There are countless methods and perspectives

on which set of conditions should be given the greatest amount of weight for identifying a minimum sample size that is deemed as optimal, many of which were previously cited in Chapter 1. Throughout this thesis, we refer to various criteria presented in select literature and cast these applications into our two-stage methodology, most of which, primarily took place in the conjugate linear model setting. We use the assurance as a guiding point to aid in determining the sample size, specifying threshold-based criteria that are tied to the positive outcome we wish to assess. From these worked out applications, we find that our method performs very well provided that we establish a well-defined study objective and construct an appropriate data generating mechanism that samples data to evaluate this objective. This is one of the key advantages that our work offers. Rather than identifying an optimal route towards determining the sample size, the two-stage Bayesian method acknowledges that each study carries its own set of associated objectives and should therefore have a tailored approach towards sample size determination.

Apart from its feasibility for user-specified criteria, the two-stage Bayesian approach poses many other advantages. The two-stage methodology addresses convergence-based limitations discussed in Chapter 2, allowing for more coherent interpretations for vague and precise prior specifications. By specifying a model that compensates for error taking place in both parameter estimation and data generation, we can avoid giving vague responses to the investigator in terms of the maximum assurance that is to be expected, e.g. 50% assurance for vague prior specification and 0% or 100% assurance for precise prior specification. Additionally, the two-stage Bayesian framework encompasses a generalized solution nested within the conjugate linear model setting that treats the frequentist setting as a special case under weak analysis stage priors and strong design stage priors. We illustrate these special cases in Chapters 3 and 4, typically demonstrated by overlaying simulated Bayesian points on top of exact frequentust power curves.

Finally, we were able to successfully compile these applications into an R package that is now available on CRAN. The R contains functions that are directly tied to the applica-

tions discussed in this thesis, including hypothesis testing under Normal and Beta-Binomial based conditions, precision-based conditions, and goal functions. The package also contains vignettes with detailed examples that users can follow on their own machine. We make use of helpful visualization tools through the implementation of **ggplot2** that we hope will aid researchers during the planning phase of their study. We discuss some areas of improvement to be made on our package in the next iteration.

6.2 Limitations and Future Direction

Our R package, though thorough, is still in its beginning stages of development. Most of the functions defined in the package follow a similar format, in which users have the option of either specifying their own design and correlation matrices or setting these parameters to `NULL` and have the function automatically generate compatible matrices based on other parameter specifications. Most of our example R code rely on the latter option for the sake of convenience. Otherwise, users would have to manually define their own design and correlation matrices, which can easily translate to a tedious and time-consuming task. A potential solution worth trying involves constructing an interface with prompts the user with a series of questions and formatting tools that aid in specifying their desired matrices. This could entail requesting the dimensions of the matrix followed by an interface that enables the user to directly enter in the specific entries. The R Shiny app is a feasible option to execute this.

To date, we have managed to extend our model to account for conducting multiple pairwise hypothesis tests and formulating the assurance with respect to Bayesian FDR restrictions. So far, we were able to investigate how the assurance and Bayesian FDR are affected for select fixed threshold values pertaining to the credible interval condition and permitted Bayesian FDR (denoted respectively as γ and ξ in Chapter 5) . Currently, these threshold values are chosen arbitrarily and treated as distinct cases for exploratory purposes. Our

next task is to incorporate an additional step that automatically identifies threshold values subject to its relationship with the assurance, taking inspiration from Müller et al., 2004.

Of course, there are many more sample size determination criteria that are yet to be tested and explored using our Bayesian paradigm, posing challenges that we have not been able to acknowledge up until this point. We wish to test our model's capabilities on other applications apart from the linear hypothesis testing setting and offers more relevance to the clinical trial setting. The Go/No-Go Paradigm to proceed into Phase 3 clinical trials is of particular interest (Pulkstenis, Patra, and Zhang, 2017). Modern day machine learning and classification approaches such as Bayesian Additive Regression Trees (Chipman, George, and McCulloch, 2010) is another area worth exploring. Advancing these areas will greatly contribute to the universal applicability and value of our work.

APPENDIX A

Appendix A: Algorithms

A.1 Algorithm 1

Algorithm 1 Bayesian assurance algorithm for known variance

procedure BAYES SIM(n, u, C, X, V_n, $V_\beta^{(d)}$, $V_\beta^{-1(a)}$, $\mu_\beta^{(d)}$, $\mu_\beta^{(a)}$, σ^2, α)
    count = 0 ▷ keeps track of iterations satisfying the analysis objective

    for i in range 1 : max number of iterations **do**
        Design Stage Starts
        $y \leftarrow$ Vector of n values each generated from $\mathrm{N}(X\mu_\beta^{(d)}, \sigma^2(XV_\beta^{(d)}X^\top + V_n))$
        Design Stage Ends

        Analysis Stage Starts ▷ Computes parameters of the β posterior:
        $M \leftarrow (V_\beta^{-1(a)} + X^\top V_n^{-1} X)^{-1}$
        $m \leftarrow V_\beta^{-1(a)}\mu_\beta^{(a)} + X^\top V_n^{-1} y$
        if $\frac{C - u^\top Mm}{\sigma\sqrt{u^\top Mu}} < Z_\alpha$ **then**
            $Z_i \leftarrow 1$
        else
            if $\frac{C - u^\top Mm}{\sigma\sqrt{u^\top Mu}} \geq Z_\alpha$ **then**
                $Z_i \leftarrow 0$
            end if
        end if

```
20:         count ← count + Z_i
21:         Analysis Stage Ends
22:     end for
23:
24:     assurance ← count / max number of iterations
25: return assurance
26:
27: end procedure
```

A.2 Algorithm 2

Algorithm 2 Bayesian assurance algorithm for unknown variance

procedure BAYES SIM2(n, u, C, R, X, V_n, $V_\beta^{(d)}$, $V_\beta^{-1(a)}$, $\mu_\beta^{(d)}$, $\mu_\beta^{(a)}$, σ^2, $a^{(d)}$, $a^{(a)}$, $b^{(a)}$, $b^{(d)}$, α)
    count1 = 0 ▷ counts iterations that meet analysis objective

    for i in range $1 : R$ **do** ▷ R denotes number of generated datasets
        Design Stage Starts
        $\gamma^2 \leftarrow \text{IG}(a^{(d)}, b^{(d)})$
        count2 = 0 ▷ tracks meeting analysis objective for generated data
        $y \leftarrow n \times 1$ vector sampled from $\text{MVN}(X\mu_\beta^{(d)}, \gamma^2(XV_\beta^{(d)}X^\top + V_n))$
        Design Stage Ends

