Design with Sampling Distribution Segments

by

Luke Hagar

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Hugh Chipman Professor, Dept. of Mathematics and Statistics, Acadia University
Supervisor:	Nathaniel Stevens Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal Members:	Christiane Lemieux Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Liqun Diao Assistant Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Ashok Chaurasia Associate Professor, School of Public Health Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The chapters of this thesis comprise Luke Hagar's work, under the supervision of Dr. Nathaniel Stevens, that has been submitted for publication in the following venues.

- Chapter 2: Hagar, L. and N. T. Stevens (2024+). Bioequivalence design with sampling distribution segments. Submitted to *Statistics in Medicine*. arXiv.
- Chapter 3: Hagar, L. and N. T. Stevens (2024+). Fast power curve approximation for posterior analyses. Major revision at *Bayesian Analysis*. arXiv.
- Chapter 4: Hagar, L. and N. T. Stevens (2024+). Posterior ramifications of prior dependence structures. Major revision at *Statistical Science*. arXiv.
- Chapter 5: Hagar, L. and N. T. Stevens (2024+). Scalable design with posterior-based operating characteristics. Major revision at *Journal of the American Statistical Association*. arXiv.

Abstract

In most settings where data-driven decisions are made, these decisions are informed by two-group comparisons. Characteristics – such as median survival times for two cancer treatments, defect rates for two assembly lines, or average satisfaction scores for two consumer products – quantify the impact of each choice available to decision makers. Given estimates for these two characteristics, such comparisons are often made via hypothesis tests. This thesis focuses on sample size determination for hypothesis tests with interval hypotheses, including standard one-sided hypothesis tests, equivalence tests, and noninferiority tests in both frequentist and Bayesian settings.

To choose sample sizes for nonstandard hypothesis tests, simulation is used to estimate sampling distributions of e.g., test statistics or posterior summaries corresponding to various sample sizes. These sampling distributions provide context as to which estimated values for the two characteristics are plausible. By considering quantiles of these distributions, one can determine whether a particular sample size satisfies criteria for the operating characteristics of the hypothesis test: power and the type I error rate.

It is standard practice to estimate *entire* sampling distributions for each sample size considered. The computational cost of doing so impedes the adoption of non-simplistic designs. However, only *quantiles* of the sampling distributions must be estimated to assess operating characteristics. To improve the scalability of simulation-based design, we could focus only on exploring the segments of the sampling distributions near the relevant quantiles. This thesis proposes methods to explore sampling distribution segments for various designs. These methods are used to determine sample sizes and decision criteria for hypothesis tests with orders of magnitude fewer simulation repetitions. Importantly, this reduction in computational complexity is achieved without compromising the consistency of the simulation results that is guaranteed when estimating entire sampling distributions.

In parametric frequentist hypothesis tests, test statistics are often constructed from exact pivotal quantities. To improve sample size determination in the absence of exact pivotal quantities, we first propose a simulation-based method for power curve approximation with such hypothesis tests. This method leverages low-discrepancy sequences of sufficient statistics and root-finding algorithms to prompt unbiased sample size recommendations using sampling distribution segments.

We also propose a framework for power curve approximation with Bayesian hypothesis tests. The corresponding methods leverage low-discrepancy sequences of maximum likelihood estimates, normal approximations to the posterior, and root-finding algorithms to explore segments of sampling distributions of posterior probabilities. The resulting sample

size recommendations are consistent in that they are suitable when the normal approximations to the posterior and sampling distribution of the maximum likelihood estimator are appropriate.

When designing Bayesian hypothesis tests, practitioners may need to specify various prior distributions to generate and analyze data for the sample size calculation. Specifying dependence structures for these priors in multivariate settings is particularly difficult. The challenges with specifying such dependence structures have been exacerbated by recommendations made alongside recent advances with copula-based priors. We prove theoretical results that can be used to help select prior dependence structures that align with one's objectives for posterior analysis.

We lastly propose a comprehensive method for sample size determination with Bayesian hypothesis tests that considers our recommendations for prior specification. Unlike our framework for power curve approximation, this method recommends probabilistic cutoffs that facilitate decision making while controlling both power and the type I error rate. This scalable approach obtains consistent sample size recommendations by estimating segments of two sampling distributions – one for each operating characteristic. We also extend our design framework to accommodate more complex two-group comparisons that account for additional covariates.

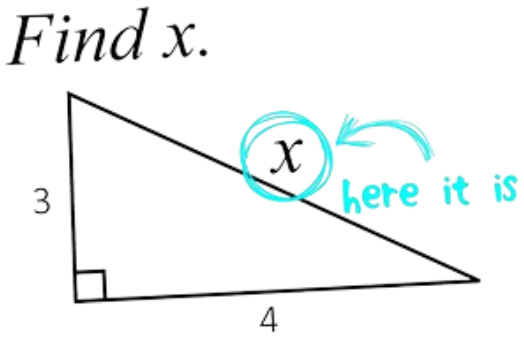
Acknowledgements

I would first like to thank my supervisor, Dr. Nathaniel Stevens, for his support during my postsecondary studies. I am grateful for the many opportunities he extended to me. Beyond those opportunities, I appreciate him consistently making time to provide helpful guidance – both on this dissertation and otherwise. The qualities that he has exhibited as a mentor are ones that I hope to emulate, wherever my career might take me.

I would also like to thank the members of my thesis committee – Dr. Christiane Lemieux, Dr. Liqun Diao, Dr. Ashok Chaurasia, and Dr. Hugh Chipman – for their helpful suggestions and insightful comments. Moreover, this thesis incorporates several concepts from a course that I took with Christiane as a PhD student, and this thesis would be radically different had I not taken her course.

Beyond the scope of this thesis, I would like to thank Dr. Christine Anderson-Cook, Dr. Lu Lu, Dr. Svetlana Kaminskaïa, and Dr. Wladyslaw Cichocki for their mentorship on collaborative research projects. My PhD experience was also enriched by the mentors and students that I met through engaging with the Statistical Society of Canada's Canadian Statistics Student Conference and Student and Recent Graduate Committee. As well, it was a privilege to work with my colleagues at the University of Waterloo's Centre for Teaching Excellence and learn from their experiences as instructors in various disciplines.

Lastly, I would like to thank my family, especially my parents and sisters. In addition to supporting my learning, my parents have encouraged my personal growth. In high school, they gifted me a math t-shirt with the image below. I refused to wear the shirt, explaining to my parents that mathematical illiteracy is no joke. Without my knowledge, my mom later wore the shirt for a parent-teacher interview with my calculus teacher. My dad immortalized the incident with a photo: a reminder to lighten up.



Dedication

This thesis is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2×2 two-sequence, two-period

ASA American Statistical Association

AUC area under the curve

BvM Bernstein-von Mises

CDF cumulative distribution function

ENIGH National Survey of Household Income and Expenses (Mexico)

FDR false discovery rate

HDI highest density interval

HPD highest posterior density

i.i.d. independently and identically distributed

MCMC Markov chain Monte Carlo

MLE maximum likelihood estimator

MXN Mexican pesos

NHST null hypothesis significance test

NOH nonoverlapping-hypotheses

PDF probability density function

SIR sampling-importance-resampling

TOST two one-sided test

Chapter 1

Introduction

1.1 Design with Sampling Distributions

1.1.1 Interval Hypothesis Specification

Hypothesis tests allow practitioners to compare scalar quantities θ_1 and θ_2, where the characteristic θ_j describes a comparison ($j = 1$) or reference ($j = 2$) group. These comparisons are typically facilitated using the difference between the characteristics: $\theta = \theta_1 - \theta_2$. The comparison can also be made with a ratio-based metric $\theta > 0$ (e.g., $\theta = \theta_1/\theta_2$), but even such metrics can be expressed as differences on the logarithmic scale. While this thesis generally considers comparisons with two groups of independent data, these hypothesis testing methods can also be simplified to consider a characteristic θ that describes a single group.

Null hypothesis significance tests (NHSTs) often assess whether θ is equal to a fixed constant θ_0. In those situations, the null hypothesis $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta \neq \theta_0$ are compared. The use of frequentist NHSTs to assess point null hypotheses has generated substantial discourse in the statistical community over the past decade. In part, this discourse can be attributed to two recent publications. The first is an official statement released by the American Statistical Association (ASA) in response to the replication crisis (Ioannidis, 2005). This statement challenged the statistical community to develop alternatives to, and extensions of, the traditional hypothesis testing framework (Wasserstein and Lazar, 2016). The second publication is a special issue of *The American Statistician* comprising 43 articles that explored the shortcomings of traditional hypothesis testing and discussed "moving to a world beyond $p < 0.05$" (Wasserstein et al., 2019).

Both publications emphasized the relationship between power and the sample size n when testing point null hypotheses. The term *power* will be precisely defined in various contexts throughout the thesis. Generally, this relationship implies that when the true value of θ is $\theta_0 + \epsilon$ for any $|\epsilon| > 0$, the probability of rejecting H_0 approaches 1 as $n \to \infty$. Hence, there is limited value in testing point null hypotheses with extremely large data sets: H_0 will almost certainly be rejected even if $\epsilon \approx 0$ and θ does not practically differ from θ_0. Given the current availability of big data, we should reconsider how hypothesis testing can meaningfully support decision making. In Bayesian settings, θ is often considered to be a continuous random variable. Bayesian hypothesis tests therefore rarely assess point null hypotheses since the probability of a continuous random variable equaling a constant is zero (Gelman et al., 2020).

Since point null hypotheses have shortcomings in frequentist settings and are implausible in most Bayesian ones, this thesis considers hypotheses of the form $H_0 : \theta_1 - \theta_2 \notin (\delta_L, \delta_U)$, where $-\infty \leq \delta_L < \delta_U \leq \infty$. The alternative hypothesis that we wish to support is therefore $H_1 : \theta_1 - \theta_2 \in (\delta_L, \delta_U)$. The interval (δ_L, δ_U) accommodates the context of comparison. Assuming larger θ_j values are preferred, the intervals $(\delta_L, \delta_U) = \{(0, \infty), (-\delta, \delta), (-\delta, \infty)\}$ for some equivalence margin $\delta > 0$ may be used to respectively assess whether θ_1 is superior, practically equivalent, or noninferior to θ_2 (Wellek, 2010; Walker and Nowacki, 2011; Spiegelhalter et al., 1994, 2004).

This thesis emphasizes the design of studies in which hypothesis tests are used to analyze data. In this thesis, we refer to the process of choosing sample sizes and decision criteria for these tests as *hypothesis test design*. By supporting scalable design for a broad suite of hypothesis tests, this thesis extends the traditional hypothesis testing framework as called for by the ASA in their official statement (Wasserstein and Lazar, 2016). We introduce an example in Section 1.1.2 to preview the comprehensive scope of the parametric frequentist and Bayesian hypothesis tests that our design methods accommodate.

1.1.2 Illustrative Example

The Jackson Laboratory recently investigated the impact of diet composition on physiology and liver gene expression in outbred mice (Gatti et al., 2017). In this experiment, mice were fed either a standard chow diet or a high-fat diet from weaning to age 26 weeks. The associated data set is available in the `dslabs` package in R (Irizarry and Gill, 2023). Here, we consider the datum y_{ij} collected for each mouse $i = 1, ..., n_j,\ j = 1, 2$ to be their body weight in grams (g) measured at 19 weeks. The scalar observations in group j are collectively denoted as $\boldsymbol{y}_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})^T$ in this chapter. When considering mice

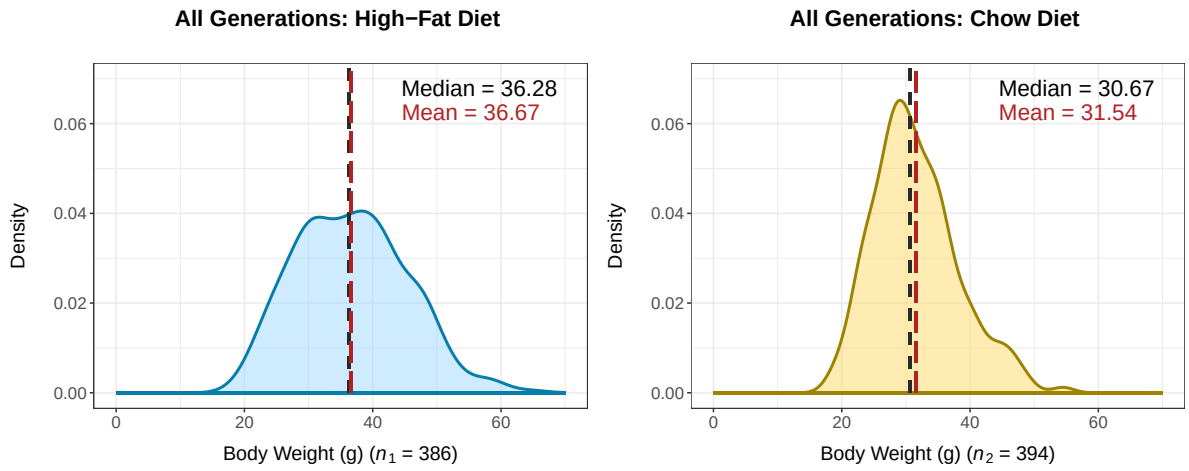


Figure 1.1: Density plots of body weight (in grams) for all generations of mice split by diet.

from all generations bred for this experiment, there are $n_1 = 386$ and $n_2 = 394$ observations in the high-fat ($j = 1$) and chow ($j = 2$) diet groups. We estimated the distribution of the body weight in each group using nonparametric density estimation (Wand and Jones, 1994). These distributions are visualized in Figure 1.1, which informally demonstrate that mice fed the high-fat diet generally have greater body weights than those fed the standard chow one. If we expect diet composition to impact body weight, it could be natural to compare the typical body weights for the two diets using a superiority or noninferiority test.

Our design framework also accommodates comparisons based on practical equivalence. The generations of mice bred for the experiment are composed of two litters, where roughly half the mice per litter are assigned to each of the two diets. As such, the body weights should not differ substantially in each of the two litters. To confirm this intuition, we also consider the subset of mice bred in generation 11. This prompts $n_1 = 100$ and $n_2 = 98$ observations in the groups for the first ($j = 1$) and second ($j = 2$) litters. The density estimates for these two body weight distributions are given in Figure 1.2. These distributions appear to be more similar than those in Figure 1.1. We use this example to provide a context to overview design with sampling distributions for frequentist and Bayesian hypothesis testing methods in Sections 1.1.3 and 1.1.4, respectively.

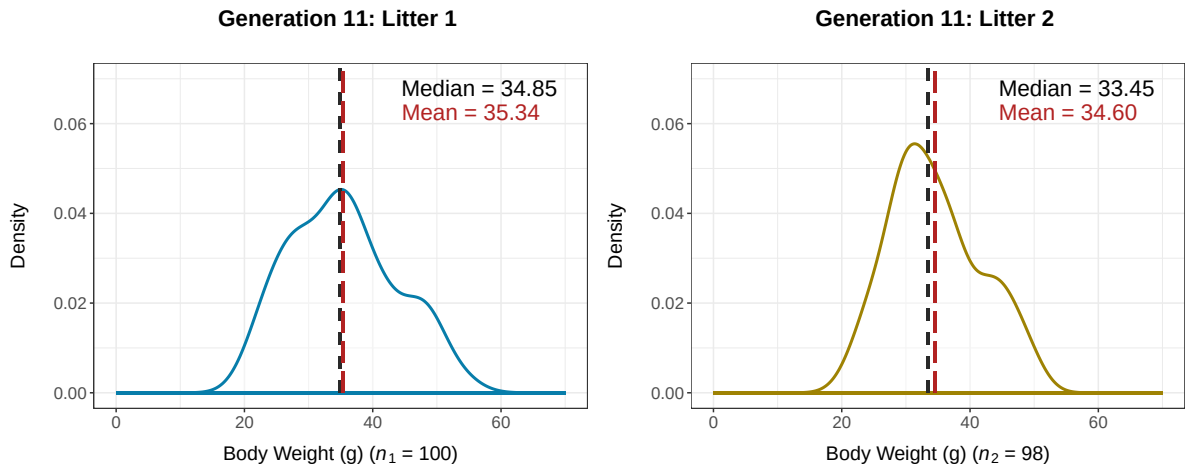


Figure 1.2: Density plots of body weight (in grams) for mice in generation 11 split by litter.

1.1.3 Frequentist Design

For parametric hypothesis tests, the characteristic of interest θ_j for group j is typically specified as a function $g(\cdot)$ of the (potentially) multivariate parameter $\boldsymbol{\eta}_j$ that parameterizes the data distribution. That is, $\theta_j = g(\boldsymbol{\eta}_j)$ for $j = 1, 2$. Various specifications of the function $g(\cdot)$ will be considered throughout this thesis. In frequentist settings, θ_1 and θ_2 are typically group means or variances. Hypothesis tests that compare proportions can be viewed as a special case of those that consider means. Alternative characteristics are typically compared using nonparametric hypothesis tests (see e.g., Pitman (1937); Wilcoxon (1992)), which are not the focus of this thesis.

When data $y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$, the model parameters are $\boldsymbol{\eta}_j = (\mu_j, \sigma_j^2)$. For normal data, means $\theta_1 = g(\boldsymbol{\eta}_1) = \mu_1$ and $\theta_2 = g(\boldsymbol{\eta}_2) = \mu_2$ are often compared using t-tests (Student, 1908; Welch, 1938). For large sample sizes n_1 and n_2, Z-tests based on the central limit theorem (Lehmann and Casella, 1998) facilitate such comparisons in most scenarios with non-normal data. Frequentist hypothesis tests also commonly compare two normal variances $\theta_1 = g(\boldsymbol{\eta}_1) = \sigma_1^2$ and $\theta_2 = g(\boldsymbol{\eta}_2) = \sigma_2^2$ using F-tests (Snedecor and Cochran, 1989).

To assess whether the mean body weight for the high-fat diet (θ_1) is greater than that for the chow diet (θ_2), we consider the hypotheses $H_0 : \theta \notin (0, \infty)$ and $H_1 : \theta \in (0, \infty)$ that are based on the difference $\theta = \theta_1 - \theta_2$. When using a Student's t-test to conduct

this hypothesis test, we obtained a p-value that was less than 2.2×10^{-16}. The hypothesis H_0 is rejected when the p-value does not exceed a significance level $\alpha \in (0, 1)$ chosen to bound the probability of making a type I error. Type I errors occur when the hypothesis $H_0 : \theta \notin (\delta_L, \delta_U)$ is incorrectly rejected. For the popular choice of $\alpha = 0.05$, we would reject $H_0 : \theta \notin (0, \infty)$ and conclude that $\theta_1 - \theta_2 \in (0, \infty)$.

We suppose there is interest in designing this straightforward hypothesis test to illustrate how sampling distributions are used to design frequentist superiority and noninferiority tests. We consider the sampling distribution of the t-test statistic:

$$t = \frac{\left(\hat{\theta}_1 - \hat{\theta}_2\right) - \theta_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1.1}$$

where $\hat{\sigma}$ is the estimate for the pooled variance of the two groups and θ_0 is some fixed value for $\theta = \theta_1 - \theta_2$ corresponding to H_0. To consider the sampling distribution of (1.1), we must choose sample sizes n_1 and n_2. Here, we select $n = n_1 = n_2 = 25$ and consider imbalanced sample size determination where $n_1 \neq n_2$ later in the thesis. Designing hypothesis tests based on Student's t-test also requires an anticipated value for the common population variance σ^2 in both groups. We let the anticipated value for σ^2 be 58.81, the pooled variance informed by the sample estimates $\hat{\sigma}_1^2 = 74.49$ and $\hat{\sigma}_2^2 = 43.13$, for illustrative purposes. These variability estimates are typically informed by previous studies.

We must consider two sampling distributions to design this t-test. The first sampling distribution we consider is that of (1.1) for the parameter values that maximize type I error under H_0. In this case, type I error is maximized when the two population means are the same. The grey curve in Figure 1.3 visualizes the sampling distribution for (1.1) under H_0 when the population means θ_1 and θ_2 are equal and when $\theta_0 = 0$. The 5% probability of making a type I error is depicted by the shaded grey area to the right of the vertical black line at the 0.95-quantile of this sampling distribution.

The second sampling distribution we consider is that of (1.1) for the parameter values that characterize the minimum effect size we would like to detect. This effect size corresponds to a positive value for $\theta_1 - \theta_2$, but $\theta_0 = 0$ once again. For illustration, we choose this effect size to be equal to the observed effect size: $\hat{\theta}_1 - \hat{\theta}_2 = 36.67 - 31.54 = 5.13$ grams. The relevant sampling distribution of (1.1) under H_1 is visualized by the pink curve in Figure 1.3. The 24.53% probability of making a type II error is represented by the pink shaded area. Type II errors occur when a false null hypothesis is not rejected. The power of a hypothesis test is $1 - \beta$, where β is the probability of making a type II error. While the type I error rate is controlled by the significance level α, the type II error rate is

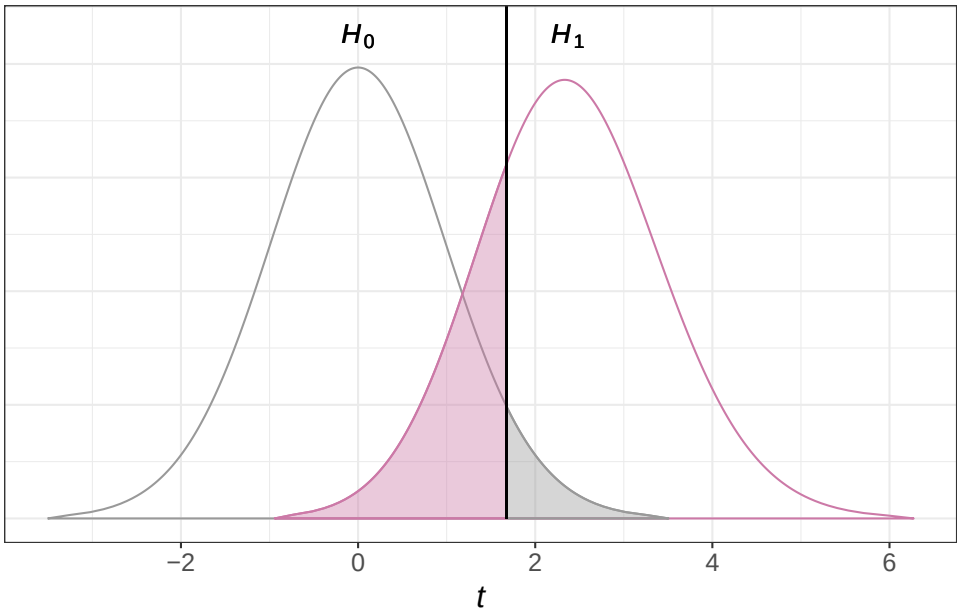


Figure 1.3: Sampling distributions for the t-test statistic under H_0 and H_1 when $n = 25$. The black line denotes the 0.95-quantile of the Student's t-distribution with $2n - 2 = 48$ degrees of freedom. The type I and II error rates are respectively depicted by the grey and pink shaded areas.

bounded by choosing a sufficiently large sample size. To design hypothesis tests, a target value for power is selected. A popular choice for the target power is $1 - \beta = 0.8$. Sampling distributions similar to those depicted in Figure 1.3 must be considered for various sample sizes n until the minimum sample size that achieves the target power is found. For this t-test, $n = 29$ is the smallest sample size that yields 80% power.

When using Student's t-tests, the sampling distributions for (1.1) follow known t-distributions under H_0 and H_1. We therefore do not require simulation to estimate the sampling distributions in this straightforward case. Nevertheless, we do not require full estimation of the sampling distributions under H_0 and H_1 to determine whether a sample size n achieves the target power. We only need to accurately estimate particular quantiles of these distributions since the target power is attained when the $(1 - \alpha)$-quantile of the sampling distribution under H_0 does not exceed the β-quantile of the sampling distribution under H_1. In Figure 1.3, a target power of 80% was not achieved for $n = 25$ since the 0.95-quantile of the sampling distribution under H_0 (1.677) exceeded the 0.2-quantile of that under H_1 (1.524). This straightfoward example with Student's t-tests illustrates that

design with sampling distribution segments near the relevant quantiles would be useful in more complex settings.

For illustration, we suppose that the equivalence margin δ is 3 grams. This choice suggests that an absolute difference of less than 3 grams between the mean body weights on the two diets is not of practical importance. To conduct frequentist equivalence tests that assess the hypotheses $H_0 : \theta \notin (-\delta, \delta)$ and $H_1 : \theta \in (-\delta, \delta)$, we use two one-sided hypothesis tests (Schuirmann, 1987). We conclude that $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$ if we reject the following two hypotheses on the basis of each of their p-values: $H_{0L} : \theta \in (-\infty, \delta_L]$ and $H_{0U} : \theta \in [\delta_U, \infty)$. We now consider the subset of mice from generation 11, where θ_1 and θ_2 are respectively the mean body weights from litter 1 and 2. When using Student's t-tests to assess H_{0L} and H_{0U}, both p-values were less than 0.021. We can therefore conclude that $\theta_1 - \theta_2 \in (-3, 3)$ at the 5% significance level. With interval hypothesis tests, our ability to conclude whether $\theta \in (\delta_L, \delta_U)$ depends on the choices for the interval endpoints. For instance, the p-value for H_{0U} would be 0.055 if we chose $\delta = 2.5$ grams. Thus, we do not have sufficient evidence to conclude that $\theta_1 - \theta_2 \in (-2.5, 2.5)$ at the 5% significance level. In practice, we should consider a single value of δ that is chosen by subject matter experts to account for the context of the comparison.

We now suppose there is interest in designing this straightforward comparison to illustrate how sampling distributions are used to design frequentist equivalence tests. We consider the joint sampling distribution of the following t-test statistics:

$$t_L = \frac{\left(\hat{\theta}_1 - \hat{\theta}_2\right) - \delta_L}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{and} \quad t_U = \frac{\delta_U - \left(\hat{\theta}_1 - \hat{\theta}_2\right)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1.2}$$

For illustration, we select $n = n_1 = n_2 = 100$ and choose the anticipated effect size to be equal to the observed effect size between the two litters: $\hat{\theta}_1 - \hat{\theta}_2 = 35.34 - 34.60 = 0.74$ grams. This effect size should be less than δ in absolute value for an equivalence test. To design this equivalence test, we let the anticipated value for σ^2 be 59.49 for illustrative purposes. This value is the pooled variance informed by generation 11's sample estimates $\hat{\sigma}_1^2 = 68.58$ and $\hat{\sigma}_2^2 = 50.41$.

We must consider three sampling distributions to design this equivalence test. The first two sampling distributions are those of t_L and t_U for the parameter values that maximize type I error under H_0. For t_L, type I error is maximized when $\theta_1 - \theta_2 = \delta_L$. Type I error is maximized when $\theta_1 - \theta_2 = \delta_U$ for t_U. The grey curves in the left and right plots of Figure 1.4 visualize these sampling distributions for t_L and t_U. The 5% probability of making a type I error in either case is depicted by the shaded grey area to the right of the solid black

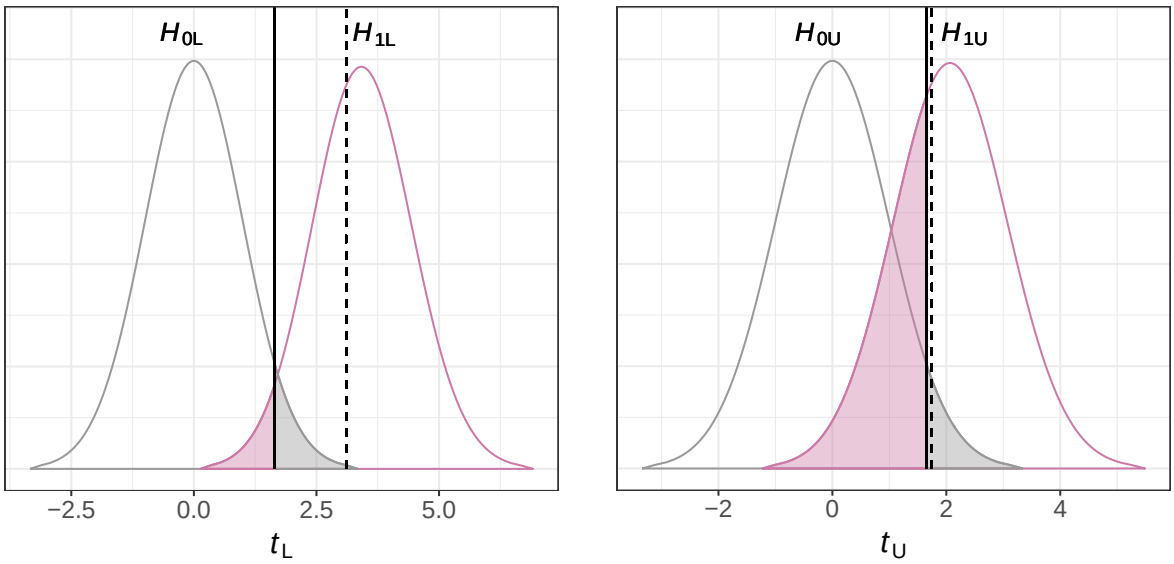


Figure 1.4: Sampling distributions for t_L (left) and t_U (right) when $n = 100$. The solid black lines denote the 0.95-quantiles of the null distributions. The type I and II error rates for each test are respectively depicted by the grey and pink shaded areas. The overall type II error rate is the proportion of the pink distributions to the left of the dotted black lines.

line at the 0.95-quantile of Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

The third sampling distribution we consider is the joint sampling distribution for t_L and t_U under the parameter values that characterize the anticipated difference for $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$, which is 0.74 grams here. The relevant joint sampling distribution is a bivariate, non-central t-distribution with singular covariance matrix. The corresponding marginal sampling distributions for t_L and t_U are visualized by the pink curves in Figure 1.4. The 3.82% probability of making a type II error with respect to H_{0L} is represented by the pink shaded area in the left plot. The probability of making a type II error with respect to H_{0U} is 33.74%, corresponding to the pink shaded area in the right plot.

For frequentist equivalence tests, the overall type II error rate that jointly accounts for t_L and t_U is bounded below by the maximum type II error rate for H_{0L} and H_{0U}. For the design in Figure 1.4, the overall type II error rate is 37.23%. This type II error rate is visualized by the proportion of the sampling distributions under H_{1L} and H_{1U} that are to the left of the dotted black lines. The power of this equivalence test is therefore 62.77%. All sampling distributions in Figure 1.4 follow known t-distributions, and analytical design

methods are suitable when using Student's t-tests. In Chapter 2, we demonstrate that analytical approaches to design such equivalence tests are unstable when relaxing the equal variance constraint that Student's t-tests impose. Simulation can be used to estimate the relevant sampling distributions in that event. Since the design of equivalence tests is predicated on quantiles of those sampling distributions, methods that consider sampling distribution segments would be efficient.

1.1.4 Bayesian Design

Bayesian methods (Gelman et al., 2020) for statistical inference treat the parameter $\boldsymbol{\eta}_j$ as a random variable for group $j = 1, 2$. As in Section 1.1.3, the characteristic θ_j is a function of the parameter(s) $\boldsymbol{\eta}_j$: $\theta_j = g(\boldsymbol{\eta}_j)$. Bayesian inference employs Bayes' theorem to update the beliefs about the random variable $\boldsymbol{\eta}_j$ as more information becomes available via the observed data. Bayesian methods also require the specification of a prior distribution for the parameter(s) $\boldsymbol{\eta}_j$, denoted by $p(\boldsymbol{\eta}_j)$. This distribution characterizes the beliefs about $\boldsymbol{\eta}_j$ prior to observing any data. For a particular statistical model, we let $L(\boldsymbol{\eta}_j; \boldsymbol{y}_j)$ be the relevant likelihood function for the parameter(s) $\boldsymbol{\eta}_j$ with respect to the observed data $\boldsymbol{y}_j$. Bayesian methods for statistical inference are facilitated via the posterior distribution of $\boldsymbol{\eta}_j$, denoted by $p(\boldsymbol{\eta}_j | \boldsymbol{y}_j)$. This distribution communicates the values of $\boldsymbol{\eta}_j$ that are plausible given the observed data and prior beliefs. By Bayes' Theorem, we have that

$$p(\boldsymbol{\eta}_j | \boldsymbol{y}_j) = \frac{L(\boldsymbol{\eta}_j; \boldsymbol{y}_j) p(\boldsymbol{\eta}_j)}{\int L(\boldsymbol{\eta}_j; \boldsymbol{y}_j) p(\boldsymbol{\eta}_j) d\boldsymbol{\eta}_j} \propto L(\boldsymbol{\eta}_j; \boldsymbol{y}_j) p(\boldsymbol{\eta}_j). \tag{1.3}$$

The posterior on $\boldsymbol{\eta}_j$ induces a posterior on the characteristic θ_j through the function $g(\cdot)$.