        Analysis Stage Starts
        Compute the components that make up the posterior distributions of β and σ^2:
        $M \leftarrow (V_\beta^{-1(a)} + X^\top V_n^{-1} X)^{-1}$
        $m \leftarrow V_\beta^{-1(a)}\mu_\beta^{(a)} + X^\top V_n^{-1} y$
        $a^* = a^{(a)} + \frac{n}{2}$
        $b^* = b^{(a)} + \frac{1}{2}\{\mu_\beta^{\top(a)} V_\beta^{-1(a)} \mu_\beta^{(a)} + y^\top V_n^{-1} y - m^\top M m\}$

        for j in range 1:J **do** ▷ J denotes number of MCMC posterior samples
            $\sigma^2 \leftarrow \text{IG}(a^*, b^*)$
            $\beta \leftarrow p \times 1$ vector sampled from $\text{MVN}(Mm, \sigma^2 M)$
            if $u^\top \beta \leq C$ **then**
                count2 $\leftarrow$ count2 + 1
            else
                if $u^\top \beta > K$ **then**
                    count2 $\leftarrow$ count2
                end if
            end if
        end for

        if count2 / J $\leq \alpha$ **then**
            count1 = count1 + 1

```
        else
            if count2 / J > α then
                count1 = count1
            end if
        end if
        Analysis Stage Ends

    end for
    assurance ← count1 / R
return assurance

end procedure
```

A.3 Algorithm 3

Algorithm 3 Bayesian assurance algorithm using Adcock's condition for known variance in the univariate case

procedure BAYES ADCOCK(n, d, $\theta_0^{(a)}$, $\theta_0^{(d)}$, n_a, n_d, σ^2, α)
    count = 0 ▷ keeps track of the iterations that satisfy the analysis obj

    maxiter = 1000 ▷ arbitrary number of iterations to loop thru
    for i in range 1 : maxiter **do**
        Design Stage Starts
        $\sigma_d^2 \leftarrow \sigma^2 \frac{n_d+n}{nn_d}$
        $\bar{x} \leftarrow$ single value generated from $N(\theta_0^{(d)}, \sigma_d^2)$
        Design Stage Ends

        Analysis Stage Starts ▷ Computes components of the posterior distribution
        $\lambda \leftarrow \frac{n_a\theta_0^{(a)}+n\bar{x}}{n_a+n}$
        $\sigma_a^2 \leftarrow \frac{\sigma^2}{n_a+n}$
        $\theta \leftarrow$ single value generated from $N(\lambda, \sigma_a^2)$

        $\phi_1 \leftarrow \frac{\sqrt{n_a+n}}{\sigma}(\theta + d - \lambda)$
        $\phi_2 \leftarrow \frac{\sqrt{n_a+n}}{\sigma}(\theta - d - \lambda)$

        if $\Phi(\phi_1) - \Phi(\phi_2) \geq 1 - \alpha$ **then**
            $Z_i \leftarrow 1$
        else
            if $\Phi(\phi_1) - \Phi(\phi_2) < 1 - \alpha$ **then**
                $Z_i \leftarrow 0$
            end if
        end if

        count $\leftarrow$ count + Z_i
        Analysis Stage Ends
    end for

    assurance $\leftarrow$ count / maxiter
return assurance

end procedure

A.4 Algorithm 4

Algorithm 4 Bayesian assurance algorithm for difference in two independent proportions under Pham-Gia's credible interval condition

procedure BAYES PHAM-GIA(n_1, n_2, $p_1 = \text{NULL}$, $p_2 = \text{NULL}$, α_1, α_2, β_1, β_2, α)
  if $\psi = 1$ **then**
    $p_1 \leftarrow$ single value generated from Unif$[p_1, p_1]$
    $p_2 \leftarrow$ single value generated from Unif$[p_2, p_2]$
  else if $\psi = 0$ **then**
    $p_1 \leftarrow \text{Beta}(\alpha_1, \beta_1)$
    $p_2 \leftarrow \text{Beta}(\alpha_2, \beta_2)$
  end if

  count = 0 ▷ keeps track of the iterations that satisfy the analysis obj
  maxiter = 1000 ▷ arbitrary number of iterations to loop thru

  for i in range 1 : maxiter **do**
    Design Stage Starts
    $x_1 \leftarrow$ single value generated from $Bin(n_1, p_1)$
    $x_2 \leftarrow$ single value generated from $Bin(n_2, p_2)$
    Design Stage Ends

    Analysis Stage Starts
    $p_{\text{post}} = \frac{\alpha_1+x_1}{\alpha_1+\beta_1+n_1} - \frac{\alpha_2+x_2}{\alpha_2+\beta_2+n_2}$ ▷ Computes posterior parameters of $p = p_1 - p_2$
    $\text{var}(p)_{\text{post}} = \frac{(\alpha_1+x_1)(\beta_1+n_1-x_1)}{(\alpha_1+\beta_1+n_1)^2(\alpha_1+\beta_1+n_1+1)} + \frac{(\alpha_2+x_2)(\beta_2+n_2-x_2)}{(\alpha_2+\beta_2+n_2)^2(\alpha_2+\beta_2+n_2+1)}$

    $\text{lb} = p_{\text{post}} - z_{1-\alpha/2}\sqrt{var(p)_{\text{post}}}$ ▷ Computes upper and lower bounds
    $\text{ub} = p_{\text{post}} + z_{1-\alpha/2}\sqrt{var(p)_{\text{post}}}$

    if 0 < lb or 0 > ub **then**
      count ← count + 1
    end if
    Analysis Stage Ends
  end for

  assurance ← count / maxiter
return assurance

end procedure

A.5 Algorithm 5

Algorithm 5 Returns the posterior probability of 0 not being contained within the credible interval bands

procedure POST.PROB.FUNC(n, u, X, $\mu_\beta^{(a)}$, $\mu_\beta^{(d)}$, σ^2, τ^2, V_y, $V_\beta^{-1(a)}$, $V_\beta^{(d)}$, alt, α, R)
    count = 0 ▷ counts datasets that meet condition
    for r in range $1 : R$ **do**
        $y_r \leftarrow$ data generated from $\mathrm{N}\Big(X_n\mu_\beta^{(d)}, \frac{\tau^2}{n}X_nV_\beta^{(d)}X_n^\top + \frac{\sigma^2}{n}V_y\Big)$

        $M^{(r)} \leftarrow \Big(\frac{\sigma^2}{n}V_\beta^{-1(a)} + \frac{\tau^2}{n}X_n^\top V_y^{-1}X_n\Big)^{-1}$ ▷ solve by Cholesky decomposition
        $m^{(r)} \leftarrow \frac{\sigma^2}{n}V_\beta^{-1(a)}\mu_\beta^{(a)} + \frac{\tau^2}{n}X_n^\top V_y^{-1}y_r$

        Determine upper and lower credible interval bounds for each alternative case
        if alt = "two.sided" **then** ▷ tests if $u^\top\beta \neq 0$
            lb $\leftarrow u^\top M^{(r)}m^{(r)} - Z_{1-\alpha/2}\sqrt{u^\top M^{(r)}u}$
            ub $\leftarrow u^\top M^{(r)}m^{(r)} + Z_{1-\alpha/2}\sqrt{u^\top M^{(r)}u}$

            if $0 <$ lb $\mid 0 >$ ub **then**
                $Z_i \leftarrow 1$
            else
                $Z_i \leftarrow 0$
            end if

        else if alt = "lower" **then** ▷ tests if $u^\top\beta < 0$
            ub $\leftarrow u^\top M^{(r)}m^{(r)} + Z_{1-\alpha}\sqrt{u^\top M^{(r)}u}$
            if $0 >$ ub **then**
                $Z_i \leftarrow 1$
            else
                $Z_i \leftarrow 0$
            end if

        else if alt = "greater" **then** ▷ tests if $u^\top\beta > 0$
            lb $\leftarrow u^\top M^{(r)}m^{(r)} - Z_{1-\alpha}\sqrt{u^\top M^{(r)}u}$
            if $0 <$ lb **then**
                $Z_i \leftarrow 1$
            else
                $Z_i \leftarrow 0$
            end if
        end if

```
        count ← count + Z_i
    end for

    v ← count / R
return v

end procedure
```

A.6 Algorithm 6

Algorithm 6 Returns the Bayesian FDR using Algorithm 5

```
procedure BAYES.FDR(u.dat, γ, ···)
    count ← 0                                          ▷ number of comparisons surpassing γ
    numerator ← 0                                      ▷ initializes numerator based on Eq. (5.10)
    k ← row dimension of u.dat                         ▷ number of comparisons

    for i in range 1 : k do
        v[i] ← POST.PROB.FUNC(u = u.dat[i, ], ···)     ▷ u.dat rows passed into Alg. 5
                                                       ▷ other fixed parameters also passed into function

        if v[i] > γ then
            reject.ind[i] ← 1                          ▷ indicates rejecting H_0
            numerator ← numerator + [reject.ind[i] * (1 - v[i])]   ▷ numerator updated
            count ← count + 1
        end if
    end for

    if count > 0 then
        bayes.fdr ← numerator / count

    else
        bayes.fdr ← 0
end if
end procedure
```

A.7 Algorithm 7

Algorithm 7 Determines assurance using Algorithms 5 and 6

procedure BAYES.ASSURANCE(mu.dat, u.dat, ξ, γ, $\cdots$)
    count $\leftarrow$ 0 ▷ number of trials that fall below FDR threshold ξ
    $q \leftarrow$ row dimension of mu.dat ▷ number of trials used to estimate assurance

    for i in range 1 : q **do**
        fdr[i] $\leftarrow$ BAYES.FDR(u = u.dat[i,], $\mu_\beta^{(d)}$ = mu.dat[i,], $\gamma = \gamma$, $\cdots$)

        if fdr[i] $< \xi$ **then**
            count $\leftarrow$ count + 1
        end if
    end for

    assurance $\leftarrow$ count / q

end procedure

APPENDIX B

Appendix B: Derivations and Specifications

B.1 Special Case Explanation in Section 2.3.1

We assume $\beta \sim N(\beta_1, \sigma^2/n_a)$ in the analysis stage and $\beta \sim N(\beta_1, \sigma^2/n_d)$ in the design stage, where $\beta_1 > \beta_0$. The data will favor H if the sample mean lies in $A_\alpha(\beta_0, \beta_1)$, where

$$A_\alpha(\beta_0, \beta_1) = \left\{ \bar{y} : \bar{y} > \beta_0 - \frac{n_a}{n}(\beta_1 - \beta_0) - \sqrt{\left(1 + \frac{n_a}{n}\right)} \frac{\sigma}{\sqrt{n}} Z_\alpha \right\} .$$

Using the design prior, we obtain the marginal distribution $\bar{y} \sim N\left(\beta_1, \left(\frac{1}{n} + \frac{1}{n_d}\right)\sigma^2\right)$. We use this distribution to calculate $\delta(n) = P_{\bar{y}}\{\bar{y} : P(\theta < \theta_0 \,|\, \bar{y}) < \alpha\}$, which produces a closed-form expression for Bayesian assurance:

$$\delta(\Delta, n, n_a, n_d) = \Phi\left(\sqrt{\frac{n n_d}{n + n_d}} \left[\frac{n + n_a}{n}\frac{\Delta}{\sigma} + Z_\alpha \frac{\sqrt{n + n_a}}{n}\right]\right) , \tag{B.1}$$

where $\Delta = \beta_1 - \beta_0$. As $n_d \to \infty$ and $n_a \to 0$, we obtain that

$$\lim_{n_a \to 0, n_d \to \infty} = \Phi\left(\sqrt{n}\frac{\Delta}{\sigma} + Z_\alpha\right) ,$$

which is precisely the frequentist power curve. Therefore, the frequentist sample size emerges as a special case of the Bayesian sample size when the design prior becomes perfectly precise and the analysis prior becomes perfectly uninformative.