Several Bayesian interval hypothesis testing methods exist, including approaches with posterior probabilities, Bayes factors, and credible intervals. While hypothesis testing with Bayes factors and credible intervals will be introduced later in this subsection, the Bayesian design methods developed in this thesis emphasize hypothesis tests with posterior probabilities. Testing methods based on posterior probabilities have been introduced in various settings (see e.g., Berry et al. (2011); Brutti et al. (2014); Stevens and Hagar (2022)). Given data observed from two groups, the posterior probability that H_1 is true for a difference-based comparison is as follows:

$$Pr(\delta_L < \theta_1 - \theta_2 < \delta_U \mid \boldsymbol{y}_1, \boldsymbol{y}_2). \tag{1.4}$$

The Bayesian paradigm readily accommodates ratio-based comparisons, and we con-

trast comparisons based on differences and ratios after introducing all three methods for Bayesian hypothesis testing. For positive characteristics θ_1 and θ_2, the posterior of θ_1/θ_2 facilitates two-group comparisons via percentage increases with the hypothesis $H_1 : \theta_1/\theta_2 \in (\delta_L, \delta_U)$. Given data observed from both groups, the relevant posterior probability is

$$Pr\left(\delta_L < \theta_1/\theta_2 < \delta_U \mid \boldsymbol{y}_1, \boldsymbol{y}_2\right). \tag{1.5}$$

For ratio-based comparisons, the intervals $(\delta_L, \delta_U) = \{(1, \infty), (\delta^{-1}, \delta)\}$ for some $\delta > 1$ may be used to respectively assess whether θ_1 is superior or practically equivalent to θ_2.

The probability in (1.4) or (1.5) is compared to a critical value $0.5 \leq \gamma < 1$. If that probability is greater than γ, one should conclude $\theta \in (\delta_L, \delta_U)$ for $\theta = \theta_1 - \theta_2$ or $\theta = \theta_1/\theta_2$. Larger values of γ allow one to draw conclusions with more conviction. In contrast to hypothesis testing with p-values, there is not a widely accepted threshold for decision making with posterior probabilities. More guidance for determining a critical value γ will be provided in Chapter 5.

Accommodating a wider variety of distributional assumptions and characteristics θ_1 and θ_2 is often more straightforward in Bayesian settings. To illustrate this, we reconsider the mice example from Section 1.1.2 under different distributional assumptions. We now assume that the body weights in litter j of generation 11 are independently distributed such that $y_{ij} \sim \text{GAMMA}(\alpha_j, \lambda_j)$ with shape parameter α_j and rate parameter λ_j for $i = 1, ..., n_j$. We also summarize the body weight distributions via their medians: $\theta_j = g(\boldsymbol{\eta}_j)$, where $g(\cdot)$ is an implicit function of the gamma model parameters $\boldsymbol{\eta}_j = (\alpha_j, \lambda_j)$. The sample medians for these data are $\hat{\theta}_1 = 35.34$ grams and $\hat{\theta}_2 = 34.60$ grams as visualized in Figure 1.2. We independently assign uninformative $\text{GAMMA}(2, 0.25)$ priors to the shape α_j and rate λ_j parameters for all posterior analyses in this subsection. We use an equivalence test with $\theta = \theta_1 - \theta_2$, $(\delta_L, \delta_U) = (-3, 3)$, and a critical value of $\gamma = 0.9$ to illustrate hypothesis testing with posterior probabilities. Using computational methods described in Section 1.3.1, we estimate $Pr(\theta \in (-3, 3) \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ as 0.9841. Because $0.9841 > 0.9$, we conclude the two median weights are practically equivalent when $\delta = 3$ grams.

We now suppose there is interest in designing this hypothesis test with posterior probabilities to illustrate how sampling distributions are used to design Bayesian hypothesis tests. The relevant posterior probabilities for such hypothesis tests can be viewed as test statistics. Unlike most frequentist test statistics, the sampling distributions of posterior probabilities do not have known parametric forms. We must estimate these sampling distributions by simulating data under various generation processes and computing the corresponding posterior probabilities $Pr(H_1 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$.

Bayesian equivalence tests are considered via a single set of hypotheses H_0 and H_1. In

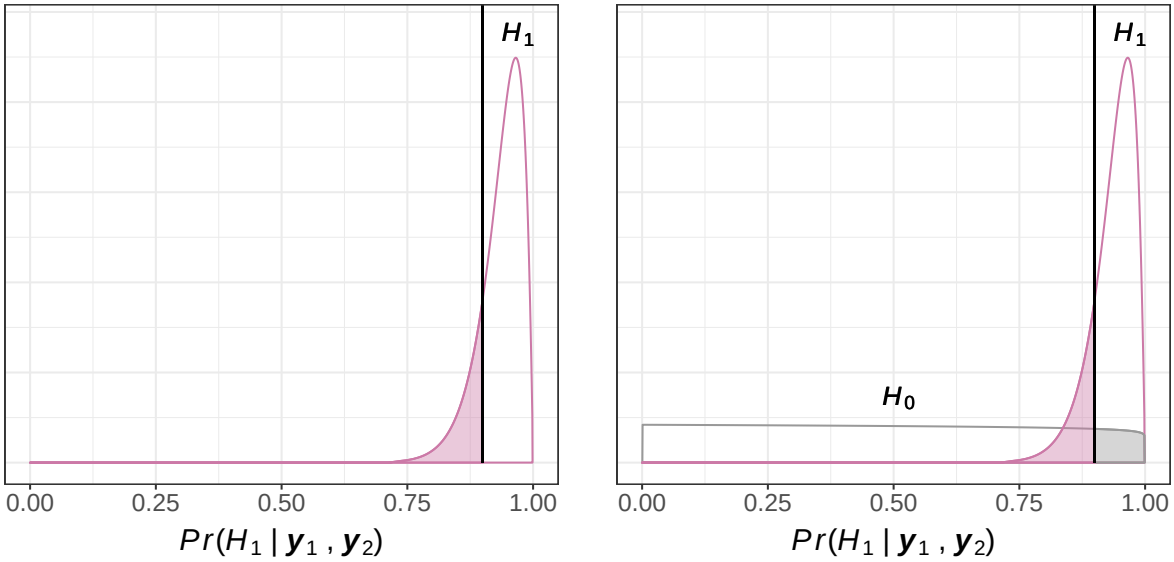


Figure 1.5: Example sampling distributions for $Pr(H_1 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ under H_1 and H_0 when $n = 100$. The solid black lines denote $\gamma = 0.9$. The type I and II error rates are respectively depicted by the grey and pink shaded areas.

the Bayesian paradigm, design for equivalence tests and design for one-sided hypothesis tests similarly leverage sampling distributions. When designing Bayesian hypothesis tests, practitioners might only be concerned with bounding the type II error rate. In that event, mice body weights would be simulated from gamma models with parameter values that align with H_1. More detail regarding how these data could be generated is provided in Chapters 3 and 5. In this case, we need only consider the sampling distribution of posterior probabilities under H_1 to assess the power of a test with respect to a critical value γ.

An example of one such sampling distribution is given by the pink curve in the left plot of Figure 1.5 for the sample size $n = n_1 = n_2 = 100$. The 18.37% probability of making a type II error is represented by the pink shaded area to the left of the black line at $\gamma = 0.9$. When $n = 100$, a target power of $1 - \beta = 0.8$ would be achieved. However, precisely estimating the right tail of the pink distribution does not help us determine whether $n = 100$ is a suitable sample size. We need only determine whether the β-quantile of the sampling distribution of posterior probabilities under H_1 exceeds $\gamma = 0.9$. Thus, it would be beneficial if we could efficiently estimate the 0.2-quantile of the pink sampling distribution without estimating the sampling distribution in its entirety.

If practitioners want to bound both the type I and II error rates of a Bayesian hypothesis test, a second sampling distribution of posterior probabilities must also be considered. For this example, that sampling distribution would leverage mice body weights simulated from gamma models with parameter values that align with H_0. While the data align with H_0, the relevant posterior probability is still $Pr(H_1 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$. The grey curve in the right plot of Figure 1.5 depicts an example sampling distribution under H_0 for $n = 100$. When $\gamma = 0.9$, the 8.91% probability of making a type I error is represented by the grey shaded area. It follows that the (n, γ) combination of $(100, 0.9)$ would attain a target power of 0.8 but not achieve a type I error rate of $\alpha = 0.05$. In practice, the sample size n and critical value γ could be chosen jointly to satisfy criteria for power and the type I error rate.

Precise estimation of the left tail of the grey sampling distribution does not help us determine whether a sample size is suitable with respect to type I error. To determine a (n, γ) combination such that criteria for both power and the type I error rate are satisfied, we need only discern whether the $(1 - \alpha)$-quantile of the sampling distribution under H_0 exceeds the β-quantile of that under H_1. In the right plot of Figure 1.5, the 0.95-quantile of the sampling distribution under H_0 (0.9423) exceeds the 0.2-quantile of that under H_1 (0.9035). It is computationally intensive to estimate two sets of entire sampling distributions of posterior probabilities for various sample sizes. Design with sampling distribution segments would therefore make Bayesian hypothesis testing more accessible.

We now briefly introduce alternative methods for Bayesian interval hypothesis testing that involve Bayes factors and credible intervals. In Chapter 3, we demonstrate that design for these hypothesis tests can also be considered using a framework based on sampling distributions of posterior probabilities. Thus, we will consider the sampling distributions associated with these hypothesis testing methods more thoroughly in that chapter. Morey and Rouder (2011) proposed the nonoverlapping hypotheses (NOH) approach to assess the plausibility of interval hypotheses with Bayes factors (Jeffreys, 1935; Kass and Raftery, 1995). This approach directly assigns a prior distribution to the effect size considered via the hypothesis test, which is $\theta = \theta_1 - \theta_2$ or $\theta = \theta_1/\theta_2$ in this case. The NOH Bayes factor is the ratio of the posterior odds of the complementary hypotheses $H_1 : \theta \in (\delta_L, \delta_U)$ and $H_0 : \theta \notin (\delta_L, \delta_U)$ to their prior odds:

$$\frac{Pr(\delta_L < \theta < \delta_U \mid \boldsymbol{y}_1, \boldsymbol{y}_2)}{1 - Pr(\delta_L < \theta < \delta_U \mid \boldsymbol{y}_1, \boldsymbol{y}_2)} \div \frac{Pr(\delta_L < \theta < \delta_U)}{1 - Pr(\delta_L < \theta < \delta_U)}. \tag{1.6}$$

The NOH Bayes factor provides support for H_1 over H_0 when its value is greater than a predetermined threshold $K \geq 1$.

Hypothesis testing methods with credible intervals have also been proposed (Gubbiotti

and De Santis, 2011; Brutti et al., 2014; Kruschke, 2018). These methods compare the credible interval for the posterior of a univariate parameter $\theta = \theta_1 - \theta_2$ or $\theta = \theta_1/\theta_2$ to the interval (δ_L, δ_U). A credible interval $(L_{\theta,1-\alpha}, U_{\theta,1-\alpha})$ has coverage of $1-\alpha$ if $Pr(\theta \in (L_{\theta,1-\alpha}, U_{\theta,1-\alpha}) \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = 1-\alpha$. Since this interval is not uniquely defined, the equal-tailed credible interval or highest density interval (HDI) is often considered. The equal-tailed interval is defined such that $Pr(\theta < L_{\theta,1-\alpha} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = Pr(\theta > U_{\theta,1-\alpha} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = 1 - \alpha/2$, and the HDI is the narrowest credible interval with coverage $1-\alpha$. If $(L_{\theta,1-\alpha}, U_{\theta,1-\alpha})$ lies entirely within (δ_L, δ_U), one should conclude $\theta \in (\delta_L, \delta_U)$.

Lastly, we contrast the conclusions prompted by Bayesian hypothesis tests based on differences with the conclusions prompted by those based on ratios. First, posterior probabilities are invariant to monotonic transformations, so the probability in (1.5) is equivalent to that from (1.4) when logarithmic transformations are applied to θ_1, θ_2, and the interval endpoints. We would therefore draw the same conclusions using hypothesis tests with (1.4) and (1.5) given appropriate logarithmic transformations. NOH Bayes factors and equal-tailed credible intervals are also invariant to monotonic transformations, but posterior HDIs are not. Thus, hypothesis tests with posterior HDIs are not guaranteed to yield the same conclusions when considering a ratio and its corresponding difference on the logarithmic scale. Ratio-based comparisons and their sampling distributions will be considered more thoroughly in Chapter 3.

1.2 Simulation-Based Design

1.2.1 The Unit Hypercube

To design studies that use flexible and realistic statistical models, computer simulation is generally required. These approaches to study design involve simulating hypothetical samples of data under various data generation processes to estimate sampling distributions. The type of data is dictated by the context for the study. For instance, these data may be nonnegative survival times for cancer treatments, binary indicators that dictate whether products from assembly lines are defective, or integer customer satisfaction scores for consumer goods.

Regardless of the data type, each observation is generated using at least one number in the interval $[0, 1]$. These data are typically generated from a distribution with a known inverse cumulative distribution function (CDF): $F^{-1}(u)$ for $u \in [0, 1]$. In such scenarios, only one number $u_i \in [0, 1]$ is required to generate an observation $y_i = F^{-1}(u_i)$ (Kroese

et al., 2013). More than one number between 0 and 1 may be required to simulate each observation in more complex settings that are not the focus of this thesis (see e.g., Casella et al. (2004)).

Simulation-based design is therefore predicated on sequences $\{\boldsymbol{u}_r\}_{r=1}^m \in [0,1]^d$, where m is the number of simulation repetitions and d is the dimension of the simulation. The *unit hypercube* $[0,1]^d$ is a closed, compact, convex hull that encompasses all possible combinations of d numbers between 0 and 1 (Balas and Jeroslow, 1972). When CDF inversion is used, the dimension d is often a function of the sample size n for the study. For each simulation repetition r, the point $\boldsymbol{u}_r$ yields a sample to be investigated via a hypothetical study. Many hypothetical studies – and hence large m – are required to reliably estimate the relevant sampling distributions. The computational cost of this computer simulation impedes the adoption of non-simplistic designs. In this thesis, we explore the unit hypercube $[0,1]^d$ as a conduit for the data space to develop general methods for study design. The remainder of this subsection overviews several approaches to generate points $\{\boldsymbol{u}_r\}_{r=1}^m$ from $[0,1]^d$.

1.2.2 Monte Carlo Simulation

Monte Carlo methods broadly leverage repeated random sampling to estimate deterministic quantities (Metropolis and Ulam, 1949). Random number generation can be facilitated using the results from physical experiments (Schindler, 2009); however, Monte Carlo simulation typically leverages deterministic pseudorandom number generators (Gentle, 2003) to construct sequences from the unit hypercube. Pseudorandom number generators aim to output sequences $\{\boldsymbol{u}_r\}_{r=1}^m \in [0,1]^d$ that are *approximately* independently and identically distributed (i.i.d.) over the unit hypercube.

Pseudorandom sequences are often created using linear congruential generators (Fishman and Moore, 1986). The quality of the approximately i.i.d. sequences output by pseudorandom number generators can be quantified using various theoretical and statistical tests, such as the spectral test (Knuth, 2014) and serial tests (L'Écuyer et al., 2002; Knuth, 2014). The Mersenne Twister (Matsumoto and Nishimura, 1998) is the default pseudorandom number generator in R, and certain alternative generators (see e.g., L'Écuyer (1999)) can also safely be used. More details on pseudorandom number generators that poorly approximate i.i.d. sequences can be found elsewhere (Entacher, 1998; L'Écuyer, 2001). Figure 1.6 visualizes an example pseudorandom sequence in two dimensions with $m = 64$ points generated using the Mersenne Twister. These points appear to reasonably approximate random scatter throughout the unit square.

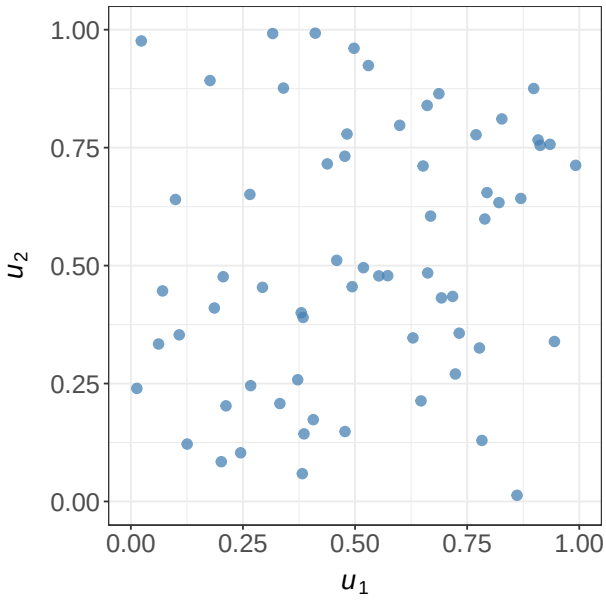


Figure 1.6: Pseudorandom sequence in $[0,1]^2$ with $m = 64$ points.

To prove theoretical results with satisfactory pseudorandom number generators, each point in the constructed sequence is assumed to be i.i.d. over the unit hypercube. That is, $\boldsymbol{U}_r \overset{\text{i.i.d.}}{\sim} U\left([0,1]^d\right)$ for $r = 1, ..., m$. It follows that pseudorandom sequences can be used in Monte Carlo simulation to prompt unbiased estimators via empirical means:

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^{m}\Psi(\boldsymbol{U}_r)\right) = \int_{[0,1]^d}\Psi(\boldsymbol{u})d\boldsymbol{u}, \tag{1.7}$$

for some function $\Psi(\cdot)$ such that the expectation in (1.7) is finite. As discussed in various contexts throughout this thesis, we take the function $\Psi(\cdot)$ to define power or the type I error rate. The unbiasedness in (1.7) makes generating data with pseudorandom sequences standard practice in simulation-based design. For design purposes, *entire* sampling distributions are estimated when using pseudorandom sequences. In the following subsection, we argue that exploring the unit hypercube $[0,1]^d$ and corresponding data space with pseudorandom sequences is computationally inefficient.

1.2.3 Quasi-Monte Carlo Methods

Low-discrepancy sequences are created to induce negative dependence between the points $\boldsymbol{U}_1, ..., \boldsymbol{U}_m$. Low-discrepancy sequences are regularly incorporated into quasi-Monte Carlo methods (Lemieux, 2009). These methods are frequently used in the financial sector but underutilized in experimental design. There are two main families of constructions used to create low-discrepancy sequences: lattices and digital sequences. Lattices construct sequences by taking select linear combinations of basis vectors (Korobov, 1959; Hickernell et al., 2000; Hickernell and Niederreiter, 2003). Digital sequences are constructed by leveraging integer expansion in particular bases (Halton, 1960; Sobol', 1967; Faure, 1982). The notion of discrepancy refers to the distance between the empirical distribution induced by a sequence and the uniform distribution over $[0,1]^d$. This discrepancy and the quality of the low-discrepancy sequence is often assessed using the star discrepancy (Niederreiter, 1992). Alternative evaluation metrics have also been proposed for lattice rules (see e.g., L'Écuyer and Lemieux (2000)) and digital sequences (see e.g., Wiart et al. (2021)).

Low-disrepancy sequences can be randomized to yield estimators with better consistency properties than those created using purely deterministic sequences. Sequences created using lattice rules are often randomized using random shifts or rotation sampling (Cranley and Patterson, 1976). The randomization of digital sequences can be carried out via a digital shift (Lemieux, 2009). For appropriately randomized low-discrepancy sequences, each point in the sequence is such that $\boldsymbol{U}_r \sim \mathcal{U}\left([0,1]^d\right)$ for $r = 1, ..., m$. These sequences are *not* i.i.d. over the unit hypercube since the points are not independent. Randomization approaches for low-discrepancy sequences should inject enough randomness to ensure $\{\boldsymbol{U}_r\}_{r=1}^m \sim \mathcal{U}\left([0,1]^d\right)$ but not too much variability to compromise the equidistribution properties.

We emphasize that quasi-Monte Carlo methods do not accurately estimate entire sampling distributions due to the negative dependence between the points in low-discrepancy sequences. However, the result in (1.7) holds for appropriately randomized low-discrepancy sequences since $\{\boldsymbol{U}_r\}_{r=1}^m \sim \mathcal{U}\left([0,1]^d\right)$. These sequences can therefore be used similarly to pseudorandom sequences in Monte Carlo simulation to prompt unbiased estimators based on empirical means. Throughout this thesis, we demonstrate how the operating characteristics of hypothesis tests can be estimated using empirical means like the one in (1.7). The result in (1.7) therefore justifies the use of quasi-Monte Carlo methods in this thesis. Throughout this thesis, we state that we *explore* – instead of *estimate* – sampling distributions when using low-discrepancy sequences.

Due to the negative dependence between the points, the variance of the estimator in

(1.7) is typically reduced by using randomized low-discrepancy sequences. We have that

$$\text{Var}\left(\frac{1}{m}\sum_{r=1}^{m}\Psi(\boldsymbol{U}_r)\right) = \frac{\text{Var}\left(\Psi(\boldsymbol{U}_r)\right)}{m} + \frac{2}{m^2}\sum_{r=1}^{m}\sum_{t=r+1}^{m}\text{Cov}\left(\Psi(\boldsymbol{U}_r), \Psi(\boldsymbol{U}_t)\right), \tag{1.8}$$

where the the first term on the right side of (1.8) is the variance of the corresponding estimator based on pseudorandom sequences of approximately i.i.d. points. Low-discrepancy sequences give rise to effective variance reduction methods when the dimension of the simulation is moderate (Kocis and Whiten, 1997). However, high-dimensional low-discrepancy sequences may have poor low-dimensional projections, which can lead to a deterioration in performance (see e.g., Braaten and Weller (1979); Fox (1986)). Substantial work on quasi-Monte Carlo methods has targeted suitable performance for $d \leq 32$ (L'Écuyer and Lemieux, 2000; Lemieux, 2009). By (1.8), randomized low-discrepancy sequences reduce the number of simulation repetitions m required to precisely and consistently estimate (1.7) and hence power and the type I error rate compared to using pseudorandom alternatives.

When implementing quasi-Monte Carlo methods, this thesis uses a particular class of low-discrepancy sequences called Sobol' sequences (Sobol', 1967) that are based on integer expansion in base 2. We use Sobol' sequences because they are well studied and can be constructed using existing software in popular programming languages. Moreover, subsequences of the Sobol' sequence are also low discrepancy, and this property will be exploited in Chapter 5. However, the methods proposed in this thesis could be implemented with various types of low-discrepancy (and pseudorandom) sequences. Sobol' sequences can be generated and randomized in R using the `qrng` package (Hofert and Lemieux, 2020). Figure 1.7 compares pseudorandom and Sobol' sequences in two dimensions with $m = 64$ points. The red points from the Sobol' sequence appear to be more evenly distributed throughout the unit square than the blue ones from the pseudorandom sequence.

Since low-discrepancy sequences are comprised of dependent points and may perform poorly in high-dimensional settings, this thesis will not use them to directly explore the data space. Low-discrepancy sequences will instead be used to explore lower-dimensional hypercubes prompted by data summaries. Both pseudorandom and low-discrepancy sequences thoroughly explore the unit hypercube for each potential design, but this thorough exploration is computationally inefficient in many design settings. In Section 1.3.2, we introduce a general framework to explore the unit hypercube in a nonuniform, targeted manner. This framework will be used to expedite simulation-based design by way of exploring sampling distribution segments.

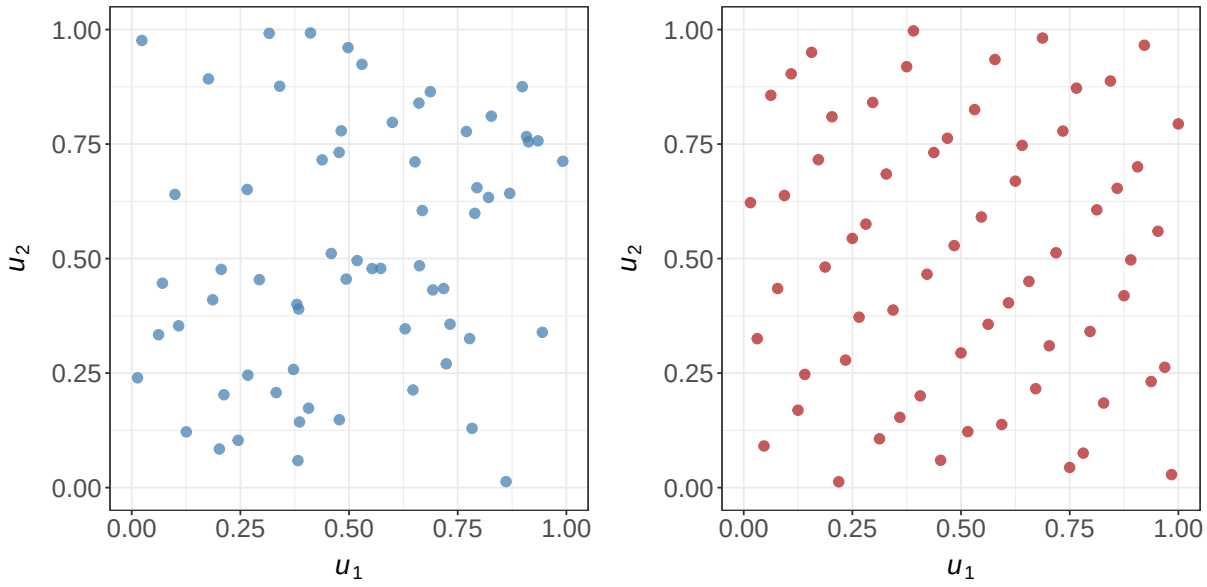


Figure 1.7: Pseudorandom (left) and Sobol' (right) sequences in $[0,1]^2$ with $m = 64$ points.

1.3 Barriers to Scalable Design

1.3.1 Computationally Complex Simulation Repetitions

The nature of the computational complexity associated with each simulation repetition $r = 1, ..., m$ is different in frequentist and Bayesian settings. Design methods for frequentist studies are generally less computationally intensive. For each simulation repetition, data $\boldsymbol{y}_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})^T$ for each group $(j = 1, 2)$ are typically generated under H_1 using a point $\boldsymbol{u} \in [0,1]^{n_1+n_2}$. The data are summarized by sufficient statistics or maximum likelihood estimates that are used to compute the relevant test statistic. Based on the value of this test statistic, we decide whether or not to reject H_0. The proportion of the m simulation repetitions in which H_0 is rejected estimates the power of a potential design.

It can be cumbersome to generate, store, and summarize the data $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ when n_1 and n_2 are large. In this thesis, we therefore simulate sufficient statistics or maximum likelihood estimates directly. This practice is advantageous in that the dimension of the unit hypercube is reduced. If the statistical model for group j is parameterized by $\boldsymbol{\eta}_j \in \mathbb{R}^d$, we can typically generate the sufficient statistics or maximum likelihood estimates for both groups using a point $\boldsymbol{u} \in [0,1]^{2d}$. The dimension of the parameter space d is generally

such that $2d << n_1 + n_2$. This dimension reduction allows for greater incorporation of quasi-Monte Carlo methods into our design framework. We can often readily find a suitable low-discrepancy sequence of dimension $2d$, but it may be challenging to find an appropriate sequence of dimension $n_1 + n_2$ for large sample sizes.

As demonstrated in Chapter 3, the test statistics for most Bayesian hypothesis tests can be expressed as posterior probabilities. Unlike for parametric frequentist hypothesis tests, these test statistics usually cannot be computed as an explicit function of the sufficient statistics or maximum likelihood estimates. Given $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ and the joint posterior $p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$, the posterior probability may be calculated as

$$\iint_{\mathcal{R}} p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{y}_1, \boldsymbol{y}_2) d\boldsymbol{\eta}_1 d\boldsymbol{\eta}_2, \tag{1.9}$$

where $\mathcal{R}$ is the region corresponding to H_1 (i.e., $\delta_L \leq \theta_1 - \theta_2 \leq \delta_U$ or $\delta_L \leq \theta_1/\theta_2 \leq \delta_U$) for the posterior probabilities in (1.4) and (1.5), respectively. When conjugate priors are used, the integral in (1.9) can be evaluated analytically. For flexible study design, computational methods can be used to approximate the posterior $p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ by way of generating posterior draws.

The most popular computational method for posterior approximation is Markov chain Monte Carlo (MCMC) simulation (Gelman et al., 2020). Where applicable, this thesis implements MCMC via Gibbs sampling (Geman and Geman, 1984) with the `rjags` package in R (Plummer, 2019). Gibbs sampling allows us to sample from joint posterior distributions using simpler conditional posteriors. We typically generate M observations from each relevant Markov chain, which yields draws $\theta_{1,1}, \theta_{1,2}, ..., \theta_{1,M}$ and $\theta_{2,1}, \theta_{2,2}, ..., \theta_{2,M}$. We can use these posterior draws to estimate relevant posterior probabilities. For instance, the posterior probability from (1.4) can be estimated as

$$\frac{1}{M} \sum_{k=1}^{M} \mathbb{I}(\delta_L < \theta_{1,k} - \theta_{2,k} < \delta_U). \tag{1.10}$$

Sampling-resampling methods can also be used to obtain an approximate sample from a *continuous* posterior distribution $p(\boldsymbol{\eta}_j | \boldsymbol{y}_j)$. One such method is the sampling-importance-resampling (SIR) algorithm (Rubin, 1987, 1988; Smith and Gelfand, 1992). In this setting, a sample can be readily generated from a continuous proposal distribution $c(\boldsymbol{x})$. However, we want to sample from a distribution $d(\boldsymbol{x})$ such that $c(\boldsymbol{x}) = 0$ implies that $d(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$. We also require a function $b(\boldsymbol{x})$ such that $d(\boldsymbol{x}) \propto b(\boldsymbol{x})$. Because $p(\boldsymbol{\eta}_j | \boldsymbol{y}_j) \propto$

$L(\boldsymbol{\eta}_j; \boldsymbol{y}_j)p(\boldsymbol{\eta}_j)$ by Bayes' Theorem, natural choices exist for the function $b(\boldsymbol{x})$ required to implement the SIR algorithm. Even so, choosing a suitable proposal distribution $c(\boldsymbol{x})$ that resembles $p(\boldsymbol{\eta}_j|\boldsymbol{y}_j) = d(\boldsymbol{x})$ is less straightforward. We therefore prefer MCMC methods over sampling-resampling approaches, but we use the SIR algorithm in Chapters 4 and 5 when MCMC methods are difficult to implement.

Although we prefer MCMC methods to the SIR algorithm, neither computational approach is ideal when we must conduct a large number of simulation repetitions m for each potential design explored. We instead recommend using large-sample normal approximations to the posterior of $\boldsymbol{\eta}_j$, such as those based on the Bernstein-von Mises (BvM) theorem (van der Vaart, 1998) or the Laplace approximation (Gelman et al., 2020). These analytical approximations to the posterior will be formally introduced later in this thesis. We will also discuss strategies to improve the quality of these large-sample normal approximations for moderate n. These analytical approximation methods mitigate the computational complexity associated with each simulation repetition in Bayesian design.

1.3.2 Inefficient Estimation of Sampling Distributions

Thorough exploration of $[0, 1]^{2d}$ prompts the estimation of test statistics from throughout the relevant sampling distributions. This standard practice is often wasteful due to the repetitive nature of simulation-based design. These simulations investigate how various design inputs impact the operating characteristics of a hypothesis test. Potential design inputs include minimum detectable effect sizes, variability estimates, parametric statistical distributions, and sample sizes. For each combination of design inputs considered, simulation is used to carry out a hypothetical study. The same procedure is used to implement each simulation repetition, and the only differences between simulation repetitions are driven by the generated data. This process is repeated *for each* combination of design inputs that are investigated.

Quasi-Monte Carlo methods allow us to reduce the number of simulation repetitions m required to estimate the operating characteristics of a hypothesis test. However, computational resources are still wasted by estimating test statistics from throughout the relevant sampling distributions for unsuitable sample sizes. Algorithmic methods are often employed to explore the sample size space. For instance, Wang and Gelfand (2002) recommended using bisection methods or grid searches to find a suitable sample size n. If we develop sound design procedures that only use points from subspaces of $[0, 1]^{2d}$ to explore each sample size, we can consider sampling distribution segments and further improve the computational savings prompted by using quasi-Monte Carlo methods alone.

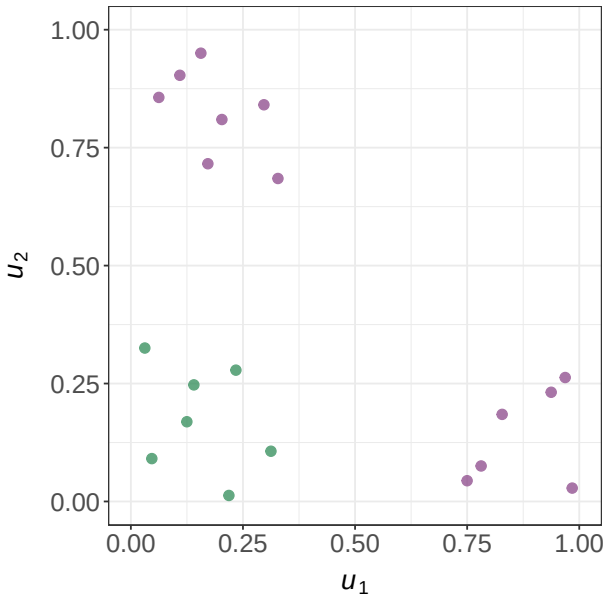


Figure 1.8: Illustration of targeted hypercube exploration.

This thesis emphasizes methods for sample size determination. In this context, uniform exploration of $[0,1]^{2d}$ is inefficient because similar points $\boldsymbol{u}_r$ and $\boldsymbol{u}_{r'}$ respectively used to consider sample sizes n_A and n_B should prompt studies with similar results when $n_A \approx n_B$. Much smaller m could be used if the points $\{\boldsymbol{u}_r\}_{r=1}^m$ selected for each sample size n were concentrated in interesting subspaces of $[0,1]^{2d}$. In this thesis, interesting subspaces of $[0,1]^{2d}$ correspond to segments of sampling distributions near the quantiles that define the operating characteristics of a hypothesis test. Figure 1.8 illustrates the notion of exploring the unit hypercube in a targeted manner. For the sample size n_A, only the green points in Figure 1.8 may be of interest, whereas perhaps the purple points prompt more interesting studies for the sample size n_B. With targeted exploration approaches, we may not need to carry out studies corresponding to uninteresting regions of $[0,1]^{2d}$ to assess the operating characteristics of a hypothesis test.