B.2 Design Prior Specifications in O'Hagan and Stevens (2001) in Section 3.1

O'Hagan and Stevens (2001) assign mean and variance design priors $\mu_\beta^{(d)} = (5, 6000, 6.5, 7200)^\top$ and $V_\beta^{(d)} = \begin{pmatrix} 4 & 0 & 3 & 0 \\ 0 & 10^7 & 0 & 0 \\ 3 & 0 & 4 & 0 \\ 0 & 0 & 0 & 10^7. \end{pmatrix}$. We factor out σ^2 in the simulation study to adhere to the conjugate Bayesian formulation such that $\sigma^2 V_\beta^{(d)} = \sigma^2 \begin{pmatrix} 4/\sigma^2 & 0 & 3 & 0 \\ 0 & 10^7/\sigma^2 & 0 & 0 \\ 3 & 0 & 4/\sigma^2 & 0 \\ 0 & 0 & 0 & 10^7/\sigma^2 \end{pmatrix}$.

B.3 Precision-Based Analysis Stage Objective in Section 3.4

The following expression denotes the assurance under precision-based conditions:

$$\delta = P_{\bar{x}}\{\bar{x} : P(|\bar{x} - \theta| \leq d) \geq 1 - \alpha\} \,. \tag{B.2}$$

We focus from the analysis objective given in the expression $P(|\bar{x} - \theta| \leq d)$. The posterior of θ can be obtained by taking the product of the prior and likelihood, giving us

$$N\left(\bar{x}\middle|\theta, \frac{\sigma^2}{n}\right) \times N\left(\theta\middle|\theta_0^{(a)}, \frac{\sigma^2}{n_a}\right) = N\left(\theta\middle|\lambda, \frac{\sigma^2}{n_a + n}\right), \tag{B.3}$$

where $\lambda = \frac{n\bar{x} + n_a\theta_0^{(a)}}{n_a + n}$. From here we can further evaluate the condition using parameters from the posterior of θ to obtain a more explicit version of the analysis stage objective. Starting from $P(|\bar{x} - \theta| \leq d) = P(\bar{x} - d \leq \theta \leq \bar{x} + d)$, we can standardize all components of the inequality using the posterior parameter values of θ, leading us to

$$\begin{aligned} P(|\bar{x} - \theta| \leq d) &= P\left(\frac{\bar{x} - d - \lambda}{\sigma/\sqrt{n_a + n}} \leq \frac{\theta - \lambda}{\sigma/\sqrt{n_a + n}} \leq \frac{\bar{x} + d - \lambda}{\sigma/\sqrt{n_a + n}}\right) \\ &= P\left(\frac{\bar{x} - d - \lambda}{\sigma/\sqrt{n_a + n}} \leq Z \leq \frac{\bar{x} + d - \lambda}{\sigma/\sqrt{n_a + n}}\right). \end{aligned}$$

Simplifying the result gives us our analysis stage objective:

$$\left\{\bar{x} : \Phi\left[\frac{\sqrt{n_a + n}}{\sigma}(\bar{x} + d - \lambda)\right] - \Phi\left[\frac{\sqrt{n_a + n}}{\sigma}(\bar{x} - d - \lambda)\right] \geq 1 - \alpha\right\}. \tag{B.4}$$

B.4 Simulation Results under Precision-Based Conditions in Section 3.4

We test our algorithm using different fixed precision parameters d with varying sample sizes n. The remaining fixed parameters, including σ^2, $\theta_0^{(a)}$, and $\theta_0^{(d)}$, are randomly drawn from the uniform distribution Unif(0, 1) for simplicity. Figure B.1 displays the results of the Bayesian-simulated points (marked in blue) in the case where weak analysis stage priors were assigned overlayed on top of the frequentist results (marked in red). Note that the Bayesian-simulated points denote the probability of observing that the posterior of θ differing from the sample mean $\bar{x}$ within a range of $\bar{x} \pm d$ exceeds $1 - \alpha$. In general, these probabilities are obtained by

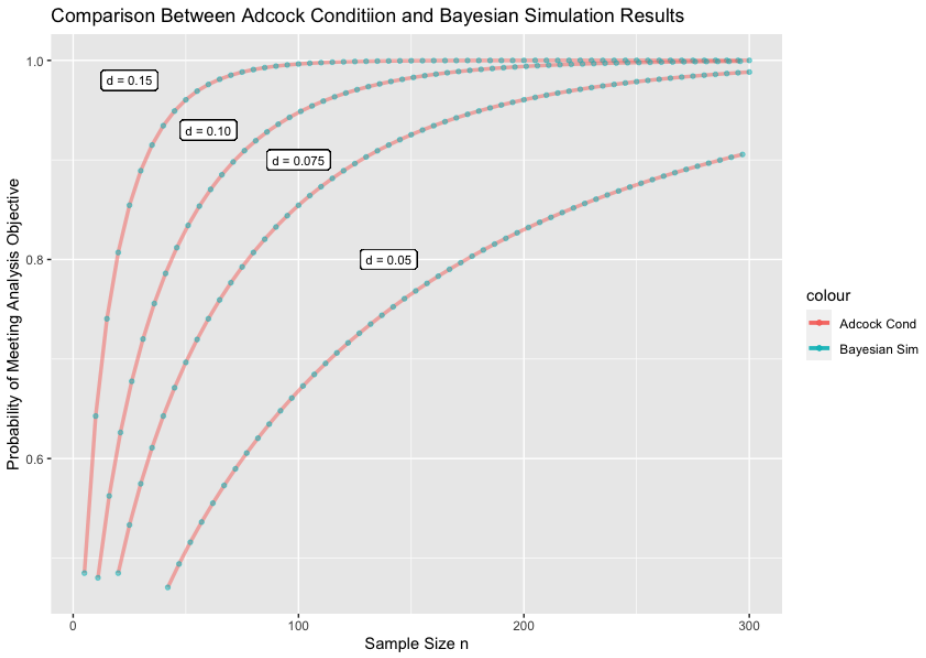


Figure B.1: Overlay of simulated results and frequentist results given a weak analysis prior such that $n_a \to 0$.

iterating through multiple samples of size n and observing the proportion of these samples that meet the analysis stage objective from Equation (B.4). As we have shown in the previous section, this becomes trivial in the case where weak analysis priors are assigned as we are left with a condition that is independent of $\bar{x}$. Hence, we are able to obtain the exact same probability values as those obtained from the frequentist formula.

B.5 Convergence to Frequentist Setting in Section 3.4

Recall the following expression for assurance in (B.4), as well as the expression for sample size,

$$n = z^2_{1-\alpha/2}\frac{\sigma^2}{d^2}\ . \tag{B.5}$$

In Equation (B.4), notice that we are ultimately assessing whether the expression on the left hand side exceeds $1-\alpha$. Hence, we can isolate $1-\alpha$ in Equation (B.5) to facilitate comparisons of the Bayesian and frequentist settings in relation to the probability of meeting the pre-specified condition. Starting from Equation (B.5), simple rearrangement reveals

$$n \geq z^2_{1-\alpha/2}\frac{\sigma^2}{d^2} \implies \frac{\sqrt{n}}{\sigma}d \geq z_{1-\alpha/2} \implies \Phi\left[\frac{\sqrt{n}}{\sigma}d\right] \geq 1-\alpha/2 \implies 2\Phi\left[\frac{\sqrt{n}}{\sigma}d\right] - 1 \geq 1-\alpha.$$

If we refer back to Equation (B.4), it becomes clear that setting $n_a = 0$ will simplify the expression down to the same expression we had previously obtained for the above frequentist scenario. Hence,

$$\begin{aligned}\delta &= \left\{\bar{x} : \Phi\left[\frac{\sqrt{n_a+n}}{\sigma}(\bar{x}+d-\lambda)\right] - \Phi\left[\frac{\sqrt{n_a+n}}{\sigma}(\bar{x}-d-\lambda)\right] \geq 1-\alpha\right\} \\ &\xrightarrow{n_a=0} \left\{\bar{x} : \Phi\left[\frac{\sqrt{n}}{\sigma}d\right] - \Phi\left[-\frac{\sqrt{n}}{\sigma}d\right] \geq 1-\alpha\right\} = \left\{\bar{x} : 2\Phi\left[\frac{\sqrt{n}}{\sigma}d\right] - 1 \geq 1-\alpha\right\}.\end{aligned}$$

In other words, if we let θ take on a weak analysis prior, we revert back to the frequentist setting in the analysis stage.

B.6 Relation to Frequentist Setting in Beta-Binomial Setting in Section 3.5

It is worth pointing out that there are no precision parameters to quantify the amount of information we have on the priors being assigned. Directly showcasing parallel behaviors between Bayesian and frequentist settings involve knowing the probabilities beforehand and passing them in as arguments into the simulation. Specifically, if p_1 and p_2 are known beforehand, we can express these "exact" priors as Uniform distributions such that $p_i \sim \text{U}[p_i, p_i]$. We can then express the overall analysis stage prior as a probability mass function:

$$p_i = \psi \text{Unif}[p_i, p_i] + (1 - \psi)\text{Beta}(\alpha_i, \beta_i), \quad i = 1, 2,$$

where ψ denotes the binary indicator variable for knowing exact values of p_i beforehand. If $\psi = 1$, we are drawing from the uniform distribution under the assumption of exact priors. Otherwise, $\psi = 0$ and we draw from the beta distribution to evaluate the analysis stage objective.

There is also an additional route we can use to showcase overlapping behaviors between the Bayesian and frequentist paradigms. Recall the sample size formula for assessing differences in proportions in the frequentist setting,

$$n = \frac{(z_{1-\alpha/2} + z_\beta)^2 (p_1(1-p_1) + p_2(1-p_2))}{(p_1 - p_2)^2},$$

where $n = n_1 = n_2$. Simple rearragements and noting that $-(z_{1-\alpha/2} + z_\beta) = z_{1-\beta} - z_{1-\alpha/2}$ lead us to obtain

$$\frac{\sqrt{n}(p_1 - p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}} + z_{1-\alpha/2} = z_{1-\beta}$$

$$\implies \text{Power} = 1 - \beta = \Phi\left(\frac{\sqrt{n}(p_1 - p_2)}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}} + z_{1-\alpha/2} \right).$$

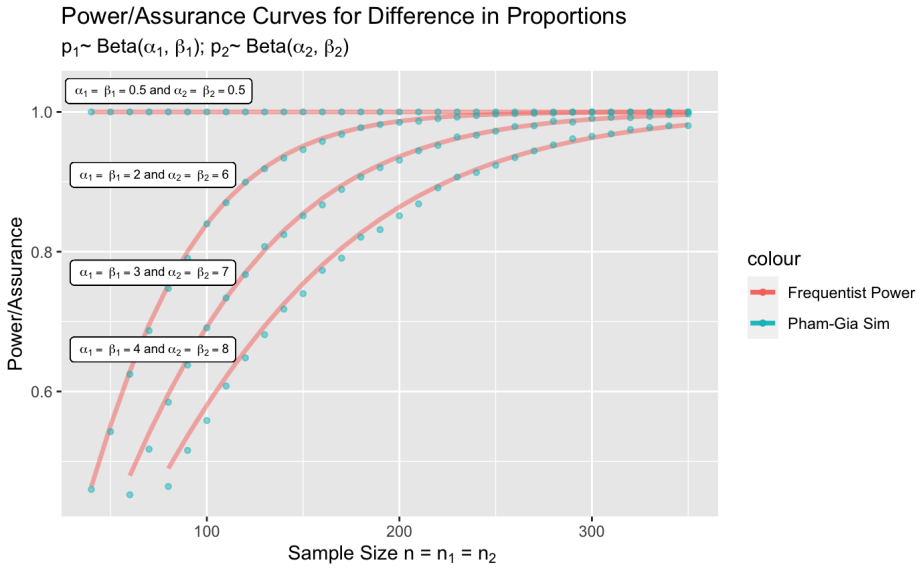


Figure B.2: Overlay of simulated assurance results using posterior credible intervals and frequentist power results based on regular confidence intervals.

In an ideal situation, we could determine suitable parameters for α_i, β_i, $i = 1, 2$ to use as our Beta priors that would enable demonstration of convergence towards the frequentist setting. However, a key relationship to recognize is that the Beta distribution is a conjugate prior of the Binomial distribution. There is a subtle advantage offered given that the Bayesian credible interval bands are based upon posterior parameters of the Beta distribution and the frequentist confidence interval bounds are based upon the Binomial distribution. Because of the conjugate relationship held by the Beta and Binomial distributions, we are essentially assigning priors to parameters in the Bayesian setting that the Binomial density in the frequentist setting is conditioned upon.It is helpful to note that the Beta distribution is approximately normal when its parameters α and β are set to be equal and large. Hence, the normal distribution can be used to approximate binomial distributions for large sample sizes. Knowing this, we can manually assign such values in our simulation study to utilize this relationship.