By reducing the number of repetitions m, such approaches would vastly improve the accessibility of simulation-based design. However, these approaches are very rarely implemented for two reasons. First, it is often unclear which $[0,1]^{2d}$-subspaces should be prioritized for arbitrary designs. Second, care must be taken to not compromise the simulation consistency since the unbiasedness in (1.7) relies on uniform exploration of the unit hypercube. The methods proposed in this thesis should therefore automate the $[0,1]^{2d}$-subspace

selection process for arbitrary designs. The simulation dimension $2d$ is not related to the sample size n, so the unit hypercube $[0,1]^{2d}$ is the same for all sample sizes (and other design inputs) considered.

Existing work on simulation with subspaces of the unit hypercube that correspond to segments of sampling distributions is limited. Thus, this thesis develops a new framework for scalable design. The contributions of this thesis are directed toward two main objectives. First, we aim to propose methods that are straightforward for practitioners to implement. We also intend to mathematically prove that simulation consistency can be maintained while ignoring certain subspaces of the unit hypercube – and the corresponding sampling distribution segments – under various conditions.

1.4 Contributions

The remainder of this chapter outlines the contributions of this thesis and the structure of this document. Chapter 2 illustrates how design with sampling distribution segments prompts power analysis in nonstandard frequentist design settings. Power analyses are typically carried out using integration when the null distributions have known parametric forms based on pivotal quantities. When the relevant test statistics cannot be constructed from pivotal quantities, their sampling distributions are approximated via repetitive, time-intensive computer simulation. We propose a novel simulation-based method to quickly approximate the power curve for many such hypothesis tests by efficiently exploring segments of the relevant sampling distributions. Despite not estimating the entire sampling distribution, this approach prompts unbiased sample size recommendations. We illustrate this method using two-group equivalence tests with unequal variances and overview its broader applicability in simulation-based design.

Chapter 3 describes a comprehensive framework for power curve approximation with Bayesian hypothesis tests conducted via posterior probabilities, Bayes factors, and credible intervals. We propose a framework for power curve approximation with such hypothesis tests that assumes data are generated using statistical models with known, fixed parameters for the purposes of sample size determination. When the conditions for the Bernstein-von Mises theorem are satisfied, we propose an approach to explore segments of the approximate sampling distribution of posterior probabilities for each sample size considered. These sampling distributions are used to construct power curves for various types of posterior analyses. Our resulting method for power curve approximation is orders of magnitude faster than conventional power curve estimation for Bayesian hypothesis tests. We also prove the consistency of the corresponding power estimates and sample size recommendations under

certain conditions. In this chapter, we only consider criteria for the power of the hypothesis test and ignore the notion of type I error.

Chapter 4 prepares the eventual relaxation of the fixed-parameter assumption for the data generation process considered in Chapter 3 by means of prior specification. In fully Bayesian analyses, prior distributions are specified before observing data. Prior elicitation methods transfigure prior information into quantifiable prior distributions. Recently, methods that leverage copulas have been proposed to accommodate more flexible dependence structures when eliciting multivariate priors. We prove that under broad conditions, the posterior cannot retain many of these flexible prior dependence structures in large-sample settings. We also overview several objectives for prior specification to help practitioners select prior dependence structures that align with their objectives for posterior analysis. Because correctly specifying the dependence structure a priori can be difficult, we consider how the choice of prior copula impacts the posterior distribution in terms of asymptotic convergence of the posterior mode. Our resulting recommendations streamline the prior elicitation process and are considered when implementing the design methods in Chapter 5.

Chapter 5 proposes a framework for design that serves as an alternative to the approach explored in Chapter 3. Unlike in Chapter 3 where we consider design for fixed parameter values, this comprehensive framework accounts for uncertainty in the parameter values used to generate the data. To design trustworthy Bayesian studies using this approach, criteria for power *and* the type I error rate are defined. These posterior-based operating characteristics are typically assessed by exploring entire sampling distributions of posterior probabilities via simulation. In this chapter, we propose a scalable method to determine optimal sample sizes and decision criteria that maps posterior probabilities to low-dimensional conduits for the data. Our method leverages this mapping and large-sample theory to consider sampling distribution segments in a targeted manner. This approach prompts consistent sample size recommendations with fewer simulation repetitions than standard methods. We repurpose the posterior probabilities computed in that approach to efficiently investigate various sample sizes and decision criteria using contour plots.

Chapter 6 concludes this thesis and discusses extensions for future work.

Chapter 2

Power Curves without Pivotal Quantities

2.1 Preamble

This chapter illustrates an initial application of design with sampling distribution segments in frequentist settings. Chapter 2 is the only chapter of this document that pertains to frequentist analyses. The inclusion of this contribution in the thesis underscores that the benefits of design with sampling distribution segments transcend paradigmatic differences in the field of statistics. The main design scenario in this chapter relaxes the equal variance assumption imposed on the equivalence test from Section 1.1.3 facilitated via Student's t-tests. Relaxing this assumption does not greatly complicate the analysis of observed data with a single equivalence test, but this relaxation creates legitimate issues for standard methods used to design such tests as detailed in this chapter. Throughout this chapter, we also emphasize how the proposed approaches based on sampling distribution segments can provide material benefit over standard methods for more complex designs. In terms of notation, we emphasize that $\bar{d}$ used in this chapter to denote the difference between two sample means is not related to the simulation dimension. This simulation dimension is generally referred to as d in this chapter instead of $2d$ as mentioned in Section 1.3.

While the Bayesian design methods proposed in the following chapters provide sample size recommendations that are *consistent*, the frequentist design methods presented in this chapter prompt *unbiased* sample size recommendations. This unbiasedness or consistency is with respect to repeated implementation of our design methods for a given two-group comparison using different pseudorandom or low-discrepancy sequences. Ideally, the ex-

pectation of the sampling distribution of these sample size recommendations is the smallest sample size that satisfies criteria for operating characteristics of a hypothesis test. When that result holds true for all possible sample sizes, the sample size recommendation is unbiased. The sample size recommendations are consistent if that result instead holds true approximately for sufficiently large sample sizes. We are able to make the stronger claim of unbiasedness for the methods in this chapter because no large-sample results are used to generate sufficient statistics or compute test statistics.

2.2 Background

Statistical studies require a substantial investment of time, funding, and human capital. It is important to ensure these resources are invested into well-designed studies that are capable of achieving their intended objectives. These objectives often involve establishing the presence or absence of meaningful effects in observational or experimental settings. In traditional hypothesis tests and equivalence tests, the study power is respectively the probability of correctly establishing the presence or absence of such effects (Chow and Liu, 2008). The study power generally increases with the sample size, and a power analysis is typically used to find the minimum sample size that achieves the desired power for a study.

A power analysis considers the sampling distributions of a relevant test statistic under two hypotheses: the null hypothesis H_0 and alternative hypothesis H_1. Under the assumption that H_0 is true, this sampling distribution is called the null distribution. For most parametric frequentist hypothesis tests, the null distribution coincides with a known statistical distribution that does not depend on the unknown model parameters. The null distribution does not depend on these parameters because the test statistics can be constructed from pivotal quantities (Shao, 2003). In contrast, the sampling distribution of the test statistic under H_1 *does* depend on the magnitude of the effect size, expressed as a function of the model parameters. Power is defined as a tail probability in the sampling distribution under H_1, where the threshold for this tail probability is called the critical value. This tail probability is straightforward to compute via integration when the null distribution is based on a pivotal quantity, but more complex methods must be used to perform power analysis otherwise.

The null distribution is not based on pivotal quantities for many studies, particularly those that leverage the Welch-Satterthwaite equation (Satterthwaite, 1946; Welch, 1947) to approximate the degrees of freedom for linear combinations of independent sample variances. This equation is applied to conduct hypothesis tests based on crossover designs (Lui, 2016), several treatments (Jan and Shieh, 2020), and sequential testing (Tartakovsky

et al., 2015). Additionally, the most common application of this equation is to compare two normal population means with unequal population variances via Welch's t-test (Welch, 1938): the default t-test in R. Even for this most basic use case, the null distribution is not based on a pivotal quantity. The null distribution for Welch's t-test approximately coincides with the standard normal distribution for large sample sizes, but this approximation based on asymptotic pivotal quantities is of limited utility since t-tests are most useful when the sample sizes are small.

This chapter presents a general framework for power analysis without the use of pivotal quantities that is primarily illustrated via two-group equivalence tests with unequal variances. We focus on this setting for three reasons. First, these tests commonly assess average bioequivalence (Chow and Liu, 2008) of two pharmaceutical drugs. Average bioequivalence compares the *mean* clinical responses for two treatments. Second, this setting allows for clear visualization of our methodology. Third, existing methods for power analysis with these designs (see e.g., *PASS* (NCSS, LLC., 2023)) produce unreliable results as demonstrated in this chapter. While this chapter emphasizes two-group equivalence tests with unequal variances, we later illustrate the use of the proposed methods with crossover designs. The methods are also generally applicable with noninferiority and one-sided hypothesis tests. The methods proposed in this chapter (as well as several extensions) can all be implemented using the `dent` package in R (Hagar and Stevens, 2024a).

Power analysis requires practitioners to choose anticipated effect sizes and variability estimates based on previous studies (Chow et al., 2008). The recommended sample sizes achieve desired statistical power when the selected response distributions, anticipated effect sizes, and variability estimates accurately characterize the underlying data generation process. Empirical power analysis prompts more flexible methods for study design when the null distribution is not based on pivotal quantities. However, simulation-based methods for power analysis have computational drawbacks. Standard practice requires simulating many samples of data to reliably approximate the sampling distribution needed to estimate power for each sample size n considered. This standard practice of estimating *entire* sampling distributions of test statistics is wasteful because study power is a tail probability in the sampling distribution under H_1 defined by a critical value. It would be more computationally efficient if we could accurately assess power for a sample size n by only exploring a *segment* of the sampling distribution that is near the critical value. The methods for power analysis proposed in this chapter adopt such an approach.

The remainder of this chapter is structured as follows. In Section 2.3, we present a method to map sampling distributions of test statistics for two-group equivalence tests with unequal variances to the unit cube. This mapping prompts unbiased power estimates given a pseudorandom or low-discrepancy sequence dispersed throughout the unit cube.

In Section 2.4, we propose a novel simulation-based method that combines the mapping from Section 2.3 with root-finding algorithms to quickly facilitate power curve approximation. This approach is fast because for a given sample size, we only explore test statistics corresponding to subspaces of the unit cube – and hence only consider a segment of the sampling distribution. Even without estimating entire sampling distributions, this method yields unbiased sample size recommendations. To illustrate the wide applicability of the proposed methodology, we extend this approach for use with hypothesis tests based on crossover designs in Section 2.5. Throughout the chapter, we also describe how the methods can be applied with more complex study designs. We provide concluding remarks and a discussion of extensions to this work in Section 2.6.

2.3 Mapping the Sampling Distribution to the Unit Cube

2.3.1 Three-Dimensional Simulation Repetitions

The results in this section are used to approximate power curves with segments of the relevant sampling distributions in Section 2.4. In this section, we describe how to map the sampling distribution of test statistics for two-group equivalence tests to $[0,1]^d$ with low dimension d. This mapping allows us to implement power analyses without necessitating the high-dimensional simulation associated with repeatedly generating data. We consider a context which prompts a three-dimensional simulation corresponding to the unit cube for illustration. In particular, suppose we collect data $y_{ij},\ i = 1, ..., n_j,\ j = 1, 2$ from the i^{th} subject in group j. We assume for illustration that the data $\boldsymbol{y}_j = \{y_{ij}\}_{i=1}^{n_j}$ for group $j = 1, 2$ are generated independently from a $\mathcal{N}(\mu_j, \sigma_j^2)$ distribution where $\sigma_1^2 \neq \sigma_2^2$. Here, $\boldsymbol{\eta}_j = (\mu_j, \sigma_j^2)$ and $\theta_j = g(\boldsymbol{\eta_j}) = \mu_j$. Interest lies in comparing θ_1 and θ_2 while accounting for unequal variances.

Given interval endpoints δ_L and δ_U, we aim to conclude that $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$ by rejecting the composite null hypothesis $H_0 : \theta_1 - \theta_2 \in (-\infty, \delta_L] \cup [\delta_U, \infty)$ in favour of the alternative hypothesis $H_1 : \theta_1 - \theta_2 \in (\delta_L, \delta_U)$. The interval is often chosen such that $(\delta_L, \delta_U) = (-\delta, \delta)$ for some equivalence margin $\delta > 0$ described in Section 1.1.3. However, the methods in this section accommodate any real $-\infty < \delta_L < \delta_U < \infty$. For such analyses, Schuirmann (1981) and Dannenberg et al. (1994) respectively proposed two one-sided test (TOST) procedures based on Student's and Welch's t-tests, with the Welch-based TOST procedure performing better than the standard version in the presence of unequal variances

(Gruman et al., 2007; Rusticus and Lovato, 2014). We henceforth refer to the Welch-based TOST procedure as the TOST procedure.

The TOST procedure decomposes the interval null hypothesis H_0 into two one-sided hypotheses. These hypotheses are $H_{0L} : \theta_1 - \theta_2 \leq \delta_L$ vs. $H_{1L} : \theta_1 - \theta_2 > \delta_L$ and $H_{0U} : \theta_1 - \theta_2 \geq \delta_U$ vs. $H_{1U} : \theta_1 - \theta_2 < \delta_U$. To conclude $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$, both H_{0L} and H_{0U} must be rejected at the nominal level of significance α. With Welch's t-tests, we therefore require that

$$t_L = \frac{(\bar{y}_1 - \bar{y}_2) - \delta_L}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \geq t_{1-\alpha}(\nu) \quad \text{and} \quad t_U = \frac{\delta_U - (\bar{y}_1 - \bar{y}_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \geq t_{1-\alpha}(\nu),$$

where s_j^2 is the sample variance for group $j = 1, 2$ and $t_{1-\alpha}(\nu)$ is the upper α-quantile of the t-distribution with ν degrees of freedom. The degrees of freedom for both t-tests are

$$\nu = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \times \left(\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}\right)^{-1}. \tag{2.1}$$

The sample variances s_1^2 and s_2^2 are unknown at the design stage. Because ν is a function of the sample variances, the null distribution of the test statistic is also unknown a priori and therefore is not based on exact pivotal quantities. The critical value $t_{1-\alpha}(\nu)$ for the test statistics t_L and t_U then depends on the to-be-observed data, which complicates an analytical power analysis that uses integration. Jan and Shieh (2017) considered analytical power analysis for the TOST procedure with unequal variances by expressing the test statistics in terms of simpler normal, chi-square, and beta random variables. However, the consistency of their power estimates depends on the numerical integration settings as demonstrated in Appendix A.2.

We instead use simulation to obtain consistent power estimates. To compute the test statistics t_L and t_U, we need only simulate three sample summary statistics: $\bar{y}_1 - \bar{y}_2$, s_1^2, and s_2^2. When the data are indeed generated from the anticipated $\mathcal{N}(\mu_j, \sigma_j^2)$ distributions, these independent sample summary statistics are sufficient and can be expressed in terms of known normal and chi-square distributions. Although each normal distribution is comprised of $d = 2$ parameters, we can use three-dimensional (3D) simulation since $\bar{y}_1$ and $\bar{y}_2$ need not be generated separately. We generate these summary statistics using 3D randomized Sobol' sequences of length m: $\boldsymbol{u}_r = (u_{1r}, u_{2r}, u_{3r}) \in [0, 1]^3$ for $r = 1, ..., m$. As detailed in Section 1.2.3, we can use fewer simulation repetitions m to obtain unbiased power estimates with Sobol' sequences in lieu of pseudorandom alternatives.

Algorithm 2.1 outlines our procedure for unbiased empirical power estimation at sample

sizes n_1 and n_2 using a Sobol' sequence of length m and significance level α for each t-test. For each of the m points from the 3D Sobol' sequence, we obtain values for the summary statistics $\bar{y}_1 - \bar{y}_2$, s_1^2, and s_2^2 using cumulative distribution function (CDF) inversion. We let $F(\cdot;\nu)$ and $\Phi^{-1}(\cdot)$ be the inverse CDFs of the $\chi^2_{(\nu)}$ and standard normal distributions, respectively. Given these summary statistics, we determine whether the sample for a given simulation repetition corresponds to the equivalence test's rejection region. The proportion of the m Sobol' sequence points for which this occurs estimates the power of the test. The test statistic for each t-test is comprised of two random components: (1) $\bar{d} = \bar{y}_1 - \bar{y}_2$ in the numerator and (2) $se = \sqrt{s_1^2/n_1 + s_2^2/n_2}$ in the denominator. The rejection region for the TOST procedure is a triangle in the $(\bar{d}, se)$-space with vertices $(\delta_L, 0)$, $(\delta_U, 0)$, and $(0.5(\delta_L + \delta_U), 0.5(\delta_U - \delta_L)/t_{1-\alpha}(\nu))$. The procedure in Algorithm 2.1 along with this rejection region is visualized in Figure 2.1.

Algorithm 2.1 Procedure to Compute Empirical Power

procedure EMPIRICALPOWER($\theta_1 - \theta_2, \sigma_1, \sigma_2, \delta_L, \delta_U, \alpha, n_1, n_2, m$)
  `reject` $\leftarrow$ NULL
  Generate a Sobol' sequence of length m: $\boldsymbol{u}_r = (u_{1r}, u_{2r}, u_{3r}) \in [0,1]^3$ for $r = 1, ..., m$.
  for r in 1:m **do**
    Let $s_{jr}^2 = (n_j - 1)^{-1}\sigma_j^2 F(u_{jr}; n_j - 1)$ for $j = 1, 2$
    Let $\bar{d}_r = (\theta_1 - \theta_2) + \Phi^{-1}(u_{3r})\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
    Use s_{1r}^2 and s_{2r}^2 to compute se_r and ν_r via (2.1)
    `reject`[w] $\leftarrow$ `ifelse(`$t_{1-\alpha}(\nu_r)se_r < \min\{\bar{d}_r - \delta_L, \delta_U - \bar{d}_r\}$`, 1, 0)`
  return `mean(reject)` as empirical power

More generally, sampling distributions for hypothesis tests can be mapped to the unit hypercube $[0,1]^d$, where d is the number of sufficient statistics required to compute the relevant test statistics. These mappings can also be implemented via CDF inversion with conditional univariate distributions when the sufficient statistics are not mutually independent. In Chapter 3, we will leverage maximum likelihood estimates when low-dimensional sufficient statistics do not exist or are difficult to generate. Those methods rely on large-sample results but could be applied in frequentist settings.

The simulation dimension d may be large if using these mappings to design sequential tests with many interim analyses or facilitate extensive multi-group comparisons. If $d \geq 32$, we recommend verifying the performance of quasi-Monte Carlo methods with additional simulation. High-dimensional low-discrepancy sequences may have poor low-dimensional projections, which can lead to a deterioration in performance depending on the effective

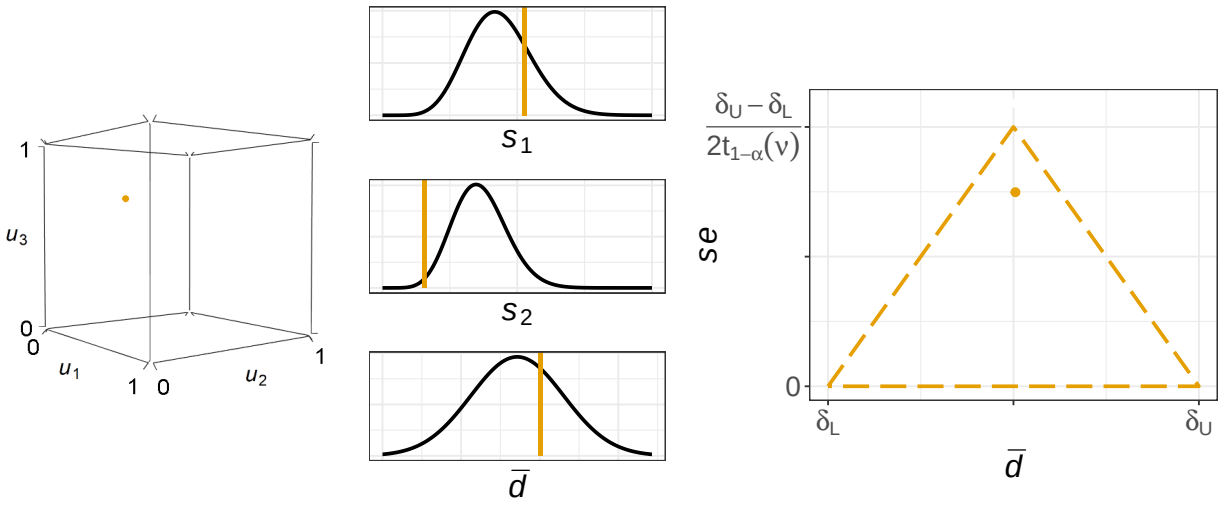


Figure 2.1: Left: Example point $(0.785, 0.009, 0.694) \in [0,1]^3$. Center: Mapping from this point to sufficient statistics. Right: The rejection region for the TOST procedure.

dimension of the simulation (Lemieux, 2009). Pseudorandom sequences could instead be used to implement such large-dimensional mappings.

For two-group equivalence tests, power analysis could be implemented by estimating power via Algorithm 2.1 at various sample sizes until the desired study power of $1 - \beta$ is achieved for some type II error rate β. However, that approach would be inefficient since we would need to thoroughly explore $[0,1]^3$ – and hence consider the entire sampling distribution – at each combination of sample sizes considered. Low-discrepancy sequences already allow us to obtain precise power estimates using fewer points from $[0,1]^3$ than pseudorandom sequences. We can further improve this efficiency by only exploring subspaces of $[0,1]^3$ that help us estimate power. We develop a methodology for this in Section 2.4. But first, we introduce an illustrative example that will be used to assess the performance of our method for power curve approximation proposed later. We illustrate the use of Algorithm 2.1 in this context.

2.3.2 Illustrative Example

This illustrative example is adapted from *PASS 2023* documentation (NCSS, LLC., 2023). *PASS* is a paid software solution that facilitates power analysis and sample size calculations for two-group equivalence tests with unequal variances. The illustrative example seeks to

	Estimated Power		
n	*PASS*	Algorithm 2.1	Naïve Simulation
3	0.1073	0.0414 (1.43×10^{-4})	0.0414 (7.85×10^{-4})
5	0.1778	0.1283 (1.70×10^{-4})	0.1282 (1.27×10^{-3})
8	0.4094	0.3801 (2.60×10^{-4})	0.3800 (2.03×10^{-3})
10	0.5527	0.5366 (2.68×10^{-4})	0.5368 (2.03×10^{-3})
15	0.7723	0.7699 (1.49×10^{-4})	0.7700 (1.88×10^{-3})
20	0.8810	0.8815 (1.65×10^{-4})	0.8816 (1.39×10^{-3})
30	0.9679	0.9687 (9.32×10^{-5})	0.9688 (6.66×10^{-4})
40	0.9924	0.9922 (5.28×10^{-5})	0.9922 (3.24×10^{-4})
50	0.9982	0.9982 (3.45×10^{-5})	0.9982 (1.67×10^{-4})
60	0.9996	0.9996 (2.10×10^{-5})	0.9996 (7.22×10^{-5})

Table 2.1: Power estimates presented in the *PASS* documentation along with the mean of 100 empirical power estimates obtained via Algorithm 2.1 and simulating normal data (Naïve Simulation). Standard deviations of the 100 empirical power estimates are given in parentheses.

compare the impact of two drugs on diastolic blood pressure, measured in mmHg (millimeters of mercury). The mean diastolic blood pressure is known to be roughly $\theta_2 = \mu_2 = 96$ mmHg with the reference drug ($j = 2$), and it is hypothesized to be about $\theta_1 = \mu_1 = 92$ mmHg with the test drug ($j = 1$). Subject matter experts use past studies to hypothesize within-group diastolic blood pressure standard deviations of $\sigma_1 = 18$ mmHg and $\sigma_2 = 15$ mmHg, respectively. The interval endpoints for the study are $\delta_U = 19.2$ and $\delta_L = -\delta_U$ to comply with guidance from the United States Food and Drug Administration (FDA, 2006). The significance level for the test is $\alpha = 0.05$.

The *PASS* documentation considers power for the illustrative example at $n = n_1 = n_2 = \{3, 5, 8, 10, 15, 20, 30, 40, 50, 60\}$. For each sample size n, we estimated power 100 times using Algorithm 2.1 with randomized Sobol' sequences of length $m = 2^{16} = 65536$. We also obtained 100 empirical power estimates for each sample size n by generating m samples of size n from the $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ distributions and recording the proportion of samples for which we concluded that $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$. Table 2.1 summarizes these numerical results and the power estimates presented in the *PASS* documentation.[1]

The two simulation-based approaches provide unbiased power estimates. However, Table 2.1 shows that the power estimates obtained via Algorithm 2.1 are much more precise

[1]In recognition of their licensing agreement, *PASS* software was not used nor accessed to confirm these power estimates.

than those obtained via naïve simulation with pseudorandom sequences. Moreover, each power estimate was obtained in roughly a quarter of a second when using Algorithm 2.1. It took between 20 and 30 seconds to obtain each power estimate using naïve simulation. This occurs because – regardless of the sample size n considered – Algorithm 2.1 reduces the power calculation to a three-dimensional problem that can be efficiently vectorized. We must use for loops to estimate power when directly generating the higher-dimensional data $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Moreover, the power estimates presented in the *PASS* documentation do not coincide with those returned via simulation for sample sizes less than 15, suggesting that it is valuable to develop more accurate methods for power analysis when the null distribution is not based on pivotal quantities.

2.4 Power Curve Approximation with Sampling Distribution Segments

2.4.1 An Efficient Approach to Power Analysis

In this section, we leverage the mapping between the unit cube and the test statistics presented in Section 2.3 to facilitate power curve approximation while exploring only segments of the sampling distribution. For given sample sizes n_1 and n_2, we previously mapped each Sobol' sequence point $\boldsymbol{u}_r$ $(r = 1, ..., m)$ to a mean difference, standard error, and degrees of freedom for its test statistic: $\bar{d}_r$, se_r, and ν_r. To compute empirical power in Algorithm 2.1, we fixed the sample sizes n_1 and n_2 and allowed the Sobol' sequence point to vary. We now specify a constant $q > 0$ such that $n = n_1 = qn_2$ to allow for imbalanced sample sizes. When approximating the power curve, we fix the Sobol' sequence point $\boldsymbol{u}_r$ and let the sample size n vary. We introduce the notation $\bar{d}_r^{(n,q)}$, $se_r^{(n,q)}$, and $\nu_r^{(n,q)}$ to make this clear. For fixed q and r, these quantities are only functions of the sample size n. As $n \to \infty$, $\bar{d}_r^{(n,q)}$, $se_r^{(n,q)}$, and $\nu_r^{(n,q)}$ approach $\theta_1 - \theta_2$, 0, and ∞, respectively.

We consider the behaviour of these functions when H_1 is true, i.e., when $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$. The upper vertex of the triangular rejection region for the TOST procedure is $\big(0.5(\delta_L + \delta_U), 0.5(\delta_U - \delta_L)/t_{1-\alpha}(\nu_r^{(n,q)})\big)$. First, $\nu_r^{(n,q)}$ almost always increases for fixed r and q as n increases. Thus as $n \to \infty$, the vertical coordinate of this rejection region vertex increases to $0.5(\delta_U - \delta_L)/\Phi^{-1}(1-\alpha)$, and the remaining two vertices do not change. The rejection region then defines a threshold for the standard error $se_r^{(n,q)}$: $\Lambda_r^{(n,q)} = \min\{\bar{d}_r^{(n,q)} - \delta_L, \delta_U - \bar{d}_r^{(n,q)}\}/t_{1-\alpha}(\nu_r^{(n,q)})$. We conclude $\theta_1 - \theta_2 \in (\delta_L, \delta_U)$ if and only if $se_r^{(n,q)}$ does not

exceed this threshold. For fixed r and q, this threshold is also a function of n:

$$\Lambda_r^{(n,q)} := \begin{cases} \dfrac{(\theta_1 - \theta_2) - \delta_L}{t_{1-\alpha}(\nu_r^{(n,q)})} + \dfrac{\Phi^{-1}(u_{3r})\sqrt{\sigma_1^2 + \sigma_2^2/q}}{\sqrt{n}\, t_{1-\alpha}(\nu_r^{(n,q)})} & \text{if } \delta_L < \bar{d}_r^{(n,q)} \leq 0.5(\delta_L + \delta_U) \\ \dfrac{\delta_U - (\theta_1 - \theta_2)}{t_{1-\alpha}(\nu_r^{(n,q)})} - \dfrac{\Phi^{-1}(u_{3r})\sqrt{\sigma_1^2 + \sigma_2^2/q}}{\sqrt{n}\, t_{1-\alpha}(\nu_r^{(n,q)})} & \text{if } 0.5(\delta_L + \delta_U) < \bar{d}_r^{(n,q)} < \delta_U \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

We suppose that a given point $\boldsymbol{u}_r$ yields $se_r^{(n,q)} \leq \Lambda_r^{(n,q)}$, which corresponds to the rejection region of the TOST procedure. In Section 2.4.2, we discuss why $se_r^{(n+1,q)} \leq \Lambda_r^{(n+1,q)}$ generally also holds true for the same point $\boldsymbol{u}_r$. In light of this, our method to approximate the power curve generates a single Sobol' sequence of length m. We use root-finding algorithms (Brent, 1973) to find the smallest value of n such that $se_r^{(n,q)} \leq \Lambda_r^{(n,q)}$ for each point $r = 1, ..., m$. We then use the empirical CDF of these m sample sizes to approximate the power curve as described in Algorithm 2.2.

Algorithm 2.2 Procedure for Power Curve Approximation

1: **procedure** PowerCurve($\theta_1 - \theta_2, \sigma_1, \sigma_2, \delta_L, \delta_U, \alpha, \beta, q, m$)
2: `sampSobol` $\leftarrow$ NULL
3: **for** r in 1:m **do**
4: Generate Sobol' sequence point $\boldsymbol{u}_r$
5: Let `sampSobol`$[r]$ solve $se_r^{(n,q)} - \Lambda_r^{(n,q)} = 0$ in terms of n
6: Let n_* be the $(1-\beta)$-quantile of `sampSobol`
7: **for** r in 1:m **do**
8: **if** `sampSobol`$[r] \leq n_*$ **then**
9: **if** $se_r^{(n_*,q)} > \Lambda_r^{(n_*,q)}$ **then**
10: Repeat Line 5, initializing the root-finding algorithm at n_*
11: **else**
12: **if** $se_r^{(n_*,q)} \leq \Lambda_r^{(n_*,q)}$ **then**
13: Repeat Line 5, initializing the root-finding algorithm at n_*
14: `powerCurve` $\leftarrow$ empirical CDF of `sampSobol`
15: Let n^* be the $(1-\beta)$-quantile of `sampSobol`
16: **return** `powerCurve`, $n_1 = \lceil n^* \rceil$ and $n_2 = \lceil qn^* \rceil$ as the recommended sample sizes

We now elaborate on several of the steps in Algorithm 2.2. Lines 2 to 6 describe a process that would yield an unbiased power curve and sample size recommendation if

$se_r^{(n,q)} = \Lambda_r^{(n,q)}$ were guaranteed to have a unique solution in terms of n for fixed r and q. However, $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ may infrequently intersect more than once. Given the reasoning in Section 2.4.2 and the numerical studies in Appendix A.1, these multiple intersections do not occur frequently enough to deter us from using root-finding algorithms to explore sample sizes. With root-finding algorithms, we explore only subspaces of $[0,1]^3$ for each sample size investigated since different values of n are considered for each point $\boldsymbol{u}_r$ in Line 4. Root-finding algorithms therefore give rise to computational efficiency as the entire sampling distribution is not estimated when exploring sample sizes n. In particular, the root-finding algorithm computes test statistics corresponding to $\mathcal{O}(\log_2 B)$ points from $[0,1]^3$, where B is the maximum sample size considered for the power curve. We would require $\mathcal{O}(B)$ such points to explore a similar range of sample sizes using power estimates from Algorithm 2.1. When $B \geq 59$, this approach reduces the number of test statistics we must estimate by at least an order of magnitude because $\mathcal{O}(\log_2 B) < \mathcal{O}(B)/10$. Using low-discrepancy sequences instead of pseudorandom ones further reduces the number of test statistics we must estimate as demonstrated in Section 2.4.3.

If we skipped Lines 7 to 14 of Algorithm 2.2, the unbiasedness of the sample size recommendation in Line 16 is not guaranteed due to the potential for multiple intersections between $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$. To ensure our sample size recommendations are unbiased despite using subspaces of $[0,1]^3$ to consider sample sizes, we estimate the entire sampling distribution of test statistics at the sample size $n = n_*$ in Lines 7 to 13. If the statements in Lines 9 or 12 are true, this implies that $se_r^{(n,q)} = \Lambda_r^{(n,q)}$ for at least two distinct sample sizes n. For these points $\boldsymbol{u}_r$, we can reinitialize the root-finding algorithm at n_* to obtain a solution for each point that will make the power curve unbiased at n_*. The resulting sample size recommendation under repeated implementation of Algorithm 2.2 is therefore also unbiased.