Figure B.2 displays the assurance curves overlayed on top of the frequentist power curves. As mentioned in the previous section, we manually set the parameters of the beta priors to

be equal as doing so results in approximately normal behavior. The horizontal line at the top of the graph corresponds to flat priors for the beta distribution known as Haldane's priors, in which the α and β parameters are all set to 0.5. Although the points do not align perfectly with the frequentist curves as we rely on an approximate relationship rather than identifying prior assignments that allow direct ties to the frequentist case, our model still performs fairly well as the points and curves are still relatively close to one another.

B.7 Deriving Goal Function Threshold in Section 3.6

Consider the linear hypothesis test $H_0 : u^\top \beta = c_0$ vs. $H_a : u^\top \beta = c_1$ under the general linear model $y = X\beta + \epsilon$, where y is $n \times 1$, X is $n \times p$, β is $p \times 1$, u is $p \times 1$, and $\epsilon \sim N(0, \sigma^2 I_n)$, implying that c_0 and c_1 are scalars.

Let us assume that $u^\top \beta$ is estimable. Then, by the fundamental principle of estimable functions, there exists a linear unbiased estimate $b^\top y$ such that $E(b^\top y) = u^\top \beta$. This suggests that u belongs in the column space of $X^\top$ since

$$E(b^\top y) = b^\top X\beta = u^\top \beta \implies b^\top X = u^\top \implies u = X^\top b \implies u \in C(X^\top).$$

Hence, u can be expressed as $u = X^\top z$ for some $z \in \mathbb{R}^n$. Letting $\tilde{y} = z^\top y$, simple linear transformation leads to $\tilde{y}|H_0 \sim N(c_0, \sigma^2 z^\top z)$ and $\tilde{y}|H_a \sim N(c_1, \sigma^2 z^\top z)$.

First, we assign a prior on $u^\top \beta$ such that $P(H_0) = 1 - P(H_a) = \pi$ and assume that the null hypothesis is not rejected if the posterior probability of H_0 is at least $1/(1+K)$, where K is the amount of utility associated with H_0 being correctly accepted. Starting from the expression $P(H_0|\tilde{y}) \geq \frac{1}{1+K}$, we apply fundamental Bayesian principles on the left hand side of the inequality to obtain

$$\frac{P(\tilde{y}|H_0)P(H_0)}{P(\tilde{y})} \geq \frac{1}{1+K}.$$

Substituting appropriate densities and assigned values results in

$$\frac{N(\tilde{y}|c_0, \sigma^2 z^\top z)(\pi)}{N(\tilde{y}|c_0, \sigma^2 z^\top z)(\pi) + N(\tilde{y}|c_1, \sigma^2 z^\top z)(1-\pi)} \geq \frac{1}{1+K}.$$

Fundamental algebra leads to the following criteria, in which H_0 is not rejected if

$$\tilde{y} \leq \frac{\sigma^2 z^\top z}{\delta} \ln\left(\frac{K\pi}{1-\pi}\right) + \frac{c_1 + c_0}{2},$$

where $\delta = c_1 - c_0$. This condition can be expressed in a more cohesive way through stan-

dardization. Given that $\tilde{y} \sim N(u^\top \beta, \sigma^2 z^\top z)$, it follows that

$$\frac{\tilde{y} - u^\top \beta}{\sigma\sqrt{z^\top z}} \sim N(0,1).$$

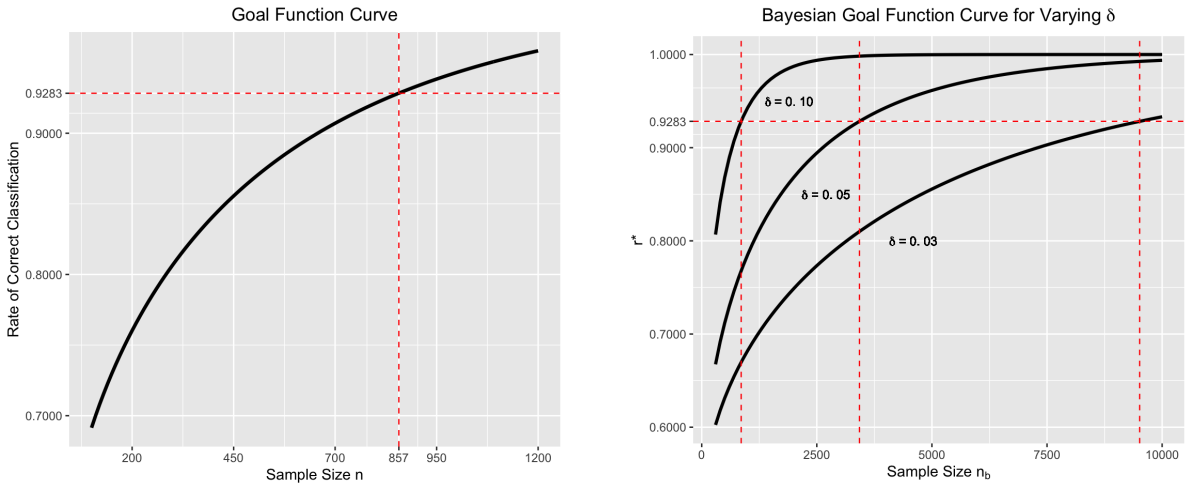
Hence, the probability of correctly accepting H_0 is given by

$$P(\text{fail to reject } H_0 |\ H_0 \text{ is true}) = \Phi\left[\frac{\sigma\sqrt{z^\top z}}{\delta}\ln\left(\frac{K\pi}{1-\pi}\right) + \frac{\delta}{2\sigma\sqrt{z^\top z}}\right],$$

where Φ denotes the cumulative distribution function of the standard normal.

B.8 Utility Function Example from Reference Paper in Section 3.6

We construct the utility curve using the set of fixed values specified in Section 2.1 of Inoue, Berry, and Parmigiani, 2005, where $\pi = 0.5$, $K = 1$, $\sigma^2 = 1$ and the critical difference, δ, is taken to be 0.1. Referring to our linear hypothesis testing framework, $H_0 : u^\top \beta = c_0$ vs. $H_a : u^\top \beta = c_1$, we let $u = 1$, $c_0 = 0.5$ and $c_1 = 0.6$ to adhere to the critical difference condition of $\delta = 0.1$. Under these specifications, the paper reports a minimum sample size of $n = 857$ to ensure a rate of correct classification of $r^* = 0.9283$. Using our linear model framework outlined in the previous section, our resulting utility curve displayed in Figure B.3a supports this claim, as indicated by the intersection of the dashed lines that highlights this exact point on the curve. We repeat the same steps for two additional critical differences, displayed in Figure B.3b. The intersections marked in the plot each correspond to the same r^* value of 0.9283, indicated by the horizontal dashed line. The figure suggests that smaller critical differences require larger sample sizes to meet the same rate of correct classification criteria. More specifically, when designing studies that require detecting critical differences of $\delta = 0.10$, $\delta = 0.05$, and $\delta = 0.03$, minimum sample sizes of 857, 3426, and 9512 are respectively needed to ensure a rate of correct classification of $r^* = 0.9283$.



(a) Rate of classification curve. Red dashed lines indicate a sample size of n=857 is needed to ensure $r^* = 0.9283$.

(b) Utility curves resulting from the same set of fixed parameters and different critical differences, δ.

Figure B.3: Utility curves using specifications provided by Inoue, Berry, and Parmigiani, 2005.

Bibliography

Adcock, C.J. (1997). "Sample Size Determination: A Review". In: *The Statistician* 46.2.

Armstrong, Richard A. (2014). "When to use Bonferroni correction". In: *Ophthalmic and Physiological Optics* 34.5.

Benjamani, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society* 57.1.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer New York.

Berry, Donald A. (2006). "Bayesian Clinical Trials". In: *Nature Reviews Drug Discovery* 5.1.

Berry, Scott M. et al. (2010). *Bayesian Adaptive Methods for Clinical Trials*. United Kingdom: Chapman & Hall/CRC Biostatistics Series.

Bickel, David R. (2003). *Selecting an Optimal Rejection Region for Multiple Testing: A Decision-Theoretic Alternative to FDR Control, With An Application to Microarrays*. Tech. rep. Medical College of Georgia.

Bland, J. Martin and Douglas G. Altman (1995). "Multiple significance tests: the Bonferroni method". In: *BMJ* 310.170.

Brutti, Pierpaolo, Fulvio Santis, and Stefania Gubbiotti (2014). "Bayesian-frequentist sample size determination: A game of two priors". In: *METRON* 72.2.

Cao, Jing, J. Jack Lee, and Susan Alber (2014). "Comparison of Bayesian Sample Size Criteria: ACC, ALC, and WOC". In: *J Stat Plan Inference* 139.12.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian Experimental Design: A Review". In: *Statistical Science* 10.3.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch (2010). "BART: Bayesian Additive Regression Trees". In: *The Annals of Applied Statistics* 4.1.

Clarke, B. and Ao Yuan (2006). "Closed Form Expressions for Bayesian Sample Size". In: *The Annals of Statistics* 34.3.

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

Desu, M.M. and Damaraju Raghavarao (1990). *Sample Size Methodology*. Massachusetts: Elsevier.

Efron, Bradley (2007). "Size, Power and False Discovery Rates". In: *The Annals of Statistics* 35.4.

— (2008). "Microrrays, Empirical Bayes and the Two-Groups Model". In: *Statistical Science* 23.1.

Efron, Bradley et al. (2001). "Empirical Bayes Analysis of Microarray Experiment". In: *Journal of the American Statistical Association* 96.456.

Gelfand, Alan E. and Fei Wang (2002). "A simulation based approach to Bayesian sample size determination for performance under a given model and for separating models". In: *Statistical Science* 17.2.

Gelman, Andrew, John B. Carlin, and Hal S. Stern (2013). *Bayesian Data Analysis (3rd ed.)* United Kingdom: Chapman and Hall/CRC.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5.

Genovese, Christopher and Larry Wasserman (2002). "Operating characteristics and extensions of the FDR procedure". In: *Journal of the Royal Statistical Society Series B* 64.3.

— (2004). "A stochastic process approach to false discovery control". In: *The Annals of Statistics* 32.3.

Gottardo, Raphael et al. (2005). "Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples". In: *Biometrics* 62.1.

Ibrahim, Joseph G. et al. (2012). "Bayesian Meta-Experimental Design: Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes". In: *Biometrics* 68.2.

Inoue, Lurdes Y.T., Donald A. Berry, and Giovanni Parmigiani (2005). "Relationship Between Bayesian and Frequentist Sample Size Determination". In: *The American Statistician* 59.1.

Joseph, Lawrence and Patrick Belisle (1997a). "Bayesian Sample Size Determination for Normal Means and DIfferences Between Normal Means". In: *Journal of the Royal Statistical Society* 46.2.

— (1997b). "Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means". In: *The Statistician* 46.2.

— (2009). *Package 'SampleSizeProportions*. R package version 1.0. URL: https://cran.r-project.org/web/packages/SampleSizeProportions/SampleSizeProportions.pdf.

— (2012). *Package 'SampleSizeMeans*. R package version 1.1. URL: https://cran.r-project.org/web/packages/SampleSizeMeans/SampleSizeMeans.pdf.

Joseph, Lawrence, Roxane Du Berger, and Patrick Belisle (1997). "Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination". In: *Statistics in Medicine* 16.7.

Joseph, Lawrence, David B.. Wolfson, and Roxane du Berger (1995a). "Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals". In: *The Statistician* 44.2.

Joseph, Lawrence, David B. Wolfson, and Roxane Du Berger (1995b). "Sample Size Calculations for Binomial Proportions via Highest Posterior Density Intervals". In: *The Statistician* 44.2.

Kendziorski, C.M. et al. (2003). "On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles". In: *Statistics in Medicine* 22.24.

Kraemer, Helena Chmura and Sue Thiemann (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park: Sage Publications.

Lee, Jack and Caleb T. Chu (2012). "Bayesian clinical trials in action". In: *Statistics in Medicine* 31.25.

Lee, Mei-Ling Ting and G.A. Whitmore (2002). "Power and sample size for DNA microarray studies". In: *Statistics in Medicine* 21.23.

Lee, Mei-Ling Ting et al. (2000). "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations". In: *PNAS* 97.18.