Our numerical studies in Section 2.4.3 show that the if statements in Lines 9 and 12 are very rarely true for any point $\boldsymbol{u}_r \in [0,1]^3$. In those situations, $n_* = n^*$ and both the power estimate at n^* and the sample size recommendations $\lceil n^* \rceil$ and $\lceil qn^* \rceil$ are unbiased. It is incredibly unlikely that n_* and n^* would differ substantially, but Lines 7 to 13 of Algorithm 2.2 could be repeated in that event, where the root-finding algorithm is initialized at n^* instead of n_*. Even when $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersect more than once, the power curves from Algorithm 2.2 are unbiased near the target power $1 - \beta$, but their global unbiasedness at all sample sizes n is not strictly guaranteed. Nevertheless, our numerical studies in Section 2.4.3 highlight good global estimation of the power curve.

The methods we leveraged to select subspaces of $[0,1]^3$ for two-group equivalence tests are tailored to the functions $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$. However, these methods rely more generally on the weak law of large numbers since most sufficient statistics are based on sample means.

Upon mapping the unit hypercube $[0,1]^d$ to sufficient statistics, the behaviour of the test statistics as a function of the sample size n can generally be studied to develop analogues to Algorithm 2.2 for other tests and designs. Root-finding algorithms are generally useful when the rejection region is convex. Rejection regions for the TOST procedure in Figure 2.1, other equivalence tests, and one-sided hypothesis tests are typically convex, whereas hypothesis tests with point null hypotheses often have non-convex rejection regions.

2.4.2 Justification for Using Root-Finding Algorithms

Here, we discuss why using root-finding algorithms to approximate the power curve yields suitable results – even though $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ can (although infrequently) intersect more than once. The threshold $\Lambda_r^{(n,q)}$ approaches $\Lambda^* = \min\{(\theta_1 - \theta_2) - \delta_L, \delta_U - (\theta_1 - \theta_2)\}/\Phi^{-1}(1-\alpha) > 0$ as n increases. The standard error $se_r^{(n,q)}$ generally decreases as $n \to \infty$, but it is not necessarily a strictly decreasing function of n. We first consider the case where $se_r^{(n,q)}$ does strictly decrease as n increases. For small sample sizes, $\Lambda_r^{(n,q)}$ is typically an increasing function of n due to the decrease in $t_{1-\alpha}(\nu_r^{(n,q)})$. Every other Sobol' sequence point $\boldsymbol{u}_r$ is such that $\text{sign}(u_{3r} - 0.5) = \text{sign}((\theta_1 - \theta_2) - 0.5(\delta_L + \delta_U))$. In this case, $\Lambda_r^{(n,q)}$ is also an increasing function of n for large sample sizes. This occurs because $\bar{d}_r^{(n,q)}$ is never closer than $\theta_1 - \theta_2$ to the horizontal center of the rejection region at $\bar{d} = 0.5(\delta_L + \delta_U)$. Therefore, the increasing $\Lambda_r^{(n,q)}$ and decreasing $se_r^{(n,q)}$ typically intersect once. If $\text{sign}(u_{3r} - 0.5) \neq \text{sign}((\theta_1 - \theta_2) - 0.5(\delta_L + \delta_U))$, then $\Lambda_r^{(n,q)}$ is a decreasing function of n for large sample sizes. This occurs because $\bar{d}_r^{(n,q)}$ approaches $\theta_1 - \theta_2$ from the horizontal center of the rejection region. However, $\Lambda_r^{(n,q)}$ decreases to a nonzero constant Λ^*, while $se_r^{(n,q)}$ decreases to 0 as $n \to \infty$. Again, the functions $\Lambda_r^{(n,q)}$ and $se_r^{(n,q)}$ typically intersect only once.

We next consider the case where $se_r^{(n,q)}$ is not a strictly decreasing function of n. Line 5 of Algorithm 2.1 prompts the first line of (2.3):

$$
\begin{aligned}
se_r^{(n,q)} &= \frac{1}{\sqrt{n}} \left[\frac{\sigma_1^2}{n-1} F(u_{1r}; n-1) + \frac{\sigma_2^2}{q(qn-1)} F(u_{2r}; qn-1) \right]^{1/2} \qquad (2.3) \\
&\approx \frac{1}{\sqrt{2n}} \left[\frac{\sigma_1^2}{n-1} \left(\Phi^{-1}(u_{1r}) + \sqrt{2(n-1)} \right)^2 + \frac{\sigma_2^2}{q(qn-1)} \left(\Phi^{-1}(u_{2r}) + \sqrt{2(qn-1)} \right)^2 \right]^{1/2}.
\end{aligned}
$$

Because quantiles from the chi-squared distribution do not have closed forms, the second line of (2.3) leverages the approximation from Fisher (1934) for illustrative purposes. When $\Phi^{-1}(u_{1r}) \in (-\sqrt{2(n-1)} \pm 1)$ or $\Phi^{-1}(u_{2r}) \in (-\sqrt{2(qn-1)} \pm 1)$, the square function respectively makes the $(\Phi^{-1}(u_{1r}) + \sqrt{2(n-1)})^2$ or $(\Phi^{-1}(u_{2r}) + \sqrt{2(qn-1)})^2$ term in (2.3)

smaller. As n increases in those situations, the relative increase in the squared terms may offset the decreasing impact of the terms in the denominators of (2.3). However, this increasing trend cannot persist as $se_r^{(n,q)}$ is $O\big(n^{-1/2}\big)$ and $Pr(\Phi^{-1}(u_{1r}) \in (-\sqrt{2(n-1)} \pm 1))$ and $Pr(\Phi^{-1}(u_{2r}) \in (-\sqrt{2(qn-1)} \pm 1))$ both approach 0 as $n \to \infty$. We show via simulation in Appendix A.1 that this increasing trend is rare for $n > 5$. For $n \leq 5$, $\Lambda_r^{(n,q)}$ is generally also an increasing function of n as mentioned in Section 2.4.1.

If $\Lambda_r^{(n,q)}$ is a decreasing function of n, it follows from (2.2) that $\bar{d}_r^{(n,q)} = 0.5(\delta_L + \delta_U)$ when

$$n = \frac{(\Phi^{-1}(u_{3r}))^2(\sigma_1^2 + \sigma_2^2/q)}{(0.5(\delta_L + \delta_U) - (\theta_1 - \theta_2))^2}. \tag{2.4}$$

The threshold $\Lambda_r^{(n,q)}$ should therefore not be decreasing for sample sizes smaller than n given in (2.4). By (2.3), $se_r^{(n,q)}$ approximates $n^{-\frac{1}{2}}\sqrt{\sigma_1^2 + \sigma_2^2/q}$ for large sample sizes n. It follows by (2.2) that

$$\Lambda_r^{(n,q)} \approx \Lambda^* + \frac{|\Phi^{-1}(u_{3r})|}{\Phi^{-1}(1-\alpha)} se_r^{(n,q)}, \tag{2.5}$$

when $\text{sign}(u_{3r} - 0.5) \neq \text{sign}((\theta_1 - \theta_2) - 0.5(\delta_L + \delta_U))$ for large n. We note that $\Lambda_r^{(n,q)}$ and $se_r^{(n,q)}$ may intersect for a value of n that is smaller than the one given in (2.4). If $se_r^{(n,q)}$ is instead larger than $\Lambda_r^{(n,q)}$ over the entire range of n values for which $\Lambda_r^{(n,q)}$ increases, then (2.5) suggests that $\Lambda_r^{(n,q)}$ and $se_r^{(n,q)}$ are likely to intersect only once when $\Lambda_r^{(n,q)}$ is decreasing. The functions $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ therefore typically have one intersection for all cases discussed here, but we illustrate an occurrence of multiple intersections below.

To do so, we reconsider the illustrative example from Section 2.3.2 with $q = 1$. In Figure 2.2, we show that $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersect more than once for the Sobol' sequence point $\boldsymbol{u}_r = (u_{1r}, u_{2r}, u_{3r}) = (0.184, 0.231, 0.449)$. We note that both $\Phi^{-1}(0.184) \in (-\sqrt{2(n-1)} \pm 1)$ and $\Phi^{-1}(0.231) \in (-\sqrt{2(n-1)} \pm 1)$ when $n = 2$; $se_r^{(n,q)}$ is therefore small for $n = 2$ and increases until $n = 4$ before decreasing to 0. This trend is evident in the right plot of Figure 2.2, which displays the functions for sample sizes n between 2 and 10. This plot shows that $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersect twice: once between $n = 2$ and 3 and again between $n = 3$ and 4. This means that for this point $\boldsymbol{u}_r$, $\bar{d}_w^{(n,q)}$, $se_r^{(n,q)}$, and $\nu_r^{(n,q)}$ correspond to the rejection region of the TOST procedure for $n = 2$ and $n \geq 4$, but not for $n = 3$. The scenario visualized in Figure 2.2 arose from using a Sobol' sequence of length $m = 1024$. Of these 1024 Sobol' sequence points, there was only one other point where $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersected more than once. The intersections for this other point were also between $n = 2$ and 3 and between $n = 3$ and 4.

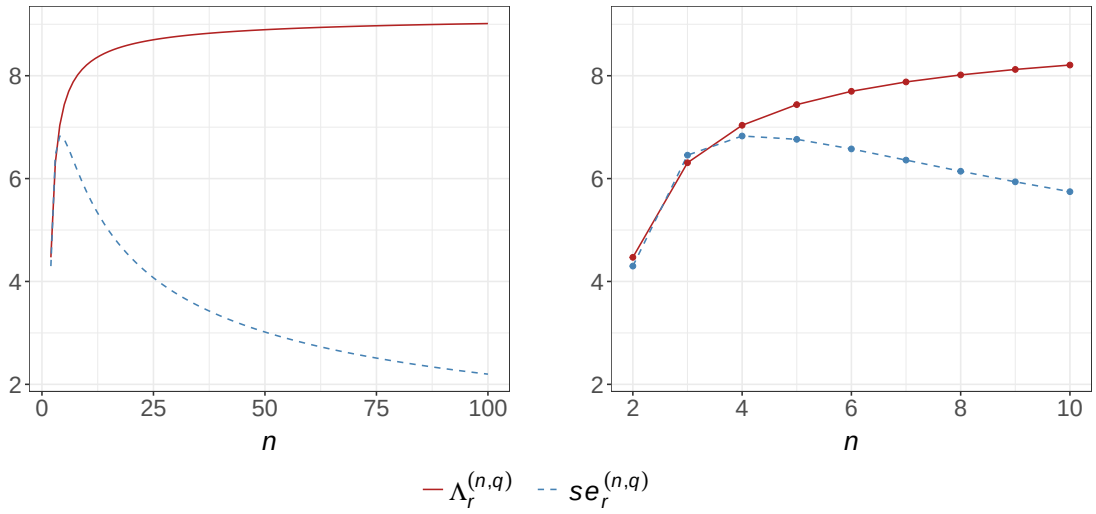


Figure 2.2: Visualization of $\Lambda_r^{(n,q)}$ and $se_r^{(n,q)}$ as functions of n for the illustrative example with $\boldsymbol{u}_r = (0.184, 0.231, 0.449)$ and $q = 1$. Left: Sample sizes 2 to 100. Right: Sample sizes 2 to 10.

2.4.3 Numerical Study with the Illustrative Example

We now consider the illustrative example from Section 2.3.2 to illustrate the reliable performance of our efficient method for power curve approximation. For this example, we approximated the power curve 1000 times with $q = 1$ (i.e., $n = n_1 = n_2$). Each of the 1000 power curves were approximated using Algorithm 2.2 with a target power of $1 - \beta = 0.8$ and a Sobol' sequence of length $m = 2^{10} = 1024$. We recommend using shorter Sobol' sequences when approximating the power curve than when computing empirical power for a specific (n_1, n_2) combination ($m = 65536$ was used in Section 2.3.2). Whereas all computations in Algorithm 2.1 can be vectorized, we must use a for loop to implement the root-finding algorithm for each Sobol' sequence point.

Although not incorporated into Algorithm 2.2, the number of points in $\{\boldsymbol{u}_r\}_{r=1}^m$ could be chosen to satisfy a criterion for the precision of a confidence interval for power or n_*, the sample size recommendation prior to rounding or reimplementing the root-finding algorithm. In that case, we would require M i.i.d. copies of shorter Sobol' sequences with length M_0 such that $M \times M_0 = m$. Each i.i.d. copy of the Sobol' sequence would prompt an independently obtained estimate for power or n_*. These estimates could be used to construct a confidence interval for the relevant unknown quantity (Lemieux, 2009). Since Sobol' sequences can be augmented with additional points (Sobol', 1967), we could

incrementally increase M_0 for each shorter Sobol' sequence until the confidence interval for power or n_* is precise enough. The approach proposed in this paragraph would be roughly as computationally efficient as Algorithm 2.2 if the Sobol' sequences were augmented prior to obtaining the confirmatory power estimates in Lines 7 to 13 of Algorithm 2.2. However, using one longer Sobol' sequence of length m typically yields greater variance reduction than using m points from M i.i.d. shorter Sobol' sequences.

We compare the 1000 generated power curves to the unbiased power estimates from Algorithm 2.1 in Table 2.1. The left plot of Figure 2.3 demonstrates that Algorithm 2.2 yields suitable global power curve approximation when comparing its results to these power estimates. Each power curve was approximated without estimating the entire sampling distribution for all sample sizes n_1 and n_2 explored as emphasized in Section 2.4.4. To further investigate the performance of Algorithm 2.2, we repeated the process from the previous paragraph to estimate 1000 power curves for the illustrative example with $1 - \beta = \{0.2, 0.3, \ldots, 0.7, 0.9\}$. In total, we approximated 8000 power curves for this example. Using the root-finding algorithm to explore the sample size space did not lead to performance issues. We did not need to reinitialize the root-finding algorithm in Lines 7 to 13 of Algorithm 2.2 for *any* of the 8.192×10^6 points used to generate these 8000 curves.

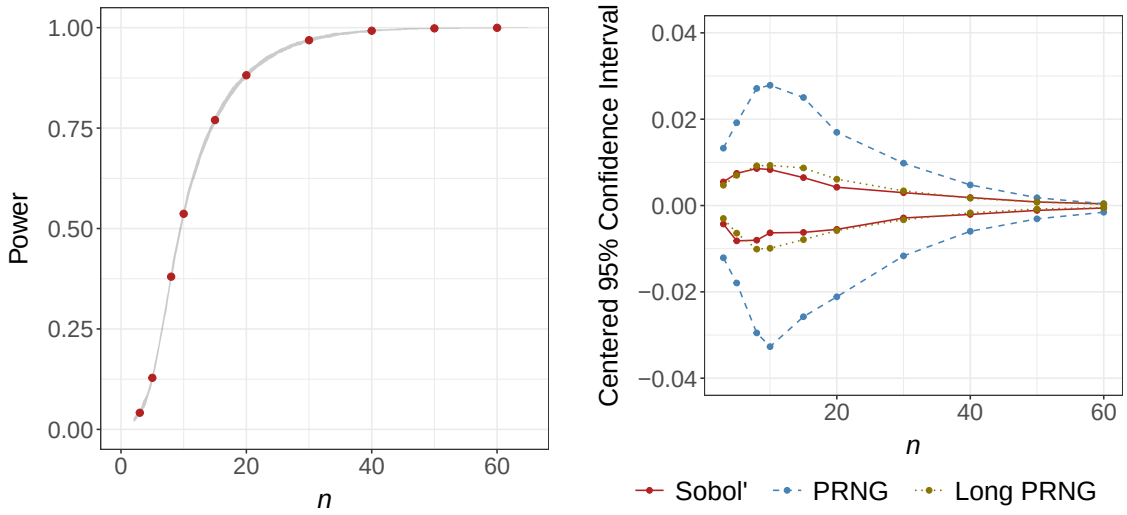


Figure 2.3: Left: 1000 power curves estimated for the illustrative example (grey) and the power estimates obtained via Algorithm 2.1 (red). Right: Endpoints of the centered 95% confidence intervals for power obtained with Sobol' ($m = 1024$) and pseudorandom (PRNG) sequences ($m = 1024, 10^4$).

The suitable performance of Algorithm 2.2 is corroborated by more extensive numerical studies detailed in Appendix A.1.

To assess the impact of using Sobol' sequences with Algorithm 2.2, we approximated 1000 power curves for the illustrative example using root-finding algorithms with $1-\beta = 0.8$ and sequences from a pseudorandom number generator. We then used the 1000 power curves corresponding to each sequence type (Sobol' and pseudorandom) to estimate power for the sample sizes considered in Section 2.3.2: $n = \{3, 5, 8, 10, 15, 20, 30, 40, 50, 60\}$. For each sample size and sequence type, we obtained a 95% confidence interval for power using the percentile bootstrap method (Efron, 1982). We then created centered confidence intervals by subtracting the power estimates produced by Algorithm 2.1 from each confidence interval endpoint. The right plot of Figure 2.3 depicts these results for the 10 sample sizes n and two sequence types considered. Figure 2.3 illustrates that the Sobol' sequence gives rise to much more precise power estimates than pseudorandom sequences – particularly when power is not near 0 or 1. We repeated this process to generate 1000 power curves via Algorithm 2.2 with pseudorandom sequences of length $m = 10^4$. The power estimates obtained using Sobol' sequences with length $m = 1024$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 10^4$. Using Sobol' sequences therefore allows us to estimate power with the same precision using approximately an order of magnitude fewer points from $[0, 1]^3$. Each power curve for this example with $m = 1024$ took just under one second to approximate. It would take roughly 10 times as long to approximate the power curve with the same precision using pseudorandom points in lieu of Sobol' sequences.

2.4.4 Exploring Subspaces of the Unit Cube

We next demonstrate how segments of the sampling distribution are considered by exploring only subspaces of the unit cube $[0, 1]^3$ for most sample sizes considered. The left plot of Figure 2.4 decomposes the results of the root-finding algorithm for one approximated power curve from Section 2.4.3 with $1 - \beta = 0.8$ for the illustrative example. Even when the root-finding algorithm is initialized at the same sample size for all $\{\boldsymbol{u}_r\}_{r=1}^m$, different n are considered for each point $\boldsymbol{u}_r$ when determining the solution to $se_r^{(n,q)} = \Lambda_r^{(n,q)}$. The value of n is noninteger in most iterations of the root-finding algorithm, and the colours in the left plot of Figure 2.4 indicate which points from the unit cube were considered for various ranges of n. For instance, the purple points were such that their test statistics corresponded to the rejection region for the smallest possible sample size of $n = 2$. Moreover, only the blue points in $[0, 1]^3$ were used to estimate test statistics for at least one sample size $n \in (2, 8)$ when exploring values of n via the root-finding procedure.

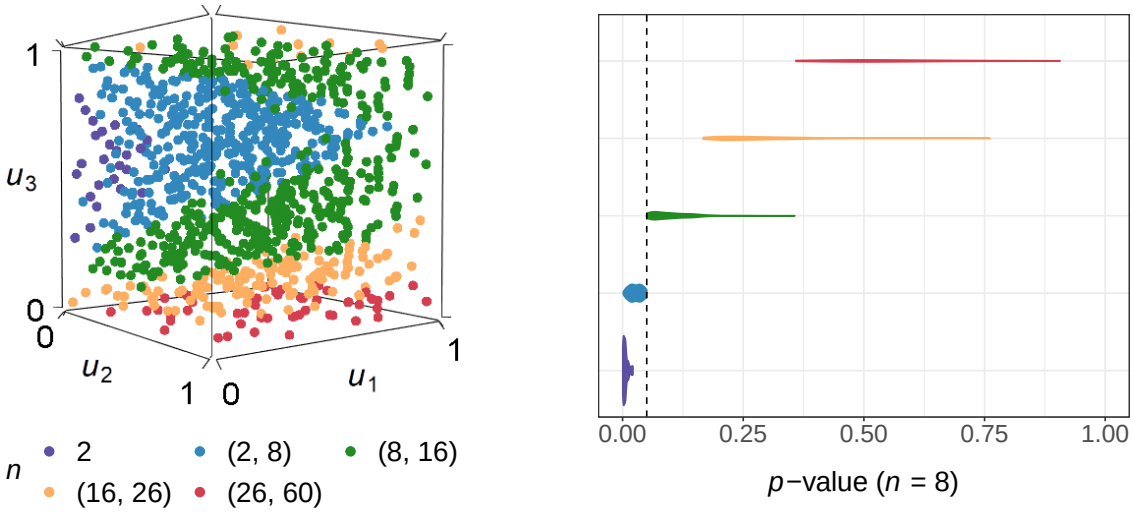


Figure 2.4: Left: Visualization of which points in $[0,1]^3$ were used to explore at least one n value in the various sample size ranges via the root-finding algorithm. Right: Violin plots for segments of the sampling distribution of p-values when $n = 8$. The dotted vertical line is at $\alpha = 0.05$.

The points that were used to explore the smallest sample sizes generally have moderate u_{3r} values and smaller u_{1r} and u_{2r} values. The mean difference $\bar{d}_r^{(n,q)}$ is therefore small in absolute value and the sample variances for groups 1 and 2 are small, which implies that the numerators of the test statistics t_L and t_U are large and their denominators are small. The points used to explore larger sample sizes generally have more extreme u_{3r} values, so $\bar{d}_r^{(n,q)}$ may not substantially differ from one of δ_L or δ_U for small sample sizes. While the pattern in the left plot of Figure 2.4 depends on the inputs for Algorithm 2.2, the root-finding algorithm correctly identifies and prioritizes subspaces of $[0,1]^3$ such that $se_r^{(n,q)} \approx \Lambda_r^{(n,q)}$ for a given sample size n with an arbitrary design. Our methods can be extended to more complex designs, but it is difficult to visualize the prioritized subspaces of the unit hypercube when the simulation dimension d is greater than 3.

The right plot of Figure 2.4 visualizes segments from the sampling distribution of p-values for $n = n_1 = n_2 = 8$ conditional on the categorizations from the left plot. For the TOST procedure, the p-value is the maximum of the p-values corresponding to t_L and t_U. This p-value does not exceed the significance level α if and only if $se_r^{(n,q)} \leq \Lambda_r^{(n,q)}$. This plot demonstrates why it is wasteful to use the purple points to consider $n \approx 8$ because those points satisfy $se_r^{(n,q)} \leq \Lambda_r^{(n,q)}$ for $n = 2$. It follows from Section 2.4.2 that $se_r^{(n,q)} \leq \Lambda_r^{(n,q)}$

generally holds true for $n \approx 8$ with those points, and the corresponding p-values are hence smaller than $\alpha = 0.05$. By similar logic, it is wasteful to consider the red points for $n \approx 8$ since the p-values for those points will be much larger than α. Although these coloured categorizations are not used in Algorithm 2.2, they illustrate the targeted nature of how we consider sample sizes n with segments of the relevant sampling distributions.

2.5 Efficient Power Analysis for Crossover Designs

The method for power curve approximation in Algorithm 2.2 was tailored to a standard parallel study with unequal variances. However, the underlying ideas generalize to other study designs. In particular, we overview here how to extend Algorithm 2.2 for use with crossover designs. Power analysis for crossover designs is of particular interest because regulatory agencies often recommend using them to conclude average bioequivalence (FDA, 2006). In crossover designs, each subject receives a different clinical treatment during different study periods (Chow and Liu, 2008). This is advantageous in that inter-subject variability is removed from between-treatment comparisons. We describe how to use our design methods based on sampling distribution segments with two-sequence, two-period (2×2) crossover designs in Appendix A.3.

For the 2×2 crossover design, we also highlight discrepancies (of up to 33%) between the sample sizes recommended by the power curves from Algorithm 2.2 and those endorsed in popular textbooks on bioequivalence study design (Chow and Liu, 2008). These discrepancies further motivate the need for our design methods. The implementation of such extensions for these and other crossover designs is supported in the `dent` package developed in conjunction with this chapter. Furthermore, our method for power analysis is flexible and could readily accommodate additional designs not discussed in this chapter.

2.6 Discussion

In this chapter, we developed a framework for power analysis when null distributions cannot be expressed in terms of exact pivotal quantities. This framework maps the unit hypercube $[0, 1]^d$ to sufficient statistics and leverages this mapping to estimate power curves using segments of sampling distributions. Using segments of sampling distributions improves the scalability of our simulation-based design procedures without compromising the unbiasedness of the sample size recommendations. Our framework is illustrated with three-dimensional simulation for two-group equivalence tests with unequal variances, but

we described how to apply our methods more generally throughout the chapter and now elaborate on several additional extensions.

Future work could apply the framework proposed in this chapter to compare more than two groups. In that case, the simulation dimension d would need to be increased, and the multiple comparisons problem would need to be considered. We could also apply this framework to efficiently design sequential analyses that allow for early termination of the study. In sequential settings, we would likely need to define analogues to $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ for each interim analysis and synthesize the results for each point $\boldsymbol{u}_r$. However, it is not trivial to create a mapping between points in $[0,1]^d$ and sufficient statistics that maintain the desired level of dependence between interim analyses for arbitrary sample sizes. We elaborate on this extension in Section 6.2.1.

Finally, we could explore how this framework might be applied to quickly and reliably recommend sample sizes for nonparametric testing methods. The exact null distributions for those tests are not based on pivotal quantities, and it is not possible to generate sufficient statistics in nonparametric settings. Sample size determination for these studies typically utilizes naïve simulation. In nonparametric settings, we may be able to map the unit hypercube $[0,1]^d$ to insufficient statistics, such as sample totals, and use low-discrepancy sequences to improve the scalability and precision of empirical power analysis. This extension is discussed in Section 6.2.2.

Chapter 3

Power Curves for Posterior Analyses

3.1 Preamble

This chapter extends design with sampling distribution segments to Bayesian settings. The approaches used to analyze data via interval hypothesis tests in this chapter are fully Bayesian. However, the design framework proposed in Chapter 3 is predominantly based on choices that are commonplace in frequentist design. For instance, the methods in this chapter assume that data are generated from statistical models with fixed parameter values for the purposes of sample size determination. Practitioners who design frequentist studies should already be accustomed to choosing these fixed parameter values. The methods from this chapter therefore provide an accessible point of entry for users who are designing their first Bayesian studies.

Since we impose restrictions on the data generation process in this chapter, we emphasize the broad applicability of our methods with posterior analyses facilitated via posterior probabilities, Bayes factors, and credible intervals. As discussed in this chapter, the sample size recommendations prompted by these methods are consistent – not unbiased. These recommendations are consistent in that they are suitable when normal approximations to the relevant posteriors and sampling distributions of the maximum likelihood estimator are appropriate. This lack of unbiasedness is a consequence of considering more complex design settings, where it is more difficult or even impossible to generate low-dimensional sufficient statistics.

The methods proposed in this chapter will be extended later to accommodate more flexible data generation processes using content from Chapters 4 and 5. Moreover, the

approaches in Chapter 3 do not formally consider Bayesian analogues to type I error. Those analogues will also be incorporated into Chapter 5. This chapter of the thesis is also the one that emphasizes the intricacies of ratio-based comparisons in most detail.

3.2 Background

Power-based approaches to Bayesian sample size determination aim for pre-experimental probabilistic control over testing procedures for a characteristic of interest θ. For two-group comparisons, $\theta = h(\theta_1, \theta_2)$ for some function $h(\cdot)$. This chapter considers $h(\theta_1, \theta_2) = \theta_1 - \theta_2$ and $h(\theta_1, \theta_2) = \theta_1/\theta_2$. The case where $\theta = \log(\theta_1) - \log(\theta_2)$ can be viewed as a generalization of $h(\theta_1, \theta_2) = \theta_1 - \theta_2$. In pre-experimental settings, the data have not been observed and are random variables. Data from a random sample are represented by $\boldsymbol{Y}^{(n)}$, consisting of observations $\{y_{i1}\}_{i=1}^n$ from group 1 and observations $\{y_{i2}\}_{i=1}^n$ from group 2. Imbalanced sample size determination in Bayesian settings will be considered in Chapter 5.

A *design* prior $p_D(\boldsymbol{\eta})$ (De Santis, 2007; Berry et al., 2011; Gubbiotti and De Santis, 2011) models uncertainty regarding the model parameters $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ from each group in pre-experimental settings. Following the convention from Chapter 1, the characteristic of interest θ_j for group j is specified as a function $g(\cdot)$ of the model parameters: $\theta_j = g(\boldsymbol{\eta}_j)$ for $j = 1, 2$. Since the (informative) design prior is concentrated on θ values that are relevant to the objective of the study, it is usually different from the *analysis* prior used to analyze the observed data. The analysis prior is specified for the model parameters $\boldsymbol{\eta}_j$ as detailed in Section 1.1.4. The design prior gives rise to the prior predictive distribution of $\boldsymbol{Y}^{(n)}$:

$$p(\boldsymbol{y}^{(n)}) = \int \prod_{i=1}^{n} f(y_{i1}; \boldsymbol{\eta}_1) \prod_{i=1}^{n} f(y_{i2}; \boldsymbol{\eta}_2) p_D(\boldsymbol{\eta})\, d\boldsymbol{\eta},$$

where $f(y; \boldsymbol{\eta}_j)$ is the model for group $j = 1, 2$. The relevant power criteria defined for the Bayesian inferential methods are considered when the data are generated from this prior predictive distribution.

Gubbiotti and De Santis (2011) defined two methodologies for choosing the sampling distribution of $\boldsymbol{Y}^{(n)}$: the conditional and predictive approaches. The conditional approach fixes *design values* $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$ for the model parameters. The power criteria are then based on the probability density or mass functions $f(y; \boldsymbol{\eta}_{1,0})$ and $f(y; \boldsymbol{\eta}_{2,0})$. The conditional approach is typically used in frequentist sample size calculations. The predictive approach uses a (nondegenerate) design prior $p_D(\boldsymbol{\eta})$, which is arguably more consistent with the

Bayesian framework. However, this chapter discusses advantages to using the conditional approach that may outweigh its weaknesses for certain practitioners.

Bayesian methods for power analysis have been proposed in a variety of contexts (Berry et al., 2011; Gubbiotti and De Santis, 2011; Brutti et al., 2014). We describe here how to control power when analyses are carried out via the methods described in Section 1.1.4. In these contexts, one aims to select a sample size n to ensure the probability of correctly concluding that $H_1 : \theta \in (\delta_L, \delta_U)$ is true is at least $1-\beta$ for some target power $1-\beta \in (0,1)$. For hypothesis tests with posterior probabilities, the selected sample size ensures that

$$\mathbb{E}\left[\mathbb{I}\{Pr(H_1 \,|\, \boldsymbol{Y}^{(n)}) \geq \gamma\}\right] \geq 1-\beta, \tag{3.1}$$

for some critical value $\gamma \in [0.5, 1)$. When $\boldsymbol{Y}^{(n)} \sim p(\boldsymbol{y}^{(n)})$, $p_D(H_1)$ provides an upper bound for the attainable target power. Power analyses for hypothesis tests with Bayes factors are related. It follows from (1.6) that the NOH Bayes factor exceeds K if and only if

$$Pr(H_1 \mid \boldsymbol{y}^{(n)}) > \frac{K \times Pr(H_1)}{1-(K-1)\times Pr(H_1)}. \tag{3.2}$$

The power criterion for NOH Bayes factors with threshold $K \geq 1$ is therefore a special case of (3.1) when the critical value γ equals the right side of (3.2).

For hypothesis tests with credible intervals, the quantity in (3.1) is replaced with

$$\mathbb{E}\left[\mathbb{I}\{L_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)}) > \delta_L \quad \cap \quad U_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)}) < \delta_U\}\right] \geq 1-\beta, \tag{3.3}$$

where the endpoints of the credible interval are henceforth denoted $L_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)})$ and $U_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)})$ to emphasize their dependence on the data. If the posterior credible interval is equal tailed, the power criterion in (3.3) simplifies to

$$\mathbb{E}\left[\mathbb{I}\{Pr(\theta < \delta_L | \boldsymbol{Y}^{(n)}) < \alpha/2 \quad \cap \quad Pr(\theta > \delta_U | \boldsymbol{Y}^{(n)}) < \alpha/2\}\right] \geq 1-\beta. \tag{3.4}$$

When $1-\alpha = \gamma$, (3.3) and (3.4) impose stricter criteria than (3.1). At least $100\times\gamma\%$ of the posterior for θ must lie within the interval (δ_L, δ_U) for $(L_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)}), U_{\theta,1-\alpha}(\boldsymbol{Y}^{(n)}))$ to also be contained in this interval. The plot of the quantity in (3.1), (3.3), or (3.4) as a function of the sample size n is called the power curve.

Minimum sample sizes that satisfy power criteria can be found analytically in certain situations where conjugate priors are used (see e.g., Spiegelhalter et al. (1994); Gubbiotti and De Santis (2011)). However, to support more flexible study design, sample sizes that satisfy power criteria can be found using simulation. Most simulation-based procedures

for power analysis with design priors follow a similar process (Wang and Gelfand, 2002). First, a sample size n is selected. Second, a value $\boldsymbol{\eta}_*$ is drawn from the design prior $p_D(\boldsymbol{\eta})$. Third, data $\boldsymbol{y}_*^{(n)}$ are generated according to the model $f(y; \boldsymbol{\eta}_*)$. Fourth, the posterior of θ given $\boldsymbol{y}_*^{(n)}$ is approximated to compute $Pr(H_1 | \boldsymbol{y}_*^{(n)})$. This process is repeated many times to estimate a sampling distribution of posterior probabilities, which is used to determine whether the power criterion is satisfied with probability $1 - \beta$ for the selected sample size n.

These simulation-based approaches can be very computationally intensive as many posteriors must be approximated to estimate the sampling distribution for each sample size n considered. Wang and Gelfand (2002) recommended using bisection methods or grid searches to streamline the exploration of sample sizes. Yet even when such methods circumvent the need for practitioners to choose which sample sizes n to explore, time is still wasted considering sample sizes that are excessively large or much too small to satisfy the power criterion. This computational inefficiency is compounded over all combinations of the design inputs that practitioners wish to consider when designing Bayesian hypothesis tests – including the critical value γ, interval (δ_L, δ_U), target power $1 - \beta$, and design and analysis priors. A fast framework for power curve approximation with posterior analyses based on sampling distribution segments would mitigate this issue and expedite study design.

The remainder of this chapter is structured as follows. We describe a food expenditure example involving the comparison of gamma tail probabilities in Section 3.3. This example is referenced throughout the chapter to motivate the proposed methods. In Section 3.4, we propose a method to map the sampling distribution of posterior probabilities to low-dimensional hypercubes under the conditional approach. We also prove that the resulting approximation to the sampling distribution gives rise to consistent power estimates under certain conditions. In Section 3.5, we exploit this mapping to quickly approximate power curves with low-discrepancy sequences. This approach is fast because for a given sample size, we consider only a segment of the approximate sampling distribution of posterior probabilities. Even without estimating entire sampling distributions, this method prompts consistent sample size recommendations. In Section 3.6, we conduct numerical studies to explore the performance of our power curve approximation method in several settings. We conclude with a discussion of extensions to this work in Section 3.7.

3.3 Motivating Example with the Gamma Model

Mexico's National Institute of Statistics and Geography conducts a biennial survey to monitor household income and expenses along with sociodemographic characteristics. We refer to this survey by its Spanish acronym ENIGH. In the ENIGH 2020 survey (INEGI, 2021), each surveyed household was assigned a socioeconomic class: lower, lower-middle, upper-middle, and upper. We use data from the lower-middle income households (the most populous class) in the Mexican state of Aguascalientes. We split the households into two groups based on the sex of the household's main provider. Each household has a weighting factor used to include its observation between one and four times in our data set. The datum y_{ij} collected for each household $i = 1, ..., n_j,\ j = 1, 2$ is its quarterly expenditure on food per person measured in thousands of Mexican pesos (MXN \$1000). We exclude the 0.41% of households that report no quarterly food expenditure to accommodate the gamma model's positive support. This respectively yields $n_1 = 759$ and $n_2 = 1959$ observations in the female ($j = 1$) and male ($j = 2$) provider groups that are visualized in Figure 3.1. We collectively refer to these data as $\boldsymbol{y}_1 = \{y_{i1}\}_{i=1}^{n_1}$ and $\boldsymbol{y}_2 = \{y_{i2}\}_{i=1}^{n_2}$ instead of $\boldsymbol{y}^{(n)}$ since $n_1 \neq n_2$.

Here, we compare tail probabilities for each distribution such that $\theta_j = Pr(y_{ij} > \kappa)$, where κ is a scalar value from the support of distribution $j = 1, 2$. The threshold of $\kappa = 4.82$ for this example is the median quarterly food expenditure per person (in MXN \$1000) for *upper* income households in Aguascalientes after accounting for weighting factors. Thus, we use the ratio θ_1/θ_2 to compare the probabilities that lower-middle income households with female and male providers spend at least as much on food per person as the typical upper income household. The observed proportions of households that spend at least \$4820 MXN on food per person are $\hat{\theta}_1 = 0.175$ and $\hat{\theta}_2 = 0.186$. We assign uninformative GAMMA$(2, 0.25)$ priors to both the shape α_j and rate λ_j parameters of the gamma model for group $j = 1, 2$. The gamma distribution is particularly well suited for design with sampling distribution segments because it is a popular statistical model that does not have a readily available conjugate prior. We let $\boldsymbol{\eta}_j = (\alpha_j, \lambda_j)$ for $j = 1, 2$. We obtain 10^5 posterior draws for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ using MCMC methods, which yields draws from the posterior of θ_1/θ_2. The gamma distributions characterized by the posterior means for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are superimposed on the histograms in Figure 3.1.

For illustration, we demonstrate two comparisons with posterior probabilities using a critical value of $\gamma = 0.8$. Figure 3.1 indicates that $Pr(\theta_1/\theta_2 \in (1, \infty) \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = 0.5429$. Because $0.5429 < 0.8$, we do not have convincing evidence that households with female providers are at least as likely to spend $\geq$ \$4820 MXN on food per person as those with male providers. We now suppose that a 10% relative increase or decrease in the gamma

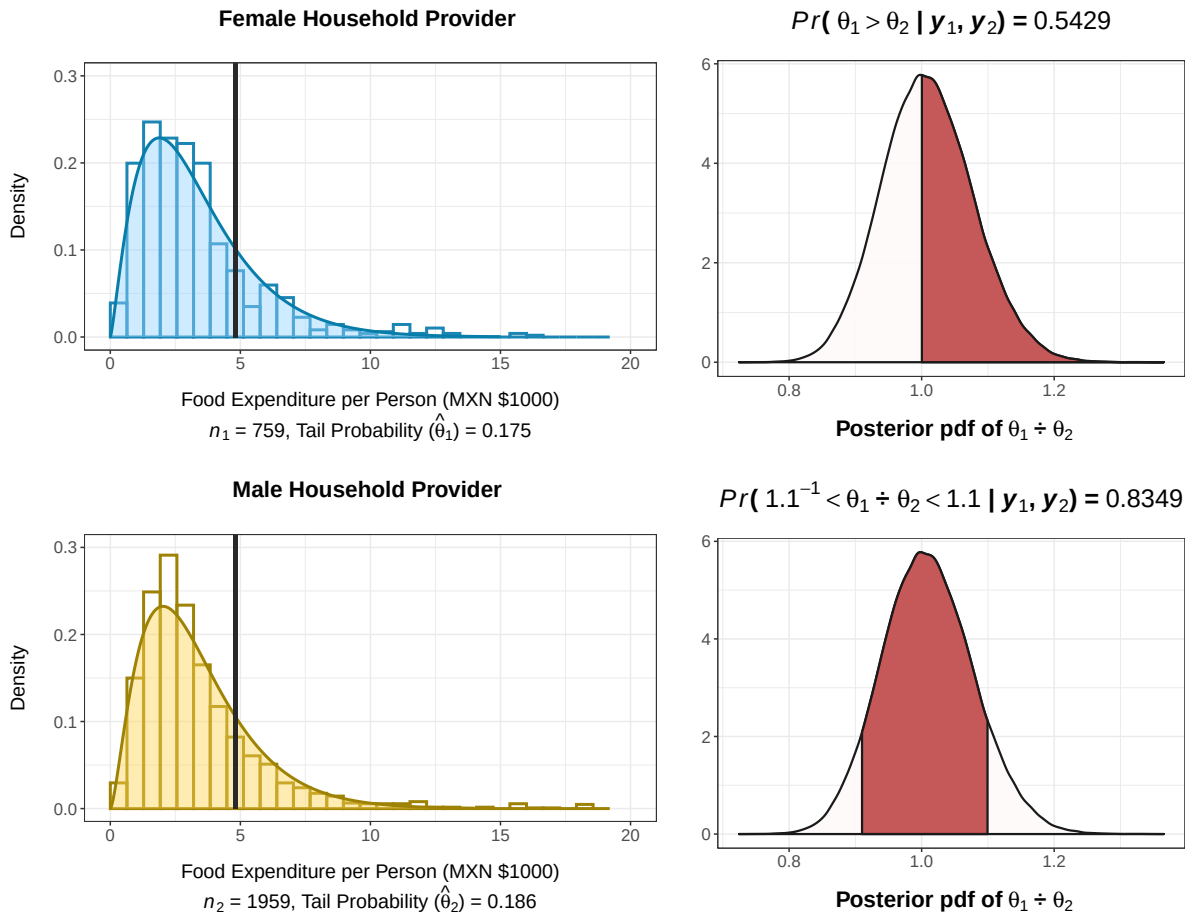


Figure 3.1: Group-specific summaries for quarterly food expenditure per person. Left: Food expenditure distributions. Right: Visualizations of the posterior probabilities.

tail probability is not of practical importance. Figure 3.1 also illustrates that $Pr(\theta_1/\theta_2 \in (1.1^{-1}, 1.1) \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = 0.8349 > 0.8$. Thus, households with female providers are practically as likely to spend $\geq$ \$4820 MXN on food per person as those with male providers.

3.4 Sampling Distributions of Posterior Probabilities

3.4.1 Analytical Approximations to the Posterior

Traditional approaches to power curve approximation for posterior analyses require that we estimate the sampling distribution of posterior probabilities for various sample sizes n. These approaches are slow because we wastefully estimate *entire* sampling distributions for sample sizes that are much too large or small when searching for a suitable sample size n. To map posterior probabilities to low-dimensional hypercubes and explore segments of sampling distributions, we leverage normal approximations to the posterior of θ based on limiting results. In this subsection, we overview several analytical posterior approximation methods and conditions that must hold for these approximations to be suitable.

The first normal approximation to the posterior that we consider follows from the Bernstein-von Mises (BvM) theorem (van der Vaart, 1998). We now describe how our framework for power curve approximation satisfies the conditions for the BvM theorem. Our framework assumes that data $\{y_{i1}\}_{i=1}^{n}$ and $\{y_{i2}\}_{i=1}^{n}$ are to be collected independently, where the data generation process for group j is characterized by the model $f(y;\boldsymbol{\eta}_{j,0})$ parameterized by $\boldsymbol{\eta}_j \in \mathbb{R}^d$. Here, $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$ are (fixed) user-specified design values for the distributional parameter(s). These distributions are herein referred to as *design* distributions. The design values $\boldsymbol{\eta}_{j,0}$ are different from the random variables $\boldsymbol{\eta}_j$ that parameterize the model for groups $j = 1, 2$ in Bayesian settings. When specifying the models $f(y;\boldsymbol{\eta}_{j,0})$, we also specify fixed values $\theta_{j,0}$ for the random variables θ_j: $\theta_{j,0} = g(\boldsymbol{\eta}_{j,0})$. We require that $g(\boldsymbol{\eta}_j)$ is differentiable at $\boldsymbol{\eta}_j = \boldsymbol{\eta}_{j,0}$ for $j = 1, 2$. A fixed value for the univariate characteristic $\theta_0 = h(\theta_{1,0}, \theta_{2,0})$ is also specified, where $h(\theta_1, \theta_2)$ is a differentiable function at $\theta_1 = \theta_{1,0}$ and $\theta_2 = \theta_{2,0}$. These derivatives of $g(\cdot)$ and $h(\cdot)$ must be nonzero.

The four assumptions that must be satisfied to invoke the BvM theorem (van der Vaart, 1998) are detailed in Appendix B.1.1. The first three assumptions involve the models $f(y;\boldsymbol{\eta}_{1,0})$ and $f(y;\boldsymbol{\eta}_{2,0})$; they are weaker than the regularity conditions for the asymptotic normality of the maximum likelihood estimator (MLE) (Lehmann and Casella, 1998), which are listed in Appendix B.1.2. The final assumption for the BvM theorem regards prior specification for the random variables. For our purposes, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the random variables for which we explicitly or implicitly assign prior distributions. We require that the prior distribution of $\boldsymbol{\eta}_j$ be continuous in a neighbourhood of $\boldsymbol{\eta}_{j,0}$ with positive density at $\boldsymbol{\eta}_{j,0}$ for $j = 1, 2$. This condition ensures that the posterior of θ converges to a neighbourhood of $\theta_0 = h(g(\boldsymbol{\eta}_{1,0}), g(\boldsymbol{\eta}_{2,0}))$. This convergence is required for our method for power curve approximation introduced in Section 3.5.

Under the conditions for the BvM theorem, the posterior of θ converges to the $\mathcal{N}(\theta_0, \mathcal{I}(\theta_0)^{-1}/n)$ distribution in the limit of infinite data (van der Vaart, 1998), where $\mathcal{I}(\theta_0)$ is the Fisher information for θ evaluated at θ_0. In practice, θ_0 is estimated from the to-be-observed data $\boldsymbol{Y}^{(n)}$ using the MLE $\hat{\theta}_n$ or the posterior mode $\tilde{\theta}_n$. In the limiting case, it does not matter which estimator for θ is used because both $\hat{\theta}_n$ and $\tilde{\theta}_n$ converge in probability to θ_0 when the conditions for the BvM theorem are satisfied. We use the MLE instead of the posterior mode for reasons discussed in Section 3.4.2. We therefore consider the following normal distribution based on the BvM theorem as one option to approximate the posterior of θ:

$$\mathcal{N}\left(h(g(\hat{\boldsymbol{\eta}}_{1,n}), g(\hat{\boldsymbol{\eta}}_{2,n})), \frac{1}{n}\sum_{j=1}^{2}\left[\frac{\partial h}{\partial \theta_j}\right]^2_{\theta_j = g(\hat{\boldsymbol{\eta}}_{j,n})}\left[\frac{\partial g}{\partial \boldsymbol{\eta}}^T \mathcal{I}(\boldsymbol{\eta})^{-1}\frac{\partial g}{\partial \boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}_{j,n}}\right). \tag{3.5}$$

To find $\mathcal{I}(\boldsymbol{\eta})^{-1}$, we find the limiting distributions for $\sqrt{n}(\hat{\boldsymbol{\eta}}_{j,n} - \boldsymbol{\eta}_{j,0})$, $j = 1, 2$. The multivariate delta method prompts the distribution in (3.5) since $\hat{\theta}_n = h(g(\hat{\boldsymbol{\eta}}_{1,n}), g(\hat{\boldsymbol{\eta}}_{2,n}))$ is a function of the MLEs $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$. We note that the variance in (3.5) is $\mathcal{I}^{-1}(\hat{\theta}_n)/n$. While it does not account for the priors, the approximation in (3.5) is useful because it prompts theoretical results about the limiting behaviour of the sampling distribution of posterior probabilities in Section 3.4.4.

We also consider the Laplace approximation to the posterior of θ that *does* account for the priors. This is useful when the sample size n is large enough to ensure the posterior is approximately normal but not large enough to guarantee the relevant priors have no substantial impact on the posterior mean and variance. For groups $j = 1$ and 2, the Laplace approximation is based on the Taylor series expansion of $\log(p_j(\boldsymbol{\eta}_j|data))$ centered at the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n} = \arg\max_{\boldsymbol{\eta}_j} p_j(\boldsymbol{\eta}_j|data)$ (Gelman et al., 2020). We henceforth consider the posteriors of $\boldsymbol{\eta}_j$, θ_j, and θ conditional on the general vector or matrix *data* instead of conditioning on $\boldsymbol{y}^{(n)}$, $\boldsymbol{y}_1$, or $\boldsymbol{y}_2$. We make this change in notation because our methods generate conduits for the data $\boldsymbol{y}^{(n)}$ as detailed in Section 3.4.2. This choice allows us to consider posterior probabilities prompted by conduits for the data and those produced by generating samples $\boldsymbol{y}^{(n)}$ using unified notation.

The multivariate delta method and the Laplace approximation prompt the following normal approximation to the posterior of θ that accounts for the priors $p_1(\boldsymbol{\eta}_1)$ and $p_2(\boldsymbol{\eta}_2)$:

$$\mathcal{N}\left(h(g(\tilde{\boldsymbol{\eta}}_{1,n}), g(\tilde{\boldsymbol{\eta}}_{2,n})), \sum_{j=1}^{2}\left[\frac{\partial h}{\partial \theta_j}\right]^2_{\theta_j = g(\tilde{\boldsymbol{\eta}}_{j,n})}\left[\frac{\partial g}{\partial \boldsymbol{\eta}}^T \mathcal{J}_j(\boldsymbol{\eta})^{-1}\frac{\partial g}{\partial \boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}_{j,n}}\right), \tag{3.6}$$

$$\text{where} \quad \mathcal{J}_j(\boldsymbol{\eta}) = -\frac{\partial^2}{\partial \boldsymbol{\eta}^2} \log(p_j(\boldsymbol{\eta}|data)).$$

We generally recommend using Laplace approximations with our framework for power curve approximation, but the approximation in (3.6) is computationally burdensome and suboptimal in certain situations as explained in Section 3.4.3.

3.4.2 Mapping Posteriors to Low-Dimensional Hypercubes

We next propose methods to map posteriors to low-dimensional hypercubes when using normal approximations to the posterior that do not and do account for the priors in this subsection and Section 3.4.3, respectively. The methods presented in the remainder of this section allow us to consider sampling distribution segments in Section 3.5. To generate samples of size n from each design distribution $f(y; \boldsymbol{\eta}_{1,0})$ and $f(y; \boldsymbol{\eta}_{2,0})$, one typically uses a pseudorandom sequence $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_m \in [0,1]^{2n}$ with length m. However, the approximation in (3.5) does not directly use the observations from the generated sample $\boldsymbol{y}^{(n)}$. Instead, $\boldsymbol{y}^{(n)}$ is used to compute maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$, which yield $\hat{\theta}_n = h(g(\hat{\boldsymbol{\eta}}_{1,n}), g(\hat{\boldsymbol{\eta}}_{2,n}))$. As such, we do not need to simulate data $\boldsymbol{Y}^{(n)}$ from the prior predictive distribution for a given sample size n. We instead recommend simulating from the approximate distributions for the MLEs of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. This reduces the dimension of the simulation from $2n$ to $2d$ (since $d << n$).

For sufficiently large n, the MLEs $\hat{\boldsymbol{\eta}}_{j,n}$ for groups $j = 1, 2$ approximately and independently follow $\mathcal{N}(\boldsymbol{\eta}_{j,0}, \mathcal{I}^{-1}(\boldsymbol{\eta}_{j,0})/n)$ distributions. The MLE – and not the posterior mode – is used with the approximation in (3.5) because we can easily simulate from its limiting distribution. Both $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$ have dimension d, so their joint limiting distribution has dimension $2d$. When using pseudorandom number generation, we now require a sequence $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_m \in [0,1]^{2d}$. Algorithm 3.1 details how we map a single point $\boldsymbol{u} = (u_1, u_2, ..., u_{2d}) \in [0,1]^{2d}$ to the posterior approximation based on the BvM theorem in (3.5), where $\hat{\boldsymbol{\eta}}_{j,n}^{(k)}$ and $\boldsymbol{\eta}_{j,0}^{(k)}$ denote the k^{th} component of these vectors.

In practice, we may require fewer observations for the sampling distributions of the MLEs to be approximately normal if we consider some transformation of $\boldsymbol{\eta}_j$. For the gamma model, both parameters in $\boldsymbol{\eta}_j = (\alpha_j, \lambda_j)$ must be positive, but the $\mathcal{N}(\boldsymbol{\eta}_{j,0}, \mathcal{I}^{-1}(\boldsymbol{\eta}_{j,0})/n)$ distribution could admit nonpositive values for small n. To obtain a sample of positive $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$ values for any sample size n with the gamma model, we exponentiate a sample of approximately normal MLEs of $\log(\boldsymbol{\eta}_1)$ and $\log(\boldsymbol{\eta}_2)$. For an arbitrary model, appropriate transformations could similarly be applied to any parameters in $\boldsymbol{\eta}_j$ that do not have support on $\mathbb{R}$. Similarly, the posterior of a monotonic transformation of θ may need to be considered

Algorithm 3.1 Mapping Posteriors to $[0,1]^{2d}$ with the BvM Theorem

procedure MapBvM($f(y;\boldsymbol{\eta}_{1,0})$, $f(y;\boldsymbol{\eta}_{2,0})$, $g(\cdot)$, $h(\cdot)$, n, $\boldsymbol{u}$)
  for j in 1:2 **do**
    for k in 1:d **do**
      Generate $\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u})^{(k)}$ as the $u_{(j-1)d+k}$-quantile of the conditional normal CDF of $\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u})^{(k)} \mid \{\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u})^{(l)}\}_{l=0}^{k-1}$ where $\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{\eta}_{j,0}, \mathcal{I}(\boldsymbol{\eta}_{j,0})^{-1}/n)$.
  Use $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u})$, $\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u})$, and the partial derivatives of $g(\cdot)$ and $h(\cdot)$ to obtain (3.5).

for the normal approximation in (3.5) to be suitable for moderate n. For instance, the posterior of $\log(\theta_1) - \log(\theta_2)$ may be better approximated by a normal distribution than that of θ_1/θ_2. Rather than introduce new notation for these untransformed and transformed variables, we assume that $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, and θ are specified to improve the quality of the relevant normal approximations in (3.5) and (3.6). Because priors are typically specified for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ before making such transformations, relevant Jacobians must be considered when using normal approximations to the posterior that account for the prior distributions.

3.4.3 Mapping Posteriors with Prior Information

The method for mapping posteriors to $[0,1]^{2d}$ proposed in Section 3.4.2 does not account for the prior distributions. The Laplace approximation to the posterior of θ in (3.6) that accounts for the priors requires an observed sample $\boldsymbol{y}^{(n)}$ – not just the maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$. In this subsection, we present two methods for posterior mapping that account for the priors. The first method is ideal when the design distributions belong to the exponential family (Lehmann and Casella, 1998), whereas the second method allows for more flexibility when specifying the models $f(y;\boldsymbol{\eta}_{1,0})$ and $f(y;\boldsymbol{\eta}_{2,0})$.

When the design distributions belong to the exponential family, the relevant probability mass or density function takes the form

$$f(y;\boldsymbol{\eta}_j) = \exp\left[\sum_{s=1}^{d} C_s(\boldsymbol{\eta}_j)T_s(y) - A(\boldsymbol{\eta}_j) + B(y)\right],$$

where $A(\boldsymbol{\eta}_j)$, $B(y)$, $C_s(\boldsymbol{\eta}_j)$, and $T_s(y)$ are known functions for $s = 1, ..., d$. For group j, $T_{j\dagger}(\boldsymbol{y}^{(n)}) = (\sum_{i=1}^{n} T_1(y_{ij}), ..., \sum_{i=1}^{n} T_d(y_{ij}))$ are called sufficient statistics that provide as much information about the parameter $\boldsymbol{\eta}_j$ as the entire sample. The first derivative of the

log-likelihood with respect to the k^{th} component of $\boldsymbol{\eta}_j$ is then

$$\frac{\partial}{\partial \boldsymbol{\eta}_j^{(k)}} l(\boldsymbol{\eta}_j; \boldsymbol{y}^{(n)}) = -n \frac{\partial}{\partial \boldsymbol{\eta}_j^{(k)}} A(\boldsymbol{\eta}_j) + \sum_{s=1}^{d} \frac{\partial}{\partial \boldsymbol{\eta}_j^{(k)}} C_s(\boldsymbol{\eta}_j) \sum_{i=1}^{n} T_s(y_{ij}). \tag{3.7}$$

At the maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{j,n}$ for an observed sample $\boldsymbol{y}^{(n)}$, all d partial derivatives in (3.7) equal 0. A d-parameter model in the exponential family has d nonredundant sufficient statistics, so all components of $T_{j\dagger}(\boldsymbol{y}^{(n)})$ can be recovered by substituting the maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{j,n}$ into the system of linear equations in (3.7). Algorithm 3.2 details how we map a single point $\boldsymbol{u} \in [0,1]^{2d}$ to the posterior approximation in (3.6) based on Laplace's method. For models $f(y; \boldsymbol{\eta}_{1,0})$ and $f(y; \boldsymbol{\eta}_{2,0})$ in the exponential family, we emphasize that $p_j(\boldsymbol{\eta}_j | data) = p_j(\boldsymbol{\eta}_j | T_{j\dagger}(\boldsymbol{y}^{(n)}))$ for $j = 1, 2$.

Algorithm 3.2 Mapping Posteriors to $[0,1]^{2d}$ with Laplace's Method

1: **procedure** MapLaplace($f(y; \boldsymbol{\eta}_{1,0})$, $f(y; \boldsymbol{\eta}_{2,0})$, $g(\cdot)$, $h(\cdot)$, n, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$)
2: Generate $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u})$ and $\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u})$ using Lines 2 to 4 of Algorithm 3.1.
3: **for** j in 1:2 **do**
4: Equate the system of equations in (3.7) to 0 with $\boldsymbol{\eta}_j = \hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u})$ to solve for $T_{j\dagger}(\boldsymbol{y}^{(n)})$.
5: Use $T_{j\dagger}(\boldsymbol{y}^{(n)})$ to obtain the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n}$ via optimization.
6: Use $\tilde{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u})$, $\tilde{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u})$, $T_{1\dagger}(\boldsymbol{y}^{(n)})$, and $T_{2\dagger}(\boldsymbol{y}^{(n)})$ along with the partial derivatives of $g(\cdot)$ and $h(\cdot)$ to obtain (3.6).

We note that Algorithm 3.2 may not provide a serviceable approach when the design distributions are not members of the exponential family. For instance, this method could not be applied if a Weibull model were chosen for the motivating example in Section 3.3 in lieu of the gamma distribution. When the shape parameter of the Weibull distribution is unknown, its minimal sufficient statistic consists of the entire sample: $\{y_{ij}\}_{i=1}^{n}$ for $j = 1, 2$. We therefore develop a hybrid approach to posterior mapping that accounts for the priors when low-dimensional sufficient statistics cannot be recovered from the maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$.

This hybrid approach leverages the following result, which holds true when $\boldsymbol{\eta}_j \approx \hat{\boldsymbol{\eta}}_{j,n}$ for sufficiently large n:

$$\log(p_j(\boldsymbol{\eta}_j | \boldsymbol{y}^{(n)})) \approx l(\hat{\boldsymbol{\eta}}_{j,n}; \boldsymbol{y}^{(n)}) - \frac{n}{2}(\boldsymbol{\eta}_j - \hat{\boldsymbol{\eta}}_{j,n})^T \mathcal{I}(\hat{\boldsymbol{\eta}}_{j,n})(\boldsymbol{\eta}_j - \hat{\boldsymbol{\eta}}_{j,n}) + \log(p_j(\boldsymbol{\eta}_j)). \tag{3.8}$$

This result follows from the second-order Taylor approximation to the log-posterior of

$\boldsymbol{\eta}_j$ around $\hat{\boldsymbol{\eta}}_{j,n}$, where the observed information is replaced with the (expected) Fisher information. The approximation to the log-likelihood function does not have a first-order term because the score function is 0 at $\hat{\boldsymbol{\eta}}_{j,n}$. We note that although the first term on the right side of (3.8) depends on the data $\boldsymbol{y}^{(n)}$, it is a constant. An approximation to the posterior mode is the value that maximizes the right side of (3.8): $\boldsymbol{\eta}^*_{j,n}$. We consider the following normal approximation to the posterior of θ:

$$\mathcal{N}\left(h(g(\boldsymbol{\eta}^*_{1,n}), g(\boldsymbol{\eta}^*_{2,n})), \sum_{j=1}^{2} \left[\frac{\partial h}{\partial \theta_j}\right]^2_{\theta_j = g(\boldsymbol{\eta}^*_{j,n})} \left[\frac{\partial g}{\partial \boldsymbol{\eta}}^T \mathcal{J}^*_j(\boldsymbol{\eta})^{-1} \frac{\partial g}{\partial \boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\boldsymbol{\eta}^*_{j,n}}\right), \tag{3.9}$$

$$\text{where} \quad \mathcal{J}^*_j(\boldsymbol{\eta}) = n\mathcal{I}(\boldsymbol{\eta}) - \frac{\partial^2}{\partial \boldsymbol{\eta}^2}\log(p_j(\boldsymbol{\eta})).$$

The observed information is again replaced with the Fisher information in $\mathcal{J}^*_j(\boldsymbol{\eta})$ of (3.9) since we do not generate samples $\boldsymbol{y}^{(n)}$. Algorithm 3.3 details how we map a single point $\boldsymbol{u} \in [0,1]^{2d}$ to the posterior approximation in (3.9).

Algorithm 3.3 Mapping Posteriors to $[0,1]^{2d}$ with a Hybrid Method

1: **procedure** MapHybrid($f(y;\boldsymbol{\eta}_{1,0})$, $f(y;\boldsymbol{\eta}_{2,0})$, $g(\cdot)$, $h(\cdot)$, n, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$)
2: Generate $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u})$ and $\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u})$ using Lines 2 to 4 of Algorithm 3.1.
3: **for** j in 1:2 **do**
4: Obtain $\boldsymbol{\eta}^*_{j,n}$ as $\arg\max_{\boldsymbol{\eta}_j}$ of the right side of (3.8) anchored at $\boldsymbol{\eta}_{j,n} = \hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u})$.
5: Use $\boldsymbol{\eta}^*_{1,n}$, $\boldsymbol{\eta}^*_{2,n}$, and the partial derivatives of $g(\cdot)$ and $h(\cdot)$ to obtain (3.9).

3.4.4 Theoretical Properties of the Power Estimates

Now that we have developed three algorithms to reduce the simulation dimension in a variety of settings, we consider the theoretical properties of the resulting power estimates. We introduce general notation to define power estimates for the simulation method ζ, where ζ is Algorithm 3.1, 3.2, or 3.3. We let $\mathcal{N}(\underline{\theta}^{(n)}_r, \underline{\tau}^{(n)}_r)$ denote the relevant normal approximation to the posterior of θ corresponding to the point $\boldsymbol{u}_r \in [0,1]^{2d}$ and sample size n for $r = 1, ..., m$. This approximation is respectively (3.5) for Algorithm 3.1, (3.6) for Algorithm 3.2, and (3.9) for Algorithm 3.3. We incorporate the sample size n into this notation because the mean $\underline{\theta}^{(n)}_r$ depends on the sample size of the joint limiting distribution for $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$. The variance $\underline{\tau}^{(n)}_r$ is also an explicit function of n in (3.5) and (3.9) and an implicit function of the sample size in (3.6).

The estimate for the posterior probability $Pr(\theta < \delta \mid data)$ is then

$$p^{\delta}_{n,\boldsymbol{u}_r,\zeta} = \Phi\left(\frac{\delta - \underline{\theta}_r^{(n)}}{\sqrt{\underline{\tau}_r^{(n)}}}\right), \tag{3.10}$$

where $\Phi(\cdot)$ is the the CDF of the standard normal distribution. The estimates from (3.10) can be used to determine whether the criterion inside the indicator function from (3.1) or (3.4) is satisfied for a particular posterior approximation $\mathcal{N}(\underline{\theta}_r^{(n)}, \underline{\tau}_r^{(n)})$. For a given sample size n, the proportion of the m approximate posteriors corresponding to $\boldsymbol{u}_1, ..., \boldsymbol{u}_m$ for which the relevant criterion is satisfied estimates power. Because the limiting posterior of θ is normal when the conditions for the BvM theorem hold, power for hypothesis testing with HDIs defined in (3.3) should be approximated by power defined in (3.4) for sufficiently large n.

No matter which of our three algorithms are used, the accuracy of these power estimates depends on the quality of the approximation to the sampling distribution of posterior probabilities. If data $\boldsymbol{Y}^{(n)}$ are generated from the design distributions and the assumptions for the BvM theorem hold true, the normal approximations to the posterior of θ given in (3.5) and (3.6) are well established. Theorem 3.1 compares the sampling distributions of posterior probabilities induced by (3.5) and (3.6) with data $\boldsymbol{Y}^{(n)}$ to the sampling distribution prompted by Algorithm 3.1 with pseudorandom sequences as $n \to \infty$.

Theorem 3.1. *Let $f(y;\boldsymbol{\eta}_{1,0})$ and $f(y;\boldsymbol{\eta}_{2,0})$ satisfy the regularity conditions from Appendix B.1.2. Let the prior $p_j(\boldsymbol{\eta}_j)$ be continuous in a neighbourhood of $\boldsymbol{\eta}_{j,0}$ with positive density at $\boldsymbol{\eta}_{j,0}$ for $j = 1, 2$. Let $g(\boldsymbol{\eta})$ and $h(\theta_1, \theta_2)$ be respectively differentiable at $\boldsymbol{\eta}_{j,0}$ and $\theta_{j,0} = g(\boldsymbol{\eta}_{j,0})$ for $j = 1, 2$ with nonzero derivatives. Let $\boldsymbol{U} \overset{i.i.d.}{\sim} \mathcal{U}([0,1]^{2d})$ and $\boldsymbol{Y}^{(n)}$ be generated independently from $f(y;\boldsymbol{\eta}_{1,0})$ and $f(y;\boldsymbol{\eta}_{2,0})$. Let $\mathcal{P}^{\delta}_{n,\Pi,\zeta}$ denote the sampling distribution of posterior probabilities for $Pr(\theta < \delta \mid data)$ given sample size n produced using input Π with method ζ. Let $\|Q_1 - Q_2\|_{TV}$ be the total variation distance between two probability measures Q_1 and Q_2. Then,*

(a) $\|\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.5)} - \mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}\|_{TV} \xrightarrow{P} 0.$

(b) $\|\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.6)} - \mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}\|_{TV} \xrightarrow{P} 0.$

The proof of Theorem 3.1 is given in Appendix B.1.3. While the approximations in (3.6) and (3.9) should better account for the prior distributions for moderate sample sizes n, they do not differ from the approximation in (3.5) in the limit of infinite data. Likewise, $\|\mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.2} - \mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}\|_{TV}$ and $\|\mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.3} - \mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}\|_{TV}$ will converge to 0 as $n \to \infty$

under the conditions for Theorem 3.1. This result is straightforward because $\tilde{\boldsymbol{\eta}}_{j,n} - \hat{\boldsymbol{\eta}}_{j,n}$ and $\boldsymbol{\eta}^*_{j,n} - \hat{\boldsymbol{\eta}}_{j,n}$ converge in probability to 0, and $\mathcal{J}_j(\tilde{\boldsymbol{\eta}}_{j,n})/n$ and $\mathcal{J}^*_j(\boldsymbol{\eta}^*_{j,n})/n$ converge in probability to $\mathcal{I}(\boldsymbol{\eta}_{j,0})$ given results from Appendix B.1.3. These results prompt Corollary 3.1, which follows from Theorem 3.1.