Lee, Sandra J. and Marvin Zelen (2000). "Clinical Trials and Sample Size Considerations: Another Perspective". In: *Statistical Science* 15.2.

Lindley, Dennis V. (1997). "The Choice of Sample Size". In: *The Statistician* 46.2.

Liu, Guanghan and Kung Yee Liang (1997). "Sample Size Calculations for Studies with Correlated Observations". In: *Biometrics* 53.3.

Mukherjee, Sayan et al. (2004). "Estimating dataset size requirements for classifying DNA microarray data". In: *Journal of Computational Biology* 119.42.

Muller, Keith E. et al. (1992). "Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications". In: *Journal of the American Statistical Association* 87.420.

Müller, Peter and Giovanni Parmigiani (1995). "Optimal Design via Curve Fitting of Monte Carlo Experiments". In: *Journal of the American Statistical Association* 90.432.

Müller, Peter et al. (2004). "Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays". In: *Journal of the American Statistical Association* 99.468.

Newton, M.A. et al. (2001). "On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data". In: *Journal of Computational Biology* 8.1.

O'Hagan, Anthony and John W. Stevens (2001). "Bayesian Assessment of Sample Size for Clinical Trials of Cost-Effectiveness". In: *Medical Decision Making* 21.3.

Pan, Jane and Sudipto Banerjee (n.d.). "Multiple Testing in the Bayesian Framework". In preparation.

— (2021a). "A Unifying Bayesian Approach for Sample Size Determination Using Design and Analysis Priors". In: *ArXiv e-prints*. arXiv: 2112.03509.

— (2021b). "bayesassurance: An R package for calculating sample size and Bayesian assurance". In: *ArXiv e-prints*. arXiv: 2203.15154.

Pan, Jane et al. (2022). "Bayesian Additive Regression Trees (BART) with covariate adjusted borrowing in subgroup analyses". In: *Journal of Biopharmaceutical Statistics* 32 (4). URL: https://doi.org/10.1080/10543406.2022.2089160.

Pan, Wei, Jizhen Lin, and Chap T. Le (2002). "How many replicates of arrays are required to detect gene expression changes in microarray experiments?" In: *Genome Biology* 3.research0022.1.

Parmigiani, Giovanni (2002). *Modeling in Medical Decision Making: A Bayesian Approach*. United States: Wiley.

Pham-Gia, T. (1997). "On Bayesian Analysis, Bayesian Decision Theory and the Sample Size Problem". In: *The Statistician* 46.2.

Pulkstenis, Erik, Kaushik Patra, and Jianliang Zhang (2017). "A Bayesian paradigm for decision-making in proof-of-concept trials". In: *Journal of Biopharmaceutical Statistics* 27.3.

Rahme, Elham, Lawrence Joseph, and Theresa W. Gyorkos (2000). "Bayesian Sample Size Determination for Estimating Binomial Parameters from Data Subject to Misclassification". In: *Journal of Royal Statistical Society* 49.1.

Raiffa, Howard and Robert Schlaifer (1961). *Applied Statistical Decision Theory*. Massachusetts: Harvard University Graduate School of Business Administration (Division of Research).

Reyes, Eric M. and Sujit K. Ghosh (2013). "Bayesian Average Error Based Approach to Sample Size Calculations for Hypothesis Testing". In: *Biopharm* 23.3.

Sahu, S.K. and T.M.F. Smith (2006). "A Bayesian Method of Sample Size Determination With Practical Applications". In: *Journal of the Royal Statistical Society* 169.2.

Santis, Fulvio De (2006). "Sample Size Determination for Robust Bayesian Analysis". In: *Journal of the American Statistical Association* 101.473.

— (2007). "Using Historical Data for Bayesian Sample Size Determination". In: *Statistics in Society* 170.1.

Scott, James G. and James O. Berger (2006). "An exploration of aspects of Bayesian. multiple testing". In: *Journal of Statistical Planning and Inference* 136.

Self, Steven G. and Robert H. Mauritsen (1988). "Power/Sample Size Calculations for Generalized Linear Models". In: *Biometrics* 44.1.

Self, Steven G., Robert H. Mauritsen, and Jill O'Hara (1992). "Power calculations for likelihood ratio tests in generalized linear models". In: *Biometrics* 48.1.

Shepel, Laurie A et al. (1998). "Genetic Identification of Multiple Loci that Control Breast Cancer Susceptibility in the Rat". In: *Genetics* 149.1.

Spiegelhalter, David J., Laurence S. Freedman, and Mahesh K.B. Parmar (1993). "Applying Bayesian ideas in drug development and clinical trials". In: *Statistics in Medicine* 12.15.

Storey, John D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value". In: *The Annals of Statistics* 31.6.

Storey, John D. and Robert Tibshirani (2003). "The Analysis of Gene Expression Data". In: Springer-Verlag. Chap. SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays, pp. 272–290.

Tseng, Chi-Hong and Yongzhao Shao (2012). "Sample size growth with an increasing number of comparisons". In: *Journal of Probability and Statistics* 2012.

Tukey, John W. (1949). "Comparing Individual Means in the Analysis of Variance". In: *Biometrics* 5.2.

VanderWheele, Tyler J. and Maya B. Mathur (2019). "Some Desirable Properties of the Bonferroni Correction: Is the Bonferroni Correction Really So Bad?" In: *American Journal of Epidemiology* 188.3.

Vickerstaff, Victoria et al. (2019). "Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomized controlled trials with multiple primary outcomes". In: *BMC Medical Research Methodology* 19.129.

Wacholder, Sholom et al. (2004). "Assessing the probability that a positive report is false: an approach for molecular epidemiology studies". In: *Journal of the National Cancer Institute* 96.6.

Weiss, Robert (2002). "Bayesian sample size calculations for hypothesis testing". In: *The Statistician* 46.2.

Wen, Xiaoquan (2018). "A unified view of false discovery rate control: reconciliation of Bayesian and Frequentist approaches". In: *ArXiv e-prints*. arXiv: 1803.05284.

Whittemore, Alice S. (2007). "A Bayesian False Discovery Rate for Multiple Testing". In: *Journal of Applied Statistics* 34.1.

Xu, ChangJiang, Antonio Ciampi, and Celia M.T. Greenwood (2014). "Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation". In: *Frontiers in Genetics* 5.11.

Zien, Alexander et al. (2002). *Microarrays: How Many Do You Need?* Tech. rep. Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.