Corollary 3.1. *Let $p^{\delta}_{n,\boldsymbol{u}_r,\zeta}$ from (3.10) be the estimate for $Pr(\theta < \delta \mid data)$ corresponding to sample size n, method ζ, and point $\boldsymbol{u}_r \in [0,1]^{2d}$. Under the conditions for Theorem 3.1, power in (3.1) and (3.4) is consistently estimated by*

$$\frac{1}{m}\sum_{r=1}^{m} \mathbb{I}\{p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} \geq \gamma\} \quad \textit{and} \quad \frac{1}{m}\sum_{r=1}^{m} \mathbb{I}\{p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} < \alpha/2 \ \cap \ 1 - p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} < \alpha/2\},$$

respectively, when ζ is Algorithm 3.1, 3.2, or 3.3 and $\boldsymbol{U}_r \overset{i.i.d.}{\sim} \mathcal{U}([0,1]^{2d})$ for $r = 1, ..., m$ as $n \to \infty$.

Corollary 3.1 ensures that our three algorithms give rise to consistent power estimates as $n \to \infty$; however, it does not guarantee that these estimators are unbiased for finite n. When the assumptions for Theorem 3.1 are satisfied, our power estimates are suitable for sufficiently large n. To optimize the performance of these design methods for moderate n, one should consider transformations of θ or certain components in $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ to improve the normal approximations to the relevant posterior and MLE distributions.

3.5 Fast Power Curve Approximation for Posterior Analyses

3.5.1 Power Estimates with Fewer Posteriors

This section details how we leverage the mappings between posterior probabilities and $[0,1]^{2d}$ to expedite power curve approximation for posterior analyses. The novelty of this computational efficiency stems from using sampling distribution segments to approximate the power curve, and our approach for power curve approximation described in Algorithm 3.4 is the main contribution of this chapter. Before introducing that approach, we justify why low-discrepancy sequences can be used instead of pseudorandom ones to reduce the number of posteriors required for precise power estimates.

In Section 1.2.3, we explained that each point in an appropriately randomized low-discrepancy sequence is such that $\boldsymbol{U}_r \sim U\left([0,1]^{2d}\right)$ for $r = 1, ..., m$. We use randomized

Sobol' sequences (Sobol', 1967) to estimate power in this chapter. Based on the discussion in Section 1.2.3, randomized Sobol' sequences prompt consistent estimators for power:

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^{m}\Psi(\boldsymbol{U}_r)\right) = \int_{[0,1]^{2d}}\Psi(\boldsymbol{u})d\boldsymbol{u}, \tag{3.11}$$

where either indicator function in Corollary 3.1 is the relevant function $\Psi(\cdot)$. Corollary 3.2 formalizes this result.

Corollary 3.2. *Under the conditions for Theorem 3.1, using a randomized low-discrepancy sequence* $\boldsymbol{U}_1, ..., \boldsymbol{U}_m$ *in lieu of a sequence* $\boldsymbol{U}_r \overset{i.i.d.}{\sim} \mathcal{U}([0,1]^{2d})$ *for* $r = 1, ..., m$ *does not impact the consistency of the power estimates from Corollary 3.1 as* $n \to \infty$.

Due to the negative dependence between the points, the variance of the estimator in (3.11) is typically reduced by using low-discrepancy sequences. By (1.8), randomized Sobol' sequences reduce the number of posteriors for θ that we must approximate to obtain precise, consistent power estimates. We leveraged similar results in Chapter 2 to reduce the number of simulation repetitions required for frequentist power estimation. Corollary 3.2 will be applied later to ensure the consistency of our sample size recommendations based on segments of approximate sampling distributions of posterior probabilities.

3.5.2 Selection of Sampling Distribution Segments

For a given sample size n, we *could* obtain power estimates using the formulas in Corollary 3.1 with randomized Sobol' sequences. Sample size determination could be conducted by repeating this process for various values of n until a suitable sample size is found. However, such a process would waste computational resources thoroughly exploring sampling distributions for unsuitable sample sizes in order to obtain an appropriate one. We note that the (θ_1, θ_2)-space such that $\theta = h(\theta_1, \theta_2) \in (\delta_L, \delta_U)$ is convex when $\theta = \theta_1 - \theta_2$ or $\theta = \theta_1/\theta_2$. If we appeal to this convexity, we argue that consistent power estimates for a given sample size n can often be obtained with only a subset of the points $\boldsymbol{u}_r \in [0,1]^{2d}$, $r = 1, ..., m$. By using only a subset of points to explore most sample sizes, we consider only segments of the relevant sampling distributions.

In each of our algorithms, the approximately normal posterior $\mathcal{N}(\underline{\theta}_r^{(n)}, \underline{\tau}_r^{(n)})$ and corresponding posterior probabilities $p^{\delta}_{n,\boldsymbol{u}_r,\zeta}$ depend on the design distributions, the sample size n, and the Sobol' sequence point $\boldsymbol{u}_r$, $r = 1, ..., m$. We have previously fixed the sample size n and allowed the point $\boldsymbol{u}_r \in [0,1]^{2d}$ to vary when estimating power. We now fix the

point $\boldsymbol{u}_r$ and let the sample size n vary. When the point $\boldsymbol{u}_r$ and design distributions are fixed, $p^{\delta}_{n,\boldsymbol{u}_r,\zeta}$ is a deterministic function of n. Lemma 3.1 motivates our approach to choose subsets of points $\boldsymbol{u}_r \in [0,1]^{2d}$ for each sample size n explored.

Lemma 3.1. *Let the conditions for Theorem 3.1 be satisfied. For a given point* $\boldsymbol{u}_r = (u_1, ..., u_{2d}) \in [0,1]^{2d}$*, we have that Algorithms 3.1, 3.2, and 3.3 prompt*

(a) $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(k)} = \boldsymbol{\eta}^{(k)}_{1,0} + \dfrac{\omega_k(u_1, ..., u_k)}{\sqrt{n}}$ *and* $\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u}_r)^{(k)} = \boldsymbol{\eta}^{(k)}_{2,0} + \dfrac{\omega_{d+k}(u_{d+1}, ..., u_{d+k})}{\sqrt{n}}$ *for* $k = 1, ..., d$*, where* $\omega_k(\cdot)$ *and* $\omega_{d+k}(\cdot)$ *are functions that do not depend on* n*.*

(b) $h(g(\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)), g(\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u}_r))) \approx h(g(\boldsymbol{\eta}_{1,0}), g(\boldsymbol{\eta}_{2,0})) + \dfrac{\omega_{\dagger}(u_1, ..., u_{2d})}{\sqrt{n}}$ *for sufficiently large* n*, where* $\omega_{\dagger}(\cdot)$ *is a function that does not depend on* n*.*

(c) $p^{\delta}_{n,\boldsymbol{u}_r,\zeta} \approx \Phi\left(a(\delta, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)\right)$ *for sufficiently large* n*, where* θ_0 *is the design value for* θ *and* $a(\cdot)$ *and* $b(\cdot)$ *are functions that do not depend on* n*.*

(d) When $\theta_0 \in (\delta_L, \delta_U)$*,* $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta}$ *is a increasing function of* n *for sufficiently large sample sizes.*

We prove Lemma 3.1 in Appendix B.2 and now consider its implications on the posterior probability of interest $p^{\delta_U}_{n_A,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n_A,\boldsymbol{u}_r,\zeta}$ when $\theta_0 \in (\delta_L, \delta_U)$ for a given point $\boldsymbol{u}_r$. If $p^{\delta_U}_{n_A,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n_A,\boldsymbol{u}_r,\zeta} \geq \gamma$, then $p^{\delta_U}_{n_B,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n_B,\boldsymbol{u}_r,\zeta} \geq \gamma$ for sufficiently large $n_A < n_B$. The (θ_1, θ_2)-space such that $\theta = h(\theta_1, \theta_2) \in (\delta_L, \delta_U)$ is convex, which limits the potential for decreasing behaviour of $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta}$ as a function of n for small and moderate sample sizes. In light of this, our method to approximate the power curve generates a single Sobol' sequence of length m. For hypothesis tests with posterior probabilities, we use root finding algorithms (Brent, 1973) to find the value for n such that $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} - \gamma = 0$ for $r = 1, ..., m$. The empirical CDF of these m sample sizes approximates the power curve. As demonstrated in Section 3.5.3, this root-finding approach facilitates targeted exploration of $[0,1]^{2d}$ based on the sample size n. Since the posterior probabilities in Corollary 3.1 are mapped to $\boldsymbol{u}_r \in [0,1]^{2d}$, this approach also allows us to consider segments of the approximate sampling distribution of posterior probabilities.

3.5.3 Power Curves with Sampling Distribution Segments

Algorithm 3.4 formally describes our method for power curve approximation with sampling distribution segments for analyses facilitated via posterior probabilities and Bayes factors.

We later discuss the necessary modifications for analyses with credible intervals. To implement this approach, we must choose a parametric statistical model $f(y;\boldsymbol{\eta})$, functions $g(\cdot)$ and $h(\cdot)$, priors $p_1(\boldsymbol{\eta}_1)$ and $p_2(\boldsymbol{\eta}_2)$, and design values $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$. We recommend using visualization techniques to choose the design values, and the literature on prior elicitation could be of use when specifying the design distributions (Chaloner, 1996; Garthwaite et al., 2005; Johnson et al., 2010). We must also select an interval (δ_L, δ_U), a critical value γ, a target power $1-\beta$, a method ζ consisting of one of the three algorithms proposed in Section 3.4, and the length of the Sobol' sequence m. We use $m = 1024$ to balance the computational efficiency and precision of the approximation to the power curve.

Algorithm 3.4 Procedure for Power Curve Approximation with Posterior Probabilities

1: **procedure** PowerCurve($f(y;\boldsymbol{\eta}_{1,0})$, $f(y;\boldsymbol{\eta}_{2,0})$, $g(\cdot)$, $h(\cdot)$, $p_j(\boldsymbol{\eta}_j)$, (δ_L, δ_U), γ, β, ζ, m)
2: Let n_0 equate the left side of (3.1) to $1-\beta$ when $p(\theta \mid data) \approx \mathcal{N}(\hat{\theta}_n, \mathcal{I}(\theta_0)^{-1}/n)$
3: `sampSobol` $\leftarrow$ NULL
4: **for** r in 1:m **do**
5: Generate Sobol' sequence point $\boldsymbol{u}_r$
6: Let `sampSobol`$[r]$ solve $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} - \gamma = 0$ in terms of n, initializing the root-finding algorithm at $\lceil n_0 \rceil$
7: `powerCurve` $\leftarrow$ empirical CDF of `sampSobol`
8: Let n_* be the $(1-\beta)$-quantile of `sampSobol`
9: **for** r in 1:m **do**
10: **if** `sampSobol`$[r] \leq n_*$ **then**
11: **if** $p^{\delta_U}_{n_*,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n_*,\boldsymbol{u}_r,\zeta} - \gamma < 0$ **then**
12: Repeat Line 6, initializing the root-finding algorithm at n_*
13: **else**
14: **if** $p^{\delta_U}_{n_*,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n_*,\boldsymbol{u}_r,\zeta} - \gamma \geq 0$ **then**
15: Repeat Line 6, initializing the root-finding algorithm at n_*
16: `powerCurveFinal` $\leftarrow$ empirical CDF of `sampSobol`
17: Let n^* be the $(1-\beta)$-quantile of `sampSobol`
18: **return** `powerCurveFinal`, $\lceil n^* \rceil$ as recommended sample size

Line 2 of Algorithm 3.4 uses the normal approximation in (3.5) with known variance to obtain a starting point n_0 for the root-finding algorithm. We can obtain this starting point in a fraction of a second under the assumption that $\hat{\theta}_n \sim \mathcal{N}(\theta_0, \mathcal{I}(\theta_0)^{-1}/n)$, and it should be close to the final sample size recommendation if uninformative priors are used. Since the root-finding algorithm is initialized at $\lceil n_0 \rceil$ for all points $\boldsymbol{u}_r$, $r = 1, ..., m$, the entire sampling distribution of posterior probabilities is explored for that sample size. In Lines

4 to 6, the root-finding algorithm then facilitates targeted exploration of the approximate distribution of posterior probabilities for all other sample sizes considered. We complete the power curve approximation procedure by exploring the entire sampling distribution of posterior probabilities at the sample size n_* in Lines 9 to 15. If the statements in Lines 11 or 14 are true, this implies that $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} = \gamma$ for at least two distinct sample sizes n. For these points $\boldsymbol{u}_r$, we can reinitialize the root-finding algorithm at n_* to obtain a solution for each point that will make the power curve consistent at n_*. This consistency is a direct consequence of Corollary 3.2.

In Section 3.6, we conduct numerical studies to demonstrate the suitability of the power curves obtained by Algorithm 3.4 in various settings. These numerical results show that the if statements in Lines 11 and 14 are very rarely true for any point $\boldsymbol{u}_r \in [0,1]^{2d}$ when sample sizes are large enough for the BvM theorem to hold. In those situations, $n_* = n^*$ and both the power estimate at n^* and the sample size recommendation $\lceil n^* \rceil$ are consistent. It is incredibly unlikely that n_* and n^* would differ substantially, but Lines 9 to 15 of Algorithm 3.4 could be repeated in that event, where the root-finding algorithm is initialized at n^* instead of n_*.

These consistent sample size recommendations are not guaranteed to be unbiased since the normal approximation to the relevant posterior and MLE distributions may introduce noticeable bias for finite n. To efficiently verify the suitability of the sample size recommendation, we could simulate various samples of size $\lceil n^* \rceil$. These samples could be used to compare (i) the empirical sampling distributions of $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$ with their normal approximations and (ii) the posterior of θ with its normal approximation. Even though the power curves from Algorithm 3.4 are consistent near the target power $1-\beta$, their global consistency at all sample sizes n is not guaranteed because part (b) of Lemma 3.1 and the BvM theorem are large-sample results. Nevertheless, our numerical studies in Section 3.6 highlight good global estimation of the power curve, particularly when the posterior approximation method ζ accounts for the priors.

We discuss how to generalize Algorithm 3.4 for analyses with credible intervals below. If one of δ_L or δ_U is not finite, Algorithm 3.4 can be used without modification where $\gamma = 1-\alpha$. Otherwise, the initial value for the root-finding algorithm n_0 found in Line 3 is based on (3.4) instead of (3.1). In Line 6, `sampSobol`$[r]$ is modified to be the maximum of the two solutions for $1 - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} - (1-\alpha/2) = 0$ and $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - (1-\alpha/2) = 0$. We lastly modify how the sampling distribution of posterior probabilities at the sample size n_* is explored. To do so, we confirm that both $1 - p^{\delta_L}_{n_*,\boldsymbol{u}_r,\zeta} - (1-\alpha/2)$ and $p^{\delta_U}_{n_*,\boldsymbol{u}_r,\zeta} - (1-\alpha/2)$ are not less than or at least 0 in Lines 11 and 14, respectively. Effectively, power curve approximation with equal-tailed credible intervals requires us to consider two hypothesis tests with posterior probabilities and intervals (δ_L, ∞) and $(-\infty, \delta_U)$.

The root-finding algorithm from Lines 4 to 6 of Algorithm 3.4 approximates posteriors corresponding to $O(\log_2 B)$ points from $[0,1]^{2d}$, where B is the maximum sample size considered for the power curve. We would require $O(B)$ such points to explore a similar range of sample sizes by thoroughly exploring sampling distributions via Corollary 3.1 with randomized Sobol' sequences. When $B \geq 59$, our approach prompts at least an order of magnitude reduction in the number of posterior approximations because $O(\log_2 B) < O(B)/10$. Moreover, using randomized Sobol' sequences in lieu of pseudorandom ones has generally allowed us to estimate power curves with similar precision using about an order of magnitude fewer points from $[0,1]^{2d}$. We visualize the insights from this paragraph using a simpler example than the four-dimensional one with gamma tail probabilities in Section 3.5.4.

3.5.4 Visualization of Computational Efficiency

The motivating example with gamma tail probabilities from Section 3.3 has dimension of $2d = 4$ since we consider the posteriors of $\boldsymbol{\eta}_j = (\alpha_j, \lambda_j)$ for $j = 1, 2$. Here, we consider a two-dimensional example involving the comparison of Bernoulli proportions that is easier to visualize. Even though this model has a conjugate prior, it is considered for illustrative purposes. We let θ_j be the Bernoulli success probability for model $j = 1, 2$. For this example, we let $\eta_j = \log(\theta_j) - \log(1-\theta_j)$ to improve the quality of the normal approximations to the posterior for η_j and the sampling distribution of its MLE. This simple transformation is based on the canonical form of the standard Bernoulli model (Lehmann and Casella, 1998).

We compare the Bernoulli probabilities via their difference: $\theta_1 - \theta_2$. Because $\theta_1 - \theta_2 \in (-1, 1)$, we similarly consider the posterior distribution of $\log(\theta_1 - \theta_2 + 1) - \log(1 - \theta_1 + \theta_2)$ to improve the quality of the normal approximations for moderate n. The selected transformation for θ is a straightforward generalization of the transformation applied to each η_j parameter. An alternative monotonic transformation could have been chosen to minimize the Kullback-Leibler divergence (Gelman et al., 2020) for the normal approximation to the posterior. However, the optimal transformation likely depends on the sample size and the data generated from the design distributions. Any transformations such that $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ have support over $\mathbb{R}^d$ and θ has support over the entire real line are suitable in the limiting case. To streamline our methods, we suggest using visualization techniques to compare the suitability of several candidate monotonic transformations if necessary.

We choose design values of $\theta_{1,0} = 0.15$ and $\theta_{2,0} = 0.14$, which gives rise to design values $\eta_{1,0}$ and $\eta_{2,0}$ on the logit scale. We specify informative priors for the Bernoulli parameters:

BETA$(3.75, 21.25)$ for θ_1 and BETA$(3.50, 21.50)$ for θ_2. These beta distributions have modes that roughly align with the design values $\theta_{1,0}$ and $\theta_{2,0}$, and these beta priors induce priors on the variables η_1 and η_2. We consider power curve approximation for analyses with posterior probabilities, where $(\delta_L, \delta_U) = (-0.05, 0.05)$ on the probability scale, $\gamma = 0.8$, and $1 - \beta = 0.6$. For concision, we only consider the method to map posteriors to $[0, 1]^{2d}$ from Algorithm 3.2 with $m = 1024$.

The left plot of Figure 3.2 decomposes the results of the root-finding algorithm for one approximated power curve for this Bernoulli example. For instance, only the pink points in the lower left corner assessed power for at least one sample size $n \in (2, 150)$ when input into the root-finding procedure. We note that in most iterations of the root-finding algorithm, the sample size n is noninteger, which does not present issues for our method. This targeted exploration approach allows us to prioritize segments of the sampling distribution of posterior probabilities based on the sample size n. Only the blue points from $[0, 1]^2$ considered at least one sample size $n \in (175, 225)$. These are the points for which the posterior probability in (3.1) is close to the critical value $\gamma = 0.8$. For reference, the final sample size recommendation for this example was $\lceil n^* \rceil = 269$.

The right plot of Figure 3.2 visualizes segments from the sampling distribution of posterior probabilities for $n = 200$ conditional on the categorizations from the left plot. It is wasteful to use the pink points to consider $n \in (175, 225)$ because those points satisfy $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta} = \gamma$ for some sample size $n \in (2, 150)$. Lemma 3.1 can be invoked for the largest sample sizes in this interval. Based on Lemma 3.1, the posterior probabilities for the pink points for $n \in (175, 225)$ should therefore be much larger than γ. This result is confirmed by comparing the pink density to the dotted vertical line at $\gamma = 0.8$. By similar logic, it is wasteful to consider the red points for $n \in (175, 225)$ since the posterior probabilities for those points will be much smaller than γ.

In Figure 3.2, the sample size categories were created to exclude most sample sizes from the first few iterations of the root-finding algorithm. This categorization limits the number of points in $[0, 1]^2$ that belong to more than one of the seven categories for clearer visualization. We emphasize that these categories are not used in our power curve approximation method. Instead, they illustrate the targeted nature of how we explore the approximate sampling distribution of posterior probabilities for each sample size n considered.

We now assess the impact of using Sobol' sequences with our power curve approximation method. To do so, we approximated 1000 power curves for this Bernoulli example using Algorithm 3.4 with sequences from a pseudorandom number generator of length $m = 1024$. As before, only the method for posterior mapping in Algorithm 3.2 was considered. We then repeated this process using Algorithm 3.4 with Sobol' sequences of length $m =$

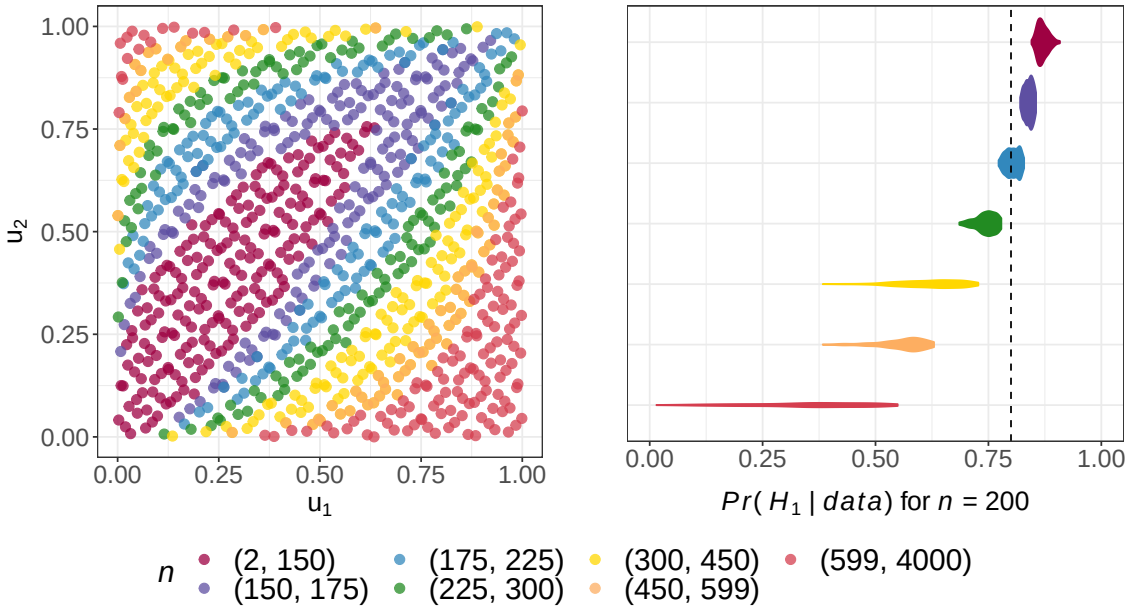


Figure 3.2: Left: Visualization of which points in $[0,1]^2$ are used to explore at least one n value in the various sample size ranges via the root-finding algorithm. Right: Violin plots for segments of the sampling distribution of posterior probabilities when $n = 200$; the vertical line is at $\gamma = 0.8$.

1024. We used the 1000 power curves corresponding to each sequence type (Sobol' and pseudorandom) to estimate power for the following sample sizes: $n = \{80, 160, ..., 2000\}$. For each sample size and sequence type, we obtained a 95% confidence interval for power using the percentile bootstrap method (Efron, 1982). We then created centered confidence intervals by subtracting the mean of the 1000 power estimates from each confidence interval endpoint. Figure 3.3 depicts these results for the sample sizes n and two sequence types considered.

Figure 3.3 illustrates that the Sobol' sequence gives rise to much more precise power estimates than pseudorandom sequences – particularly for sample sizes where power is not near 0 or 1. We repeated the process detailed in the previous paragraph to generate 1000 power curves via Algorithm 3.4 with pseudorandom sequences of length $m = 10^4$. The power estimates obtained using Sobol' sequences with length $m = 1024$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 10^4$. Similar results were observed for more extensive numerical studies as detailed in the remainder of this chapter. Using Sobol' sequences therefore allows us to estimate power with the same

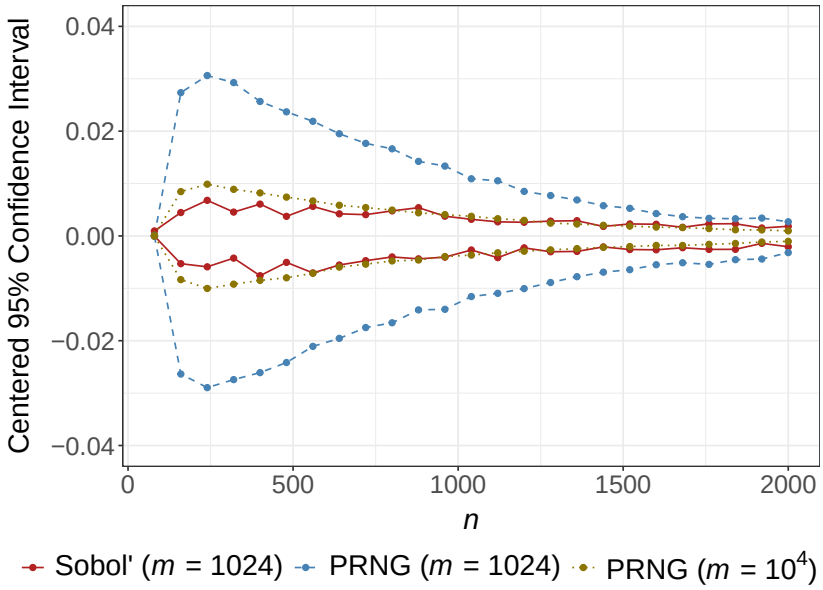


Figure 3.3: Endpoints of the centered 95% confidence intervals for power obtained with Sobol' and pseudorandom (PRNG) sequences of various lengths.

precision using approximately an order of magnitude fewer points.

To conclude this subsection, we concretely detail the magnitude of the gains in computational efficiency that are attributable to the use of (i) sampling distribution segments, (ii) Sobol' sequences, and (iii) both. The following runtimes for this Bernoulli example were measured on a standard laptop without parallelization. When using Algorithm 3.4 with Sobol' sequences of length $m = 1024$, it took just under 0.3 seconds to approximate the power curve. The 0.99-quantile of this power curve is roughly $n = 1620$. It took about 25 seconds to construct a power curve by obtaining power estimates with a single Sobol' sequence ($m = 1024$) at $n = \{2, 3, \ldots, 1620\}$. For this example, using sampling distribution segments to approximate the power curve is roughly 83 times more computationally efficient than considering entire sampling distributions. That is, we approximate the power curve $(83 - 1) \times 100 = 8200\%$ faster.

When using Algorithm 3.4 with pseudorandom sequences of length $m = 10^4$ as informed by Figure 3.3, it took roughly 2.6 seconds to approximate the power curve. These pseudorandom sequences are 9.77 times longer than the Sobol' sequences considered in the previous paragraph. Yet, the use of Sobol' sequences with this example only reduces the runtime by a factor of roughly 9 (or by roughly 800%) since there is computational overhead

associated with choosing the initial sample size n_0 in Line 2 of Algorithm 3.4. Moreover, it took just over 4 minutes to construct a power curve by obtaining power estimates with a single pseudorandom sequence ($m = 10^4$) at $n = \{2, 3, \ldots, 1620\}$. We therefore reduced the runtime by a factor of 800 when combining the use of sampling distribution segments and Sobol' sequences.

We emphasize that the gains in computational efficiency detailed above are specific to the example in this subsection. In general, the extent of the computational savings depends on the statistical models, design inputs, and the magnitude of the sample sizes that correspond to high study power. Furthermore, this discussion surrounding efficiency gains did not account for the computational savings that arise from using analytical posterior approximation instead of computational approximation methods. Those computational savings are discussed for statistical models that do not have conjugate priors in our numerical studies in Section 3.6.

3.6 Numerical Studies

3.6.1 Power Curve Approximation with the Gamma Distribution

We now compare the performance of our power curve approximation procedure across several scenarios. For each scenario, we specify design values for the gamma tail probability example from Section 3.3. Because the ENIGH survey is conducted biennially, we choose design values for both gamma distributions using data from the ENIGH 2018 survey (INEGI, 2019). We repeat the process detailed in Section 3.3 to create a similar data set of 2018 quarterly food expenditure per person. We adjust each expenditure to account for inflation, compounding 2% annually, between 2018 and 2020. We find the posterior means for the gamma shape and rate parameters to be $\bar{\alpha}_1 = 2.11$ and $\bar{\lambda}_1 = 0.69$ for the female provider group and $\bar{\alpha}_2 = 2.43$ and $\bar{\lambda}_2 = 0.79$ for the male provider group. These posterior means comprise the design values $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$. After accounting for inflation, the 2018 estimate for the median quarterly food expenditure per person in upper income households is 4.29 (MXN \$1000). For the purposes of sample size determination, we use $\kappa_0 = 4.29$ as the threshold for the gamma tail probabilities.

The scenarios we consider are based on two sets of prior distributions. For the first set, we specify uninformative GAMMA$(2, 0.25)$ priors for the gamma parameters α_j and λ_j for group $j = 1, 2$. To choose the second set of priors, we reconsider the approximately gamma distributed posteriors used to obtain design values for α_1, λ_1, α_2, and λ_2. To

incorporate prior information, we consider gamma distributions that have the same modes with variances that are larger by a factor of 10. In comparison to the GAMMA$(2, 0.25)$ prior, these distributions are quite informative. These distributions – which we use as the set of informative priors – are GAMMA$(34.23, 15.85)$ for α_1, GAMMA$(27.20, 38.15)$ for λ_1, GAMMA$(105.31, 42.96)$ for α_2, and GAMMA$(85.49, 106.58)$ for λ_2.

For each prior specification, we first consider the quality of power curve estimation for analyses with posterior probabilities. We consider three $((\delta_L, \delta_U), \gamma, 1-\beta)$ combinations: $\{a, b, c\} = \{((1.25^{-1}, 1.25), 0.5, 0.6),\ ((1.3^{-1}, \infty), 0.9, 0.7),\ ((1.15^{-1}, 1.15), 0.8, 0.8)\}$. The first combination explores moderate sample sizes. The second combination considers a one-sided noninferiority hypothesis for θ_1.[2] The third combination explores larger sample sizes. This gives rise to six settings, each consisting of a prior specification (Setting 1 = uninformative, Setting 2 = informative) and $((\delta_L, \delta_U), \gamma, 1-\beta)$ combination.

For each setting, we generated 100 power curves using Algorithm 3.4 with $\zeta = \{$Alg. 3.1, Alg. 3.2$\}$. We do not consider Algorithm 3.3 for this example because the gamma model belongs to the exponential family. We used the following transformations to improve the quality of the normal approximations for moderate n: $\theta = \log(\theta_1) - \log(\theta_2)$ and $\boldsymbol{\eta}_j = (\log(\alpha_j), \log(\lambda_j))$ for $j = 1, 2$. We then selected an appropriate array of sample sizes n. For each value of n, we generated 10000 samples of that size from $f(y; \boldsymbol{\eta}_{1,0})$ and $f(y; \boldsymbol{\eta}_{2,0})$. We approximated the corresponding posterior of $\theta = \theta_1/\theta_2$ using MCMC methods and determined whether $100 \times \gamma\%$ of the posterior was contained within (δ_L, δ_U). For each n explored, we computed the proportion of the 10000 samples in which this occurred to approximate the power curve based on entire sampling distributions. Figure 3.4 depicts these results.

For all settings considered, the alignment between the blue and red curves in Figure 3.4 indicates that the power curves generated by our method with Algorithm 3.2 perform well. Considering sampling distribution segments via the root-finding algorithm has not led to performance issues – even with moderate sample sizes. In total, we approximated 2800 power curves for the gamma tail probability example in this section and Appendix B.3. We did not need to reinitialize the root-finding algorithm in Lines 9 to 15 of Algorithm 3.4 for *any* of the 2.867×10^6 points used to generate these 2800 curves. Moreover, we considered the precision of power estimation with Sobol' and pseudorandom sequences for setting 2a with this gamma example using the process to create Figure 3.3. The results from that numerical study suggested the power estimates obtained using Sobol' sequences with length

[2] We do not use the interval $(\delta_L, \delta_U) = (1, \infty)$ corresponding to the superiority of θ_1 from Section 3.3. The design values are such that $\theta_{1,0} = 0.174 \approx \theta_{2,0} = 0.168$, so we would require an impractically large sample to support the hypothesis $H_0 : \theta_1/\theta_2 \in (1, \infty)$.

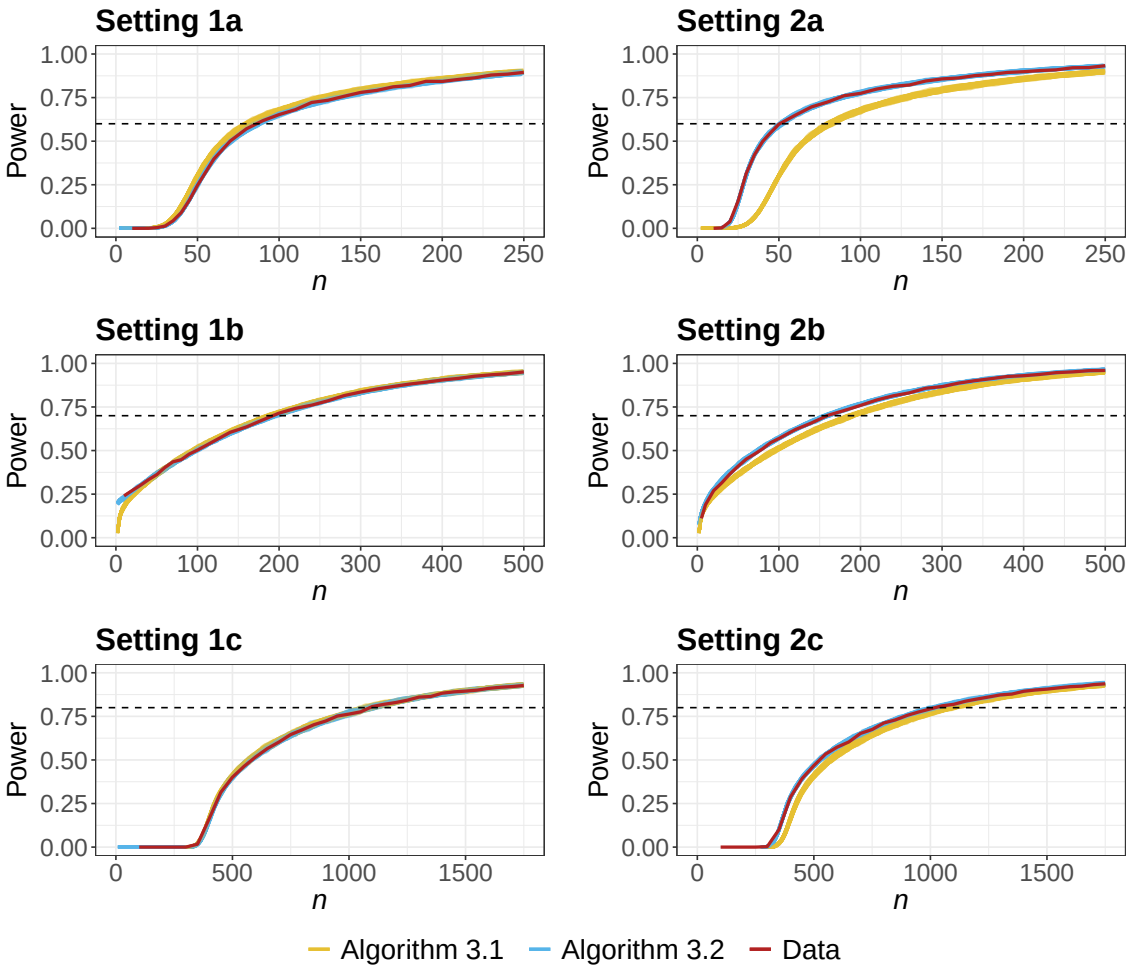


Figure 3.4: 100 power curves obtained via Algorithms 3.1 (yellow) and 3.2 (blue), power curve estimated via simulated data (red), and target power $1 - \beta$ (dotted line) for each setting with hypothesis tests facilitated via posterior probabilities.

$m = 1024$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 10^4$.

The power curves obtained using our method with Algorithm 3.1 require large sample sizes n or uninformative priors to yield good performance. This is evident as the yellow curves do not approximate the red curves for the settings with informative priors. Even in Setting 1a with uninformative priors and moderate sample sizes, the yellow power curves are noticeably shifted to the left. The bias is a result of the approximation in (3.5) not

accounting for the priors and not the root-finding algorithm. Algorithm 3.2 also performs better for smaller sample sizes in Setting 1b as the sample sizes corresponding to low power are too small for the BvM theorem to be invoked. For Setting 1c, neither the blue nor yellow power curves differ substantially from the red curve. This is a direct consequence of the BvM theorem.

Each yellow power curve in Figure 3.4 was estimated in 2 to 3 seconds without parallelization, whereas each blue curve took roughly 5 seconds to approximate without parallelization. Each red curve took between 2 and 4 hours to estimate using heavy parallelization with 72 cores. This longer runtime for the red curves even takes into account not estimating power at every sample size n and an efficient implementation of MCMC methods for the gamma distribution that is mentioned in Section 6.2.4. The blue curves take slightly longer to estimate than the yellow ones because we must find the posterior modes $\tilde{\boldsymbol{\eta}}_{1,n}$ and $\tilde{\boldsymbol{\eta}}_{2,n}$ using optimization methods. We therefore recommend using Algorithm 3.4 with the posterior approximation method from Algorithm 3.2 whenever possible since this method accounts for the prior distributions without a substantial increase in runtime. With the same computational resources for this example, we can approximate a handful of posteriors using standard computational methods. This is not sufficient to produce even a crude power estimate for a single sample size n on the red power curve.

Because power curve approximation for hypothesis tests with Bayes factors just requires that we choose γ to align with the right side of (3.2), we consider the performance of our method for such analyses in Appendix B.3.1. We now reconsider Settings 1a and 2a with analyses facilitated via equal-tailed credible intervals. We choose $\alpha = 1 - \gamma = 0.4$ for this analysis. We again implemented Algorithm 3.4 with $\zeta = \{\text{Alg. } 3.1, \text{Alg. } 3.2\}$ to obtain 100 power curves with each method. When approximating the power curve by simulating data to estimate entire sampling distributions, we computed power as the proportion of simulation repetitions in which $100\times(1-\alpha/2)\%$ of the posterior for $\theta = \theta_1/\theta_2$ was contained within each of the following intervals: (δ_L, ∞) and $(-\infty, \delta_U)$. These results for Settings 1a and 2a are visualized in Figure 3.5.

For analyses with credible intervals, we draw similar conclusions about the performance of our approach with Algorithms 3.1 and 3.2 as in Figure 3.4. Each blue curve in Figure 3.5 took roughly 7 seconds to estimate since we must examine the intervals corresponding to two one-sided tests with posterior probabilities as discussed in Section 3.5.3. Because sufficient statistics can be readily computed for this example, we did not consider the performance of our approach to power analysis with the approximation method in Algorithm 3.3. We consider the performance of that approach in Section 3.6.2.

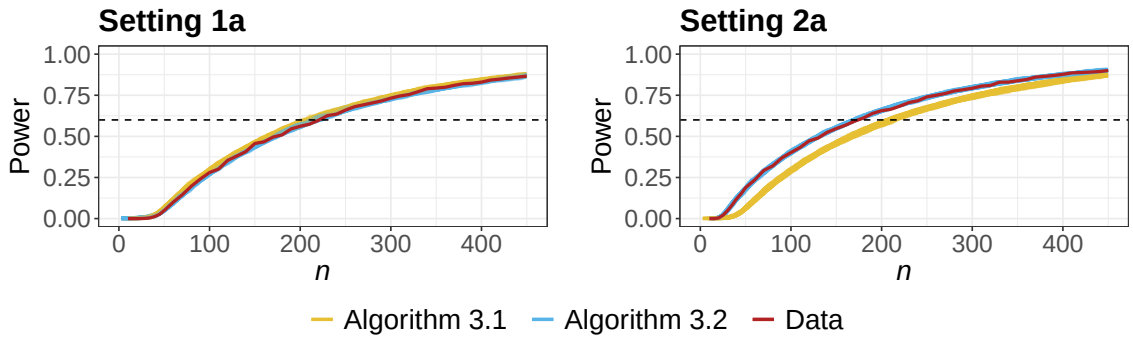


Figure 3.5: 100 power curves obtained via Algorithms 3.1 (yellow) and 3.2 (blue), power curve estimated via simulated data (red), and target power $1-\beta$ (dotted line) for Settings 1a and 2a with hypothesis tests facilitated via credible intervals.

3.6.2 Power Curve Approximation with the Weibull Distribution

To further explore the performance of our power curve approximation procedure, we reconsider the food expenditure example with Weibull distributions. We find design values for the Weibull distributions using the data from the ENIGH 2018 survey processed in Section 3.6.1. We find the posterior means for the Weibull shape (ν_j) and scale (ι_j) parameters to be $\bar{\nu}_1 = 1.41$ and $\bar{\iota}_1 = 3.39$ for the female provider group and $\bar{\nu}_2 = 1.49$ and $\bar{\iota}_2 = 3.42$ for the male provider group, where GAMMA$(2, 1)$ priors were assigned to each parameter. These posterior means comprise the new design values $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$. As in Section 3.6.1, a threshold of $\kappa_0 = 4.29$ defines the Weibull tail probabilities.

We again consider two sets of prior distributions. For the first set, we specify uninformative GAMMA$(2, 1)$ priors for the Weibull parameters ν_j and ι_j for group $j = 1, 2$. To choose the second set of priors, we reconsider the approximately gamma distributed posteriors used to obtain design values for ν_1, ι_1, ν_2, and ι_2. To incorporate prior information, we consider gamma distributions that have the same modes with variances that are larger by a factor of 100. These distributions prompt the following informative priors: GAMMA$(12.73, 8.28)$ for ν_1, GAMMA$(11.81, 3.20)$ for ι_1, GAMMA$(38.35, 25.09)$ for ν_2, and GAMMA$(37.91, 10.79)$ for ι_2.

For each prior specification, we consider power curve estimation for analyses facilitated via posterior probabilities with Settings 1a and 2a from Section 3.6.1, where $((\delta_L, \delta_U), \gamma, 1-\beta) = ((1.25^{-1}, 1.25), 0.5, 0.6)$. For each setting, we generated 100 power curves using Algorithm 3.4 with $\zeta = \{\text{Alg. } 3.1, \text{Alg. } 3.3\}$. We used the following transformations to improve the quality of the normal approximations: $\theta = \log(\theta_1) - \log(\theta_2)$ and $\boldsymbol{\eta}_j = (\log(\nu_j), \log(\iota_j))$

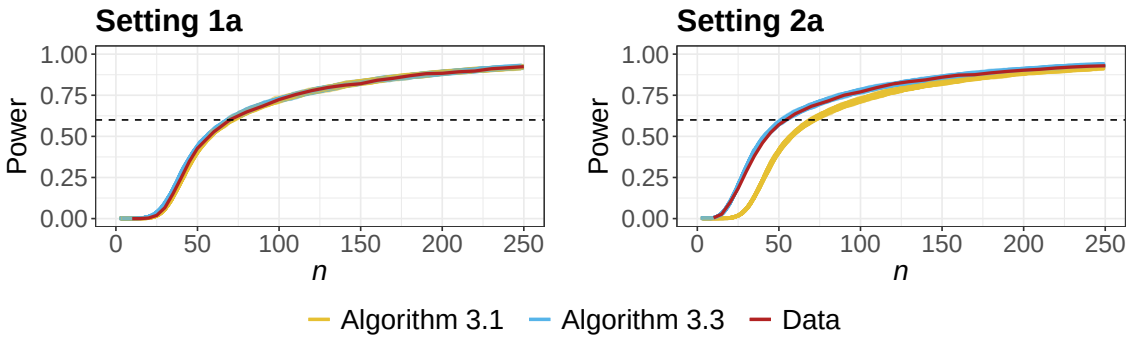


Figure 3.6: 100 power curves obtained via Algorithms 3.1 (yellow) and 3.3 (blue), power curve estimated via simulated data (red), and target power $1-\beta$ (dotted line) for Settings 1a and 2a with hypothesis tests using the Weibull distribution.

for $j = 1, 2$. As in Section 3.6.1, we also estimated the power curves by generating data from the design distributions and approximating the posterior of $\theta = \theta_1/\theta_2$ via MCMC methods. Figure 3.6 depicts these results. Following the process to create Figure 3.3, we also considered the precision of power estimation with Sobol' and pseudorandom sequences for setting 1a with this Weibull example. The results from that numerical study suggested the power estimates obtained using Sobol' sequences with length $m = 1024$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 6 \times 10^3$.

For Settings 1a and 2a in Figure 3.6, we draw similar conclusions for the blue and yellow power curves as in Figure 3.4 with the gamma model. Algorithm 3.3 therefore yields suitable performance for this example when low-dimensional sufficient statistics cannot be calculated. Each yellow and blue power curve in Figure 3.6 was estimated in less than five seconds without parallelization. For this example with the Weibull model, we cannot approximate a single posterior of θ using the same computing resources. Each red power curve took over 12 hours to estimate using parallelization with 72 cores. More computational resources are required to approximate the red power curves for the Weibull distribution than for the gamma distribution because implementing MCMC methods with Weibull data is more costly. While our power curve approximation method does not require parallel computing for fast performance, all yellow and blue power curves in our numerical studies could be estimated in a second or two if parallelized on a standard laptop with four cores. As such, our method for power curve approximation allows users to quickly explore potential designs for their study in real time, expediting communication between stakeholders of the study.

3.7 Discussion

In this chapter, we developed a framework for fast power curve approximation with hypothesis tests facilitated via posterior probabilities, Bayes factors, and credible intervals. The computational efficiency of this framework stems from exploring segments of the sampling distribution of posterior probabilities for each sample size considered when the conditions for the BvM theorem are satisfied. The numerical studies conducted show that our fast method yields suitable power curve approximation for moderate and large sample sizes. While this method is not appropriate for small sample sizes, it informs practitioners when their required sample sizes are small. More traditional simulation-based design methods can be used in these scenarios since they are often less cumbersome to implement with small sample sizes.

In Chapter 5, we make this framework more flexible by extending it to the predictive approach for choosing the sampling distribution of $\boldsymbol{Y}^{(n)}$. This extension prevents us from directly applying the BvM theorem. However, the results from the BvM theorem (where we treat a draw from the design prior as the fixed parameter $\boldsymbol{\eta}_{j,0}$) are still combined with targeted exploration of sampling distributions to yield fast sample size recommendations. This framework is also extended to accommodate multiple study objectives. For instance, we may require an (n, γ) combination that both satisfies a power criterion and bounds a type I error or false discovery rate. If the critical value γ were chosen algorithmically to bound the type I error rate, we would not be able to use root-finding algorithms as in this chapter to select sampling distribution segments. As such, Chapter 5 considers alternative methods to select these segments.

Moreover, the framework presented in the main portion of this chapter does not support imbalanced two-group sample size determination (i.e., where $n_2 = qn_1$ for some constant $q > 0$). It may be inefficient or impractical to force $q = 1$ when prior information for one group is much more precise, when it is more difficult to sample from one of the groups, or in scenarios where one treatment is much riskier. In Appendix B.3.2, we extend this framework to settings where practitioners specify this constant q. Imbalanced sample size determination for Bayesian hypothesis tests is also considered more thoroughly in Chapter 5.

Chapter 4

Posterior Ramifications of Prior Dependence Structures

4.1 Preamble

In Chapter 3, we developed methods to design Bayesian hypothesis tests facilitated via posterior probabilities, Bayes factors, and credible intervals. For design purposes, these methods assume data are generated from statistical models with known, fixed parameter values. This simplifying assumption allowed us to bypass specifying a design prior for the relevant parameter(s). However, methods for Bayesian analysis incorporate uncertainty about the parameter values used to generate the data. To promote better coherence between experimental design and analysis, we develop methods for study design that incorporate such uncertainty in Chapter 5. This development requires us to specify nondegenerate design priors that yield prior predictive distributions for the to-be-observed data in pre-experimental settings.

Specifying such priors is not a trivial task, particularly when the parameter is multivariate because we must consider the dependence structure between its components. The extent of the effort required to specify these priors suggests that the simpler design procedures proposed in Chapter 3 may be ideal for certain practitioners. In this thesis chapter only, we use $\boldsymbol{\theta} \in \mathbb{R}^d$ instead of $\boldsymbol{\eta}$ to denote the (possibly multivariate) parameter(s) for which we assign a prior distribution – which may not coincide with the characteristic(s) of interest. This notation promotes better alignment with the standard nomenclature used in the prior elicitation literature. The notation for the posterior mode also differs in this chapter as detailed later.

In this chapter, we explore recent advances to elicit multivariate prior distributions using copula models. Under broad conditions, we demonstrate that the posterior of $\boldsymbol{\theta}$ cannot retain many of these flexible prior dependence structures in large-sample settings. We also overview several objectives for prior specification to help practitioners determine whether the inability to retain the prior dependence structure presents practical issues for their posterior objectives. Our resulting recommendations for prior dependence specification are generally useful for Bayesian analyses, and they will be referenced when reconsidering design with sampling distribution segments in Chapter 5.

4.2 Background

4.2.1 Overview of Prior Elicitation

Statistical methods leverage data to infer properties about an unobservable, possibly multivariate parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$. Bayesian methods for statistical inference (see e.g., Gelman et al. (2020)) require the specification of a prior distribution for the parameter $\boldsymbol{\theta}$, denoted by $p(\boldsymbol{\theta})$ in this chapter. This distribution characterizes the beliefs about $\boldsymbol{\theta}$ prior to observing any data. For a particular statistical model, $L(\boldsymbol{\theta}; \boldsymbol{y})$ denotes the relevant likelihood function for the parameter $\boldsymbol{\theta}$ with respect to the observed data, now referred to as the vector or matrix $\boldsymbol{y}$. In the Bayesian paradigm, inference is facilitated via the posterior distribution of $\boldsymbol{\theta}$, denoted by $p(\boldsymbol{\theta}|\boldsymbol{y})$ in this chapter. By Bayes' Theorem, we have that

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{\theta}; \boldsymbol{y})p(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}; \boldsymbol{y})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto L(\boldsymbol{\theta}; \boldsymbol{y})p(\boldsymbol{\theta}). \tag{4.1}$$

To implement fully Bayesian analyses, prior distributions must be specified before observing data. This is in contrast to empirical Bayes methods that set the parameters for the prior distributions to their most likely values given the observed data (Casella, 1985; Carlin and Louis, 2000). In the absence of prior information, uninformative or diffuse priors are often used. When relevant prior information from subject matter experts or previous statistical analyses is available, it rarely takes the form of quantifiable prior distributions on the unobservable parameters of a statistical model. Prior elicitation procedures are used to transfigure prior information into quantifiable prior distributions.

Winkler (1967) conducted some of the initial work on prior elicitation, citing the siloed nature of prior specification and posterior analysis. Most previous Bayesian research had investigated how to leverage sample information from the likelihood function to obtain the

posterior from the prior distribution, which was assumed to have been already assessed. Winkler (1967) explored elicitation methods for Bernoulli processes, two of which involved directly asking questions about the prior CDF or probability distribution function (PDF) for the Bernoulli parameter. These elicitation methods were for univariate priors, yet Winkler (1967) acknowledged that the assessment of prior distributions in multivariate contexts was an important and nontrivial problem.

Decades later, Chaloner (1996) overviewed methods for subjective prior specification with an emphasis on methods for clinical trials. In this context, Freedman and Spiegelhalter (1983) elicited a prior on the difference in cancer recurrence probability between treatment groups by asking experts about the mode and likely lower and upper bounds for this difference. This prior was not combined with observed data and instead used for design purposes to choose the number of interim analyses in a sequential trial. Other contributions in clinical settings advocated for soliciting prior beliefs from several experts to form a community of prior distributions and basing inference on a consensus of posterior conclusions (Kadane, 1986; Chaloner et al., 1993). To reduce cognitive and computational complexity, most elicitation methods overviewed by Chaloner (1996) relied on parametric assumptions and solicited information about potentially observable conduits for the unobservable parameter $\boldsymbol{\theta}$ that are easier to conceptualize. For instance, priors for Bayesian regression model coefficients were elicited using information about quantiles of predictive distributions (Kadane et al., 1980) and survival probabilities (Chaloner et al., 1993).

Additional reviews of prior elicitation methods have since been conducted. Garthwaite et al. (2005) divided the elicitation process into four stages: setup, elicitation, fitting, and adequacy. These stages involve preparing the expert for elicitation, eliciting distributional summaries of expert knowledge, fitting a probability distribution to these summaries, and assessing the adequacy of elicitation. O'Hagan et al. (2006) focused on elicitation of subjective probabilities and noted that experts tend to characterize events as impossible instead of assigning them small probabilities. Because this gives rise to hard boundaries in the support of the elicited priors, O'Hagan et al. (2006) overviewed methods to correct for expert overconfidence. Johnson et al. (2010) conducted a systematic review of prior elicitation methods with an emphasis on the feasibility of the process in terms of the required time, cost, personnel, and equipment. They found that although expert fatigue and lack of understanding compromise the reliability of elicitation procedures (Winkler, 1971; Garthwaite et al., 2005), the reviewed methods were rarely formally evaluated on their feasibility.

Recent contributions have aimed to reduce expert fatigue and improve understanding by developing iterative elicitation procedures that provide experts with instant feedback via a graphical interface (Jones and Johnson, 2014; Casement and Kahle, 2018; Williams et al., 2021; Casement and Kahle, 2023). Prior elicitation procedures have also been developed

for more complex settings – including methods for rank analysis (Crispino and Antoniano-Villalobos, 2023), nonparametric models (Seo and Kim, 2022), sequential analysis (Santos and Costa, 2019), power priors (Ye et al., 2022), and mixture models (Fúquene et al., 2019; Feroze and Aslam, 2021). One such area of research involves using copulas to accommodate more flexible dependence structures when eliciting multivariate priors (Elfadaly and Garthwaite, 2017; Wilson, 2018; Wilson et al., 2021). These contributions are novel given that the dependence structure is an afterthought in many recent prior elicitation methods. In particular, many recent methods consider elicitation for univariate parameters (Casement and Kahle, 2018, 2023), fix the dependence structure to leverage conjugacy (Santos and Costa, 2019; Srivastava et al., 2019), or assume that the components of $\boldsymbol{\theta}$ are independent a priori (Garthwaite et al., 2013; Jones and Johnson, 2014; Fúquene et al., 2019; Seo and Kim, 2022). Prominent recent advances with copula-based priors are overviewed in Section 4.2.3. Copula-based priors are the focus of this chapter, and a primer on copulas is provided in Section 4.2.2.

4.2.2 Background on Copula Models

The behaviour of random variables $\boldsymbol{X} = \{X_j\}_{j=1}^d$ is often characterized by their joint distribution function $H(\boldsymbol{x})$. Each component of $\boldsymbol{X}$ also has a marginal distribution function $F_j(x_j) = Pr(X_j \leq x_j),\ j = 1, \ldots, d$. Copulas flexibly allow for the dependence structure of $\boldsymbol{X}$ to be considered separately from its marginals when eliciting multivariate distributions. Let $U_1, U_2, \ldots, U_d$ be uniformly-distributed random variables over the unit interval $[0, 1]$. The distribution function

$$C(u_1, \ldots, u_d) = Pr(U_1 \leq u_1, \ldots, U_d \leq u_d)$$

is such that $C : [0, 1]^d \to [0, 1]$ is a copula (Nelsen, 2006). Sklar's theorem (Sklar, 1959; Schweizer and Sklar, 2011) explicates the relationship between the copula C, the multivariate joint distribution function $H(\boldsymbol{x})$, and the univariate marginal CDFs $F_j(x_j)$ for $j = 1, \ldots, d$:

$$H(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)),$$

where $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$. Copulas are therefore incorporated into multivariate distributions even if they are not explicitly defined. If $F_1, \ldots, F_d$ are continuous, the copula C is unique.

A copula C can be represented as the sum of its absolutely continuous component A_C

and singular component S_C (Nelsen, 2006). For $\boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d$,

$$A_C(\boldsymbol{u}) = \int_0^{u_1} \cdots \int_0^{u_d} \frac{\partial^d}{\partial t_1 \cdots \partial t_d} C(t_1, \ldots, t_d) dt_d \cdots dt_1,$$

and $S_C(\boldsymbol{u}) = C(\boldsymbol{u}) - A_C(\boldsymbol{u})$. If $C = A_C$ on $[0,1]^d$, the copula C is absolutely continuous and admits a density function

$$c(\boldsymbol{u}) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(u_1, \ldots, u_d).$$

Moreover, if the support of C is $[0,1]^d$, the copula is deemed to have full support (Nelsen, 2006). All copulas considered in this chapter adhere to this definition.

Copulas can be defined using common probability distributions. For instance, the Gaussian copula (Clemen and Reilly, 1999) with correlation matrix $\boldsymbol{R}$ is defined such that

$$C^{Ga}_{\boldsymbol{R}}(\boldsymbol{u}) = \Phi_{\boldsymbol{R}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)),$$

where Φ is the CDF of the $\mathcal{N}(0,1)$ distribution and $\Phi_{\boldsymbol{R}} : \mathbb{R}^d \to [0,1]$ is the CDF of the d-dimensional $\mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$ distribution with correlation and covariance matrix $\boldsymbol{R}$. The t-copula (Demarta and McNeil, 2005) can be similarly defined given a multivariate t-distribution with correlation matrix $\boldsymbol{R}$ and degrees of freedom ν. In contrast, Archimedean copulas are a commonly used class of copulas that are efficiently parameterized via generator functions (Nelsen, 2006). This chapter focuses on parametric copula models, but copulas can also be leveraged in a nonparametric framework (Wong and Ma, 2010; Wu et al., 2015; Ning and Shephard, 2018; Barone and Dalla Valle, 2023).

For $d = 2$ dimensions, Figure 4.1 visualizes samples from two copulas. In the left plot, the blue points are generated from an Archimedean Clayton copula that characterizes positive dependence with dependence parameter $\phi \geq 0$, set to $\phi = 3$. The extent of the dependence between two random quantities is constrained by the Fréchet-Hoeffding bounds (Nelsen, 2006). The dotted line in the left plot of Figure 4.1 is the upper Fréchet-Hoeffding bound. This bound characterizes dependence for comonotonic variables. In the right plot, the blue points are generated from a Gaussian copula parameterized by Pearson's $\rho = -0.8$. The dotted line in the right plot is the lower Fréchet-Hoeffding bound. This bound characterizes dependence for countermonotonic variables. As the strength of the positive (negative) dependence between two variables increases, their (u_1, u_2) combinations tend to cluster around the upper (lower) Fréchet-Hoeffding bound. The dependence structure between two independent random variables is characterized by the independence copula:

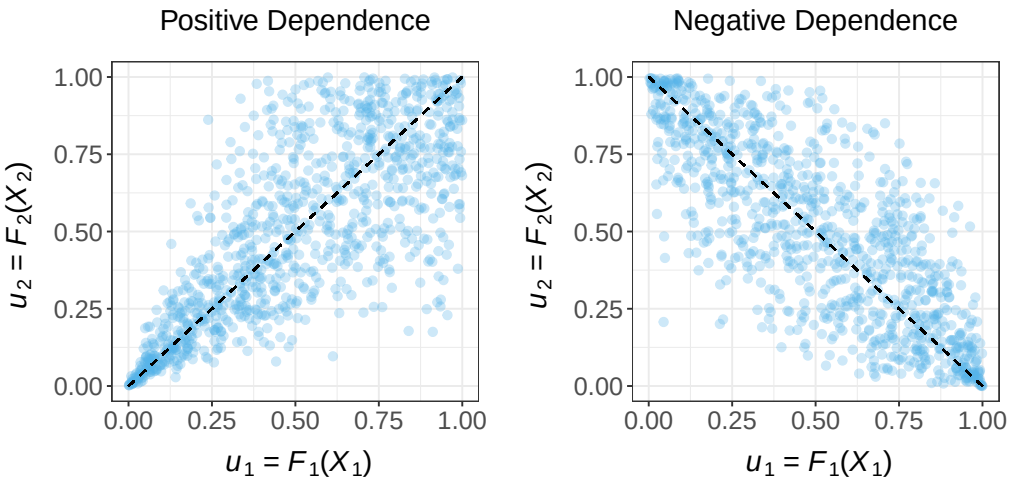


Figure 4.1: Samples of 1000 points from a Clayton copula with $\phi = 3$ (left) and a Gaussian copula with Pearson's $\rho = -0.8$ (right). The upper and lower Fréchet-Hoeffding bounds are given by the dotted lines.

$c(\boldsymbol{u}) = 1$ for $\boldsymbol{u} \in [0,1]^2$. In two dimensions, a sample from this copula approximates random scatter over $[0,1]^2$.

The upper Fréchet-Hoeffding bound for positive dependence generalizes to settings with more than two dimensions. However, negative dependence in higher dimensions is more complicated since many random variables cannot exhibit strong mutual negative dependence. In addition to characterizing the strength and direction of dependence between random variables, copula models also account for symmetry and tail dependence. For instance, the Clayton copula accounts for lower tail dependence since the points in the left plot of Figure 4.1 are tightly clustered in the bottom left corner.

4.2.3 Recent Developments with Copula-Based Priors

Copulas can be incorporated into the prior elicitation process as illustrated in recent developments for the multinomial model. Elfadaly and Garthwaite (2017) proposed one such method to elicit Gaussian copula prior distributions. The standard multinomial model assumes that data $y_i \in \{1, 2, \ldots, w\}$, $i = 1, \ldots, n$ are collected independently and that the outcome v occurs with probability $0 < p_v < 1$ for $v = 1, \ldots, w$ such that $\sum_{v=1}^{w} p_v = 1$. The multinomial model is parameterized by $\boldsymbol{p} = (p_1, \ldots, p_w)$. However, Elfadaly and Garth-

waite (2017) did not directly assign a Gaussian copula to the multinomial probabilities since that approach would not enforce the unit-sum constraint. Instead, they defined new variables $Z_1, \ldots, Z_w$ such that

$$Z_1 = p_1, \quad Z_v = \frac{p_v}{1 - \sum_{t=1}^{v-1} p_t} \text{ for } v = 2, \ldots, w-1, \quad \text{and} \quad Z_w = 1. \tag{4.2}$$

The corresponding inverse transformations are given by

$$p_1 = Z_1 \quad \text{and} \quad p_v = Z_v \prod_{t=1}^{v-1}(1 - Z_t) \text{ for } v = 2, \ldots, w. \tag{4.3}$$

The variable Z_v represents the probability that an observation is assigned to category v given that it has not been assigned to categories $1, \ldots, v-1$. Elfadaly and Garthwaite (2017) assigned marginal BETA(α_v, β_v) priors to Z_v, $v = 1, \ldots, w-1$. A joint prior for $\boldsymbol{\theta} = \boldsymbol{Z}_{w-1} = (Z_1, \ldots, Z_{w-1})$ that satisfies the unit-sum constraint was created by joining the marginal beta priors with a Gaussian copula. If $Z_1, \ldots, Z_{w-1}$ are independent, $\boldsymbol{p}_w$ follows a generalized Dirichlet distribution (Connor and Mosimann, 1969). In this scenario, Gaussian copulas were leveraged to construct priors that were more flexible than standard alternatives.

Elfadaly and Garthwaite (2017) constructed marginal beta distributions for $Z_v, v = 1, \ldots, w-1$ by soliciting estimates for the quartiles of each variable. For each variable, the three quartile estimates were reconciled into a two-parameter beta distribution using least-squares optimization. They formed the correlation matrix $\boldsymbol{R}$ for the Gaussian copula by soliciting further estimates. For $v = 2, \ldots, w-1$, experts were asked to update their estimate for the median of p_v under the assumption that the median of p_{v-1} was equal to the lower quartile specified in the previous step. These additional estimates were used in conjunction with a method from Kadane et al. (1980) to ensure $\boldsymbol{R}$ was positive definite.

If the prior must admit a density function, only absolutely continuous copulas should be considered during the elicitation process. Unless certain combinations of the variables in $\boldsymbol{\theta}$ are invalid, the candidate copulas should have full support so as to not inadvertently restrict the domain of the parameter space $\boldsymbol{\Theta}$ a priori. When $\boldsymbol{R}$ has full rank, the Gaussian copula is absolutely continuous with full support.

Wilson (2018) extended this method for use with vine copulas (Joe, 1996; Bedford and Cooke, 2002; Joe and Kurowicka, 2011), using a similar process as Elfadaly and Garthwaite (2017) to elicit marginal beta distributions for $Z_v, v = 1, \ldots, w-1$. This extended method leveraged D-vines (Kurowicka and Cooke, 2005) to incorporate more flexibility into the

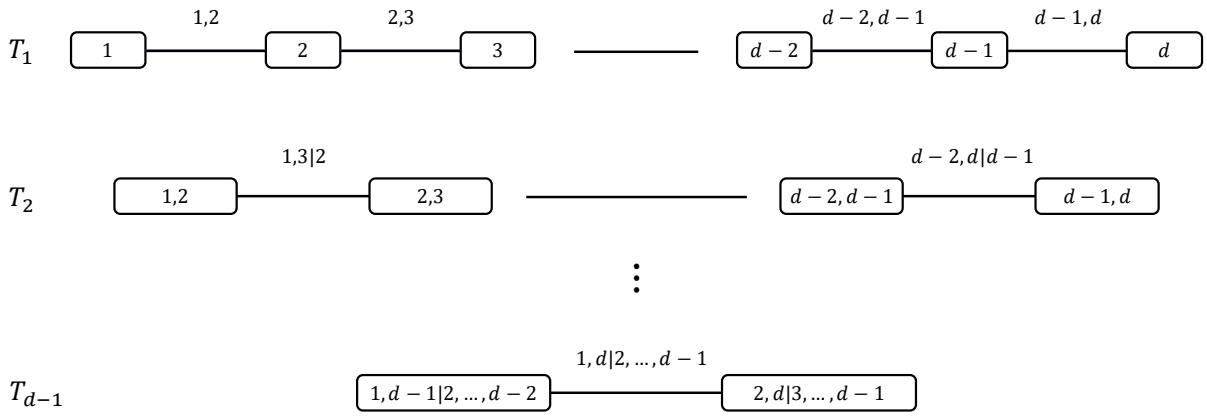


Figure 4.2: The structure of a general D-Vine on d variables.

prior dependence structure than the Gaussian copula can accommodate. For a model with d variables, D-vines utilize the graphical structure in Figure 4.2 to characterize dependence between the variables in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ using $d-1$ trees: $T_1, \ldots, T_{d-1}$. T_1 consists of a node set $N_1 = \{1, 2, \ldots, d\}$ and an edge set $E_1 = \{(1,2), (2,3), \ldots, (d-1, d)\}$, where the integers in the node and edge sets refer to indices in $\boldsymbol{\theta}$. For $j = 2, 3, \ldots, d-1$, the node set of T_j is $N_j = E_{j-1}$, and two edges in E_{j-1} are connected with an edge in T_j only if they share a common node. D-vines characterize dependence in higher-dimensional settings using (un)conditional bivariate copulas. Each edge $\boldsymbol{e}$ in the edge set $E(\mathcal{V}) = \cup_{j=1}^{d-1} E_j$ considers the dependence between two variables in $\boldsymbol{\theta}$, denoted e_1 and e_2, conditional on the variables in a potentially empty set $D_{\boldsymbol{e}} \subset \{1, 2, \ldots, d\}$.

Wilson (2018) proposed considering Gaussian and t-copulas as candidate copulas, along with several Archimedean copulas that are absolutely continuous with full support. These bivariate copulas were selected by soliciting estimates for the conditional quartiles of the p_v and Z_v variables. Across all candidate copula families considered, the best-fitting copula was parameterized to minimize least squares between solicited and induced prior quantiles on the Z_v variables.

4.2.4 Contributions

This chapter provides general guidance for prior dependence specification in multivariate settings. These recommendations are topical given that recent advances in copula-based priors allow for the incorporation of unprecedented flexibility into the prior dependence

structure. While this additional flexibility could give rise to priors that more accurately characterize real-life phenomena, practitioners may be subject to choice overload when deciding between the vast number of potential dependence structures. We argue that this additional flexibility is only useful in certain contexts. Our recommendations help practitioners discard prior dependence structures that do not align with their objectives for posterior analysis – simplifying the prior specification process. These recommendations also clarify *how* the prior dependence structure can impact the posterior distribution. Our recommendations are illustrated using several models for which copula-based priors have been proposed. Unlike situations that require the specification of a community of prior distributions, we restrict our discussion to the case where a single prior is elicited.

The remainder of this chapter is structured as follows. In Section 4.3, we define general conditions under which the prior dependence structure is incompatible with that induced by the likelihood function and unable to be retained by the posterior distribution as data are observed. We formally prove this large-sample result and demonstrate that priors elicited using the methods proposed by Elfadaly and Garthwaite (2017) and Wilson (2018) are generally incompatible with standard likelihood functions. We discuss small-sample scenarios where this additional flexibility in the prior dependence structure is nevertheless useful in Section 4.4. In Section 4.5, we prove asymptotic results about the impact of the prior dependence structure on the convergence of the posterior mode, which we contrast with previous work on copula-based priors (Michimae and Emura, 2022). We then conduct numerical studies with both small and large sample sizes to contextualize these theoretical results. These simulations prompt further recommendations regarding choosing a prior dependence structure for posterior analysis. We provide concluding remarks and a discussion of extensions to this work in Section 4.6.

4.3 Retention of Prior Dependence

4.3.1 Background

In this section, we examine situations where the prior dependence structure cannot carry over into the posterior distribution as data are collected. It is unrealistic to expect the prior dependence structure for $\boldsymbol{\theta}$ to be perfectly specified, but we should be mindful of whether the prior dependence structure can be retained a posteriori before using the posterior of $\boldsymbol{\theta}$ to draw conclusions about its dependence structure. Otherwise, we may draw different conclusions about the dependence structure given small and large samples generated from the same data generation process. Drawing such conclusions in an uninformed way presents

practical issues and complicates the consideration of the posterior of a function of several parameters in $\boldsymbol{\theta}$.

The concept of chronic rejection (Libby and Pober, 2001; Vos et al., 2011) frames this section's main result. The term chronic rejection describes the process in which a transplanted organ is rejected by the recipient's immune system over a long period of time. The recipient's persistent immune response against the transplanted organ causes gradual damage. Similar to screening an organ donor, we argue that one should consider whether the prior dependence structure for $\boldsymbol{\theta}$ is compatible with that induced by the likelihood function. Such incompatibilities imply that the prior dependence structure cannot be retained a posteriori – and the prior dependence structure is hence a chronically rejected one. We emphasize that the term *rejection* as used in this chapter does not imply the formal rejection of a statistical hypothesis test. In Section 4.3.2, we define the notion of chronically rejected prior dependence structures. We illustrate this definition using simulation in Section 4.3.3. Section 4.4 discusses situations where using chronically rejected prior dependence structures may be sensible, so practitioners should consider their posterior objectives before choosing a prior dependence structure.

4.3.2 Chronically Rejected Prior Dependence Structures

Here, we define general conditions under which prior dependence structures cannot be retained by the posterior with enough data. These sufficient conditions can be readily verified prior to observing data. When planning posterior analyses, practitioners can discard prior dependence structures that satisfy these conditions for certain posterior objectives. This simplification reduces the number of potential dependence structures and hence streamlines the prior elicitation process.

We consider the limiting behaviour of the posterior distribution for $\boldsymbol{\theta}$. Because we consider this behaviour under various data generation processes, the data are random variables. Data from a random sample of size n are represented by $\boldsymbol{Y}^{(n)}$. Realizations of these samples are denoted by $\boldsymbol{y}^{(n)}$. Much of the work on limiting posteriors (see e.g., Ghosal et al. (1995), Le Cam and Yang (2000), Gao et al. (2020), or Schillings et al. (2020)) appeals to the Bernstein-von Mises theorem (van der Vaart, 1998). We let $m(\cdot|\boldsymbol{\theta})$ be the statistical model corresponding to the likelihood function in (4.1). When data $\boldsymbol{Y}^{(n)}$ are generated independently and identically from $m(\cdot|\boldsymbol{\theta}_0)$, the BvM theorem dictates that the posterior for $\boldsymbol{\theta}$ converges to the $\mathcal{N}(\boldsymbol{\theta}_0, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}/n)$ distribution in the limit of infinite data, where $\mathcal{I}(\cdot)$ is the Fisher information.

In addition to the independently and identically distributed assumption, there are three

conditions that must be satisfied to invoke the BvM theorem. The first two conditions involve the model $m(\cdot|\boldsymbol{\theta}_0)$, and they are collectively weaker than the conditions for the asymptotic normality of the maximum likelihood estimator (MLE) (Lehmann and Casella, 1998). Condition 1 ensures the model $m(\cdot|\boldsymbol{\theta}_0)$ is differentiable in quadratic mean with nonsingular $\mathcal{I}(\boldsymbol{\theta}_0)$. Condition 2 requires that there exist a sequence of uniformly consistent tests for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $H_1 : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \geq \varepsilon$ for every $\varepsilon > 0$. Condition 3 concerns the prior distribution $p(\boldsymbol{\theta})$ used to analyze the observed data. This prior must be absolutely continuous in a neighbourhood of $\boldsymbol{\theta}_0$ with $p(\boldsymbol{\theta}_0) > 0$. We consider priors defined such that

$$p(\boldsymbol{\theta}) = c(F_1(\theta_1), \ldots, F_d(\theta_d)) \times \prod_{j=1}^{d} f_j(\theta_j), \tag{4.4}$$

where $c(u_1, \ldots, u_d)$ is the copula density function for an absolutely continuous copula C and $F_j(\theta_j)$ and $f_j(\theta_j)$ are respectively the marginal prior CDF and PDF for θ_j, $j = 1, \ldots, d$.

To directly apply the BvM theorem, a single value of $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ must be selected. It may not be realistic to expect practitioners to correctly identify this value for $\boldsymbol{\theta}_0$ a priori. As such, our results incorporate uncertainty about the value of $\boldsymbol{\theta}_0$ used to generate $\boldsymbol{Y}^{(n)}$. We do so by introducing a prior $p_D(\boldsymbol{\theta})$ that defines the prior predictive distribution of $\boldsymbol{Y}^{(n)}$:

$$p(\boldsymbol{y}^{(n)}) = \int \prod_{i}^{n} m(y_i|\boldsymbol{\theta}) p_D(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{4.5}$$

In (4.5), data $\boldsymbol{Y}^{(n)}$ are generated independently and identically from $m(\cdot|\boldsymbol{\theta}_0)$ given $\boldsymbol{\theta}_0 \sim p_D(\boldsymbol{\theta})$. This data generation process allows us to compare various objectives for prior specification using repeated simulation in Section 4.4. The prior in (4.5) may or may not be the same prior as $p(\boldsymbol{\theta})$ used to analyze the observed data in (4.4), which is often called the analysis prior. To explore the limiting behaviour of the posterior when data are generated via (4.5), we note that the BvM theorem considers a special case of $p(\boldsymbol{y}^{(n)})$ in which the prior $p_D(\boldsymbol{\theta})$ is degenerate. That is, $p_D(\boldsymbol{\theta}_0) = 1$ for a particular $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, and 0 otherwise. In light of this, we emphasize that the prior that must satisfy condition 3 for the BvM theorem is the analysis prior $p(\boldsymbol{\theta})$. Theorem 4.1 generalizes results from the BvM theorem to nondegenerate priors $p_D(\boldsymbol{\theta})$ under certain conditions.

Theorem 4.1. *Let $\boldsymbol{\Theta}^*$ be the set of interior points in $\boldsymbol{\Theta}$. Suppose conditions 1, 2, and 3 for the BvM theorem are satisfied for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}^*$. Let data $\boldsymbol{Y}^{(n)}$ be generated via (4.5) such that $\boldsymbol{\theta}_0 \sim p_D(\boldsymbol{\theta})$ and $p_D(\boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \notin \boldsymbol{\Theta}^*$. The posterior dependence structure of $\boldsymbol{\theta}$ given observed $\boldsymbol{y}^{(n)} \xrightarrow{d} C^{Ga}_{\boldsymbol{R}}(\cdot)$ corresponding to the covariance matrix $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ as $n \to \infty$.*

Proof of Theorem 4.1. When the conditions for Theorem 4.1 hold, $p(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}) \xrightarrow{d} \mathcal{N}(\theta_0, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}/n)$. The posterior dependence structure of $\boldsymbol{\theta}$ is hence reasonably characterized by a Gaussian copula with correlation matrix $\boldsymbol{R}$ corresponding to $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ for large enough n. This result follows by the BvM theorem because we restrict the value of $\boldsymbol{\theta}_0$ to be contained in $\boldsymbol{\Theta}^*$; the BvM theorem is applicable when $\boldsymbol{\theta}_0$ is not a boundary point of the parameter space $\boldsymbol{\Theta}$. □

Under the conditions in Theorem 4.1, the posterior copula for $\boldsymbol{\theta}$ is approximately Gaussian for large samples. However, we may not collect nearly enough data for this copula to be Gaussian in practice. Because it is unrealistic to expect the prior dependence structure of $\boldsymbol{\theta}$ to be perfectly specified, we consider a partial characterization of this prior (and posterior) dependence structure using D-vines in Corollary 4.1. Even if the copula in (4.4) is not specified using a D-vine, such a structure can be induced. D-vines are often specified via the set of Kendall's τ (Kendall, 1938) values for each of the D-vine's bivariate copulas (Kurowicka and Cooke, 2005). Kendall's τ measures rank correlation in terms of how similar the orderings of bivariate data are when ranked by each quantity. We let $\{\tau^p_{e_1,e_2|D_{\boldsymbol{e}}}\}_{\boldsymbol{e}\in E(\mathcal{V})}$ characterize the magnitude and direction of the prior dependence structure on $\boldsymbol{\theta}$.

Corollary 4.1. *Under the conditions for Theorem 4.1, let $\{\tau^p_{e_1,e_2|D_{\boldsymbol{e}}}\}_{\boldsymbol{e}\in E(\mathcal{V})}$ describe prior dependence on $\boldsymbol{\theta}$. Suppose no $\boldsymbol{\theta}_0$ with $p_D(\boldsymbol{\theta}_0) > 0$ is such that $C^{Ga}_{\boldsymbol{R}}(\cdot)$ corresponding to the covariance matrix $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ induces a dependence structure $\{\tau(\boldsymbol{\theta})_{e_1,e_2|D_{\boldsymbol{e}}}\}_{\boldsymbol{e}\in E(\mathcal{V})}$ such that $\tau(\boldsymbol{\theta}_0)_{e_1,e_2|D_{\boldsymbol{e}}} = \tau^p_{e_1,e_2|D_{\boldsymbol{e}}}$ for all $\boldsymbol{e} \in E(\mathcal{V})$. Then, the posterior of $\boldsymbol{\theta}|\boldsymbol{Y}^{(n)}$ cannot retain the magnitude and direction of prior dependence as $n \to \infty$.*

Corollary 4.1 follows directly from Theorem 4.1. We suppose there is no $\boldsymbol{\theta}_0$ with $p_D(\boldsymbol{\theta}_0) > 0$ such that the Gaussian copula corresponding to the covariance matrix $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ induces a dependence structure characterized by $\{\tau^p_{e_1,e_2|D_{\boldsymbol{e}}}\}_{\boldsymbol{e}\in E(\mathcal{V})}$. It follows that the magnitude and direction of the dependence structure for $p(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ cannot be characterized by $\{\tau^p_{e_1,e_2|D_{\boldsymbol{e}}}\}_{\boldsymbol{e}\in E(\mathcal{V})}$ for sufficiently large n. We emphasize that considering dependence structures via Kendall's τ on the D-vine structure of the copula does *not* fully specify the dependence structure. This allows for more flexibility in the choices for the copula families; it also facilitates the consideration of prior dependence for subvectors of $\boldsymbol{\theta}$, which is useful because it may require too much cognitive complexity to assess the full prior dependence structure. However, two bivariate copulas may have the same Kendall's τ but different properties in terms of symmetry and tail dependence. We focus on the magnitude and direction of dependence to present generally applicable guidance for prior specification.

Corollary 4.1 therefore provides a straightforward result that can be used to discard potential dependence structures for $\boldsymbol{\theta}$ that cannot be retained as data $\boldsymbol{y}^{(n)}$ are observed.

We refer to prior dependence structures that satisfy the conditions for Corollary 4.1 as chronically rejected dependence structures. The conditions for Corollary 4.1 are sufficient in that there is *no* value for $\boldsymbol{\theta}_0$ with $p_D(\boldsymbol{\theta}_0) > 0$ such that these dependence structures will be retained as data are generated via (4.5). However, it is not guaranteed that the prior dependence structure will be retained when the conditions for Corollary 4.1 are not satisfied. That is, the true value of $\boldsymbol{\theta}_0$ might not be one of those that prevent the conditions for Corollary 4.1 from being satisfied. Thus, a prior dependence structure that is not retained is not necessarily a chronically rejected one. In this chapter, the notion of chronic rejection defines a specific class of prior dependence structures, whereas retention of the prior dependence structure is viewed more generally as a posterior outcome.

4.3.3 Illustration with Copula-Based Priors for the Multinomial Model

We now apply Corollary 4.1 with the multinomial model. In Appendix C.1, we show that the inverse Fisher information matrix for the multinomial model parameterized in terms of $\boldsymbol{\theta} = (Z_1, Z_2, \ldots, Z_{w-1})$ from (4.2) is diagonal for all possible $(Z_1, Z_2, \ldots, Z_{w-1}) \in (0,1)^{w-1}$. If $p(\boldsymbol{\theta})$ incorporates any positive or negative dependence between the conditional multinomial probabilities, this dependence structure is therefore a chronically rejected one.

For $w = 3$ categories, we illustrate this phenomenon when combining such priors with the standard multinomial likelihood, which is parameterized by conditional probabilities $\boldsymbol{\theta} = (Z_1, Z_2)$. We specify $p(\boldsymbol{\theta})$ as in (4.4) where the marginal prior for Z_1 is BETA$(20, 40)$, the marginal prior for Z_2 is BETA$(30, 30)$, and these marginal priors are joined using a Gaussian copula parameterized with Pearson's $\rho = -0.9$. This prior distribution conveys that we expect the multinomial probabilities to be roughly equal for all three categories. This prior specification could be facilitated using either of the methods by Elfadaly and Garthwaite (2017) or Wilson (2018). We note that the prior copula gives rise to a value for Kendall's τ of $2\sin^{-1}(-0.9)/\pi = -0.713$.

For each of 10000 simulation repetitions, we generated $\boldsymbol{\theta}_0 = (Z_{1,0}, Z_{2,0})$ from the prior specified in the previous paragraph. We generated $\boldsymbol{Y}^{(n)}$ for $n = 10$ from the multinomial model parameterized by $\boldsymbol{\theta}_0$. For each sample, we approximated the posterior of $\boldsymbol{\theta}|\boldsymbol{y}^{(n)}$ using sampling-resampling methods (Rubin, 1988), where the proposal distribution was the posterior of $\boldsymbol{\theta}|\boldsymbol{y}^{(n)}$ obtained by independently joining the marginal priors for Z_1 and Z_2. For each posterior sample, we estimated Kendall's τ for Z_1 and Z_2. This process was repeated for $n = \{10^2, 10^3, 10^4, 10^5\}$. The range of Kendall's τ values observed in this numerical study is summarized in Table 4.1.

Table 4.1: Kendall's τ values for 10000 posteriors of Z_1 and Z_2 across various sample sizes n.

	Kendall's τ		
n	Minimum	Median	Maximum
10^1	-0.7128	-0.6806	-0.6638
10^2	-0.5488	-0.5049	-0.4815
10^3	-0.2103	-0.1611	-0.1288
10^4	-0.0464	-0.0214	0.0062
10^5	-0.0326	-0.0022	0.0247

This table shows that the posterior cannot retain the negative dependence structure present in the prior as data are observed from the prior predictive distribution. Even though the analysis prior $p(\boldsymbol{\theta})$ coincides with that used to define the prior predictive distribution of $\boldsymbol{Y}^{(n)}$, the multinomial likelihood function cannot accommodate this dependence between Z_1 and Z_2. As the sample size increases, we observe that the impact of the likelihood function on the posterior overwhelms that of the prior, and the prior dependence structure will eventually not be retained when the conditions for Corollary 4.1 hold.

The copula-based priors proposed by Elfadaly and Garthwaite (2017) and Wilson (2018) define valid posteriors for $\boldsymbol{\theta}$ when combined with the multinomial likelihood, but Table 4.1 corroborates that we would draw vastly different conclusions about the posterior dependence structure of $\boldsymbol{\theta}$ for small and large samples. Elfadaly and Garthwaite (2017) provided R code to combine their priors with the multinomial likelihood function, and Wilson et al. (2021) stated that these priors could be used for Bayesian analysis. However, neither contribution clearly disclosed that these more flexible prior dependence structures cannot be retained as multinomial data are collected. Given a candidate prior dependence structure, Corollary 4.1 helps assess whether this dependence structure for $\boldsymbol{\theta}$ can be retained a posteriori. We elaborate on how to determine whether the inability to retain the prior dependence structure presents practical issues in Section 4.4.

4.4 Objectives for Dependence Structure Specification

This section outlines several objectives that practitioners may wish to achieve when specifying prior distributions and their dependence structures. For each objective, we discuss

whether retention of the prior dependence structure as defined in Corollary 4.1 is important. This knowledge can be used to assess the adequacy of an elicited prior distribution.

4.4.1 Supplementation of Small Samples

Even if a prior dependence structure cannot be retained as the sample size n increases, it may be used to supplement information from small samples. The following example that illustrates the utility of chronically rejected dependence structures for small samples was helpfully provided by a reviewer of the paper associated with this chapter (Hagar and Stevens, 2024b (under review)). We suppose that a practitioner aims to model the heights of a new animal species with a $\mathcal{N}(\mu, \sigma^2)$ distribution. The practitioner does not know the typical heights for members of this species, but it is reasonable to expect that μ and σ are on the same scale. Hence, positive prior dependence between μ and σ^2 is sensible. Upon observing a single member of this species, the practitioner can estimate μ – even though the single observation does not directly provide information about σ^2. If μ and σ^2 were positively correlated a priori, the posterior could convey that a value of one nanometer for σ would be unlikely given an observed height of one meter.

For this model with $\boldsymbol{\theta} = (\mu, \sigma^2)$, the inverse Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$ is diagonal for all possible values of $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. It follows by Corollary 4.1 that the joint posterior of μ and σ^2 cannot retain a prior dependence structure for which Kendall's τ is not 0. Under the conditions for Corollary 4.1, any attempt to inject positive or negative dependence into this posterior will be unsuccessful in the limit of infinite data. That is, once both μ and σ^2 are precisely estimated from the data, there will be no practical correlation between their estimates. However, imposing prior dependence is nevertheless useful for small sample sizes.

While chronically rejected dependence structures can supplement information from small samples, it would be ideal if the source of posterior dependence for $\boldsymbol{\theta}$ were transparent. Corollary 4.1 can be used to determine whether the prior or likelihood function gives rise to this dependence. This knowledge would provide additional context with which to interpret posterior analyses. This example also suggests that it is not worthwhile to spend time eliciting complicated prior dependence structures that cannot be retained when collecting large samples.

The use of informative priors to improve model identifiability is related to supplementing small samples with prior information. Although unidentifiability arises in other contexts, the sample size is often small with respect to the complexity of models that cannot be fully identified. Previous contributions detailed how using informative priors to

enhance model identifiability impacts model interpretation and the convergence of Markov chain Monte Carlo methods (Gelfand and Sahu, 1999; Eberly and Carlin, 2000; Gustafson, 2005). These contributions mainly considered informative marginal priors, and discussion of informative prior dependence structures was limited.

4.4.2 Coverage of Credible Sets

Prior dependence structures may also be specified to calibrate Bayesian credible sets. In this context, Gustafson (2012) referred to the data generating prior as "nature's" prior. The data generation prior in our framework is $p_D(\boldsymbol{\theta})$ from (4.5). If the analysis prior $p(\boldsymbol{\theta})$ coincides with "nature's" prior and the model $m(\cdot|\boldsymbol{\theta})$ is correctly specified, credible sets attain their nominal coverage (Gustafson, 2012). We let $\mathcal{J}_{\boldsymbol{\theta},1-\alpha}(\boldsymbol{y}^{(n)})$ denote a credible set for $\boldsymbol{\theta}$ given the observed data $\boldsymbol{y}^{(n)}$ with coverage $1-\alpha$. A popular choice for the credible set is that of highest posterior density (HPD). Credible sets attain their nominal coverage when

$$Pr(\boldsymbol{\theta_0} \in \mathcal{J}_{\boldsymbol{\theta},1-\alpha}(\boldsymbol{Y}^{(n)})) = 1-\alpha, \tag{4.6}$$

where $\boldsymbol{Y}^{(n)} \sim m(\cdot|\boldsymbol{\theta}_0)$ such that $\boldsymbol{\theta}_0$ is drawn from $p_D(\boldsymbol{\theta})$. The probabilistic statement in (4.6) is therefore made with respect to repeated sampling from the prior predictive distribution of $\boldsymbol{Y}^{(n)}$. Once $\boldsymbol{y}^{(n)}$ is observed, the credible set is not random and either contains or does not contain a given value $\boldsymbol{\theta_0}$.

We now investigate how the prior dependence structure impacts the calibration of credible sets for the multinomial example from Section 4.3.3. We used the same prior predictive distribution of $\boldsymbol{Y}^{(n)}$ for this numerical study, joining the BETA$(20, 40)$ prior for Z_1 and BETA$(30, 30)$ prior for Z_2 with a Gaussian copula parameterized with Pearson's $\rho = -0.9$. For each of 10000 simulation repetitions, we approximated the posterior of $\boldsymbol{\theta}|\boldsymbol{y}^{(n)}$ as described in Section 4.3.3 with "nature's" analysis prior. For each posterior, we approximated its 95% HPD set for Z_1 and Z_2 using two-dimensional kernel density estimation (Ripley, 2002). Empirical coverage was estimated as the proportion of simulation repetitions for which the parameter value $\boldsymbol{\theta}_0 = (Z_{1,0}, Z_{2,0})$ used to generate the multinomial data was contained in this HPD set. We implemented this process for $n = \{10^1, 10^2, 10^3, 10^4, 10^5\}$. We then repeated this process for analysis priors $p(\boldsymbol{\theta})$ that joined the marginal beta priors from "nature's" prior with a Gaussian copula parameterized by Pearson's $\rho = \{-0.95, -0.85, -0.80, \ldots, 0.95\}$. The results from this numerical study are visualized in Figure 4.3.

The fact that all five coloured curves in Figure 4.3 roughly intersect the horizontal dotted line when $\rho = -0.9$ confirms the result in (4.6). For small sample sizes n, empirical

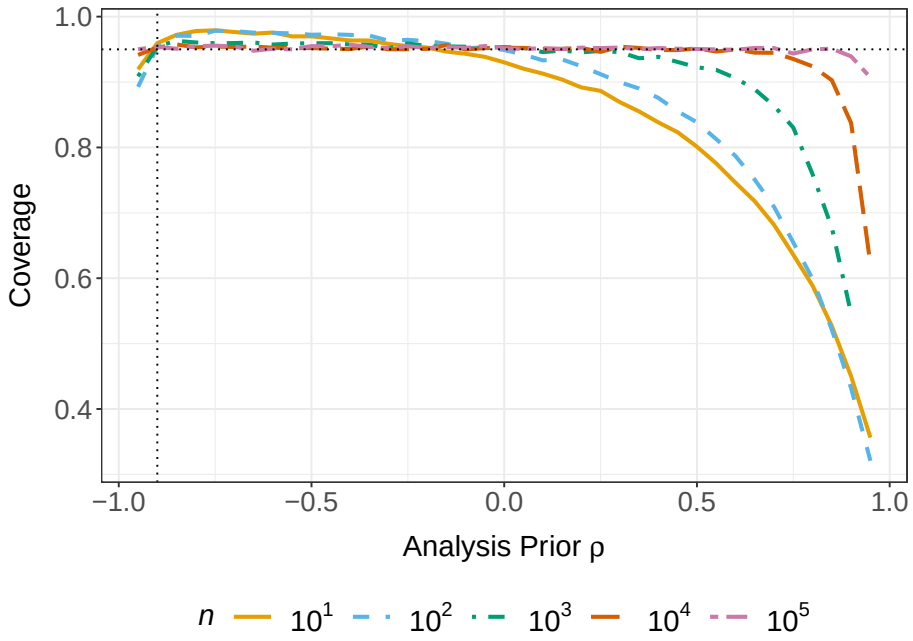


Figure 4.3: Empirical coverage of 95% HPD sets for the multinomial parameter $\boldsymbol{\theta} = (Z_1, Z_2)$ across 10000 posteriors. The horizontal dotted line denotes the nominal coverage, and the vertical one denotes "nature's" prior.

coverage generally deviated from the nominal coverage when $\rho \neq -0.9$. These deviations resulted in worse coverage when $p(\boldsymbol{\theta})$ was specified such that the magnitude of the negative dependence between Z_1 and Z_2 was overstated or when the direction of the dependence was inaccurate. The empirical coverage was greater than the nominal coverage when $p(\boldsymbol{\theta})$ incorporated slightly weaker negative dependence between Z_1 and Z_2 than "nature's" prior. For those settings, the copula density function for $p(\boldsymbol{\theta})$ is flatter on $[0, 1]^2$ than the copula density function that defines "nature's" prior. The resulting credible sets cover a greater range of $\boldsymbol{\theta}$ values but still account for the correct direction of dependence between Z_1 and Z_2.

As the sample size n increases, the impact of prior dependence is reduced and the empirical coverage approaches the nominal coverage for all ρ values considered. This conclusion is to be expected given the asymptotic theory discussed in Section 4.3. The insights drawn from Figure 4.3 are not solely applicable to chronically rejected prior dependence structures. We observed similar results for a numerical study with the gamma model involving analysis priors that do not satisfy the conditions for chronic rejection in Corollary

4.1. Those results are detailed in Appendix C.2.

Given the results from this subsection, we suggest erring on the side of understating the strength of dependence structures a priori. This recommendation is not made to encourage practitioners to engineer credible sets with coverages that exceed their nominal values. This excess coverage may not be desirable and cannot persist as the sample size increases. We instead make this recommendation to mitigate the potential ramifications of overstating the strength of dependence structures on the calibration of credible sets. The notion of harmful priors has been discussed elsewhere (see e.g., Reimherr et al. (2021)). However, priors that are harmful to the calibration of credible sets are not necessarily harmful priors in alternative contexts.

4.4.3 Inference Regarding Dependence Structures

One may also conduct posterior analyses to draw conclusions about the dependence structure of $\boldsymbol{\theta}$. The dependence structure is also relevant when considering the posterior of a function of several parameters in $\boldsymbol{\theta}$. To achieve these objectives with posterior analysis, we recommend using Corollary 4.1 to determine whether the prior dependence structure can be retained. Per Table 4.1, we may draw vastly different conclusions about chronically rejected dependence structures given small and large samples obtained from the same data generation process. Corollary 4.1 provides context to help gauge the reliability of our conclusions for small samples. For large samples, Corollary 4.1 dictates if time should be invested eliciting a dependence structure based on whether it can be retained a posteriori.

Corollary 4.1 could also be used to determine *whether* it is sensible to make inferences about dependence structures. While copula-based priors may be able to accommodate flexible dependence structures, the statistical models chosen for the likelihood function may not have this capability. For the multinomial model, there is a single component for each observation of $\boldsymbol{y}^{(n)} = \{y_i\}_{i=1}^n$. It would be difficult to specify a likelihood function that could accommodate complex dependence structures between the conditional multinomial probabilities given the available data. As such, Corollary 4.1 could diagnose philosophical issues with objectives for posterior analyses.

If each observation of $\boldsymbol{y}^{(n)}$ is comprised of multiple components, incorporating a copula into the likelihood function may promote greater coherence between the prior and posterior dependence structures. For illustration, we suppose that $\boldsymbol{y}^{(n)} = \{(y_i, y_i^*)\}_{i=1}^n$, where $y_i \sim \text{EXP}(\kappa)$, $y_i^* \sim \text{EXP}(\kappa^*)$, and κ and κ^* are rates. The likelihood function for this example

is such that

$$L(\boldsymbol{\theta}; \boldsymbol{y}^{(n)}) \propto \prod_{i=1}^{n} \kappa e^{-\kappa y_i} \times \kappa^* e^{-\kappa^* y_i^*} \times c(1 - e^{-\kappa y_i}, 1 - e^{-\kappa^* y_i^*}; \boldsymbol{\upsilon}), \tag{4.7}$$

where the copula density function is parameterized by $\boldsymbol{\upsilon}$ and $\boldsymbol{\theta} = (\kappa, \kappa^*, \boldsymbol{\upsilon})$. If $c(u_1, u_2; \boldsymbol{\upsilon})$ corresponds to the independence copula, the posterior correlation between κ and κ^* based on the inverse Fisher information will approach 0 as $n \to \infty$. More flexible dependence structures for κ and κ^* could be accommodated a posteriori given different choices for $c(u_1, u_2; \boldsymbol{\upsilon})$. In those settings, eliciting dependence structures for κ and κ^* could be worthwhile – even for large samples since the dependence structure might be retained.

4.4.4 Design Priors

Not all prior distributions are elicited with the intent of defining a posterior. Prior distributions are regularly used for design purposes. For instance, priors might be used to summarize expert opinion to choose inputs for a decision model (Garthwaite et al., 2005). Design priors are also used in experimental settings to conduct sample size determination (De Santis, 2007; Berry et al., 2011; Gubbiotti and De Santis, 2011). In these settings, design priors are often informative and concentrated on $\boldsymbol{\theta}$ values that are relevant to the objective of the study. Data generated according to the design prior are often combined with an uninformative analysis prior to assess whether a posterior criterion is satisfied. Design priors will be used for this purpose in Chapter 5. In the context of this chapter, the prior $p_D(\boldsymbol{\theta})$ from (4.5) could be considered as a design prior.

Design priors are not directly combined with a likelihood function. It is therefore not an issue if the dependence structure in the *design* prior satisfies the conditions for chronic rejection outlined in Corollary 4.1. Generally, we do not need to make separate considerations for small and large samples when using design priors. It is possible that the design prior might coincide with an analysis prior $p(\boldsymbol{\theta})$ that defines a posterior of $\boldsymbol{\theta}$. In that event, the design prior is subject to the previous recommendations in this section. The discussion on the coverage of credible sets in Section 4.4.2 may be relevant if the design and analysis priors do not coincide. We detail additional considerations for design prior specification when reconsidering design with sampling distribution segments in Chapter 5.

4.4.5 Posterior Concentration

Recent work has suggested that the choice of prior dependence structure can expedite the convergence of the posterior around a fixed parameter value $\boldsymbol{\theta}_0$. In particular, Michimae and Emura (2022) suggested that joint priors with vine structures based on Archimedean copulas lead to more accurate and concentrated posterior distributions in the context of Bayesian ridge regression. The accuracy and concentration of the posterior of $\boldsymbol{\theta}$ were considered via the total mean absolute error between the posterior median and a fixed parameter value $\boldsymbol{\theta}_0$. Michimae and Emura's (2022) numerical studies showed that the posterior was more concentrated around $\boldsymbol{\theta}_0$ when the marginal priors for the regression coefficients were joined using (Archimedean) Clayton and Gumbel copulas (Nelsen, 2006) instead of more standard Gaussian copulas.

The asymptotic theory presented in Section 4.3 indirectly suggests the choice of prior dependence structure cannot give rise to increased posterior concentration for large samples. For small samples, further investigation into how the prior dependence structure prompts increased posterior concentration is required. This investigation would disclose whether improving posterior concentration is a sensible objective for prior dependence specification. These recent recommendations serve as one source of motivation to study the impact of the prior copula on the posterior distribution, which we do in Section 4.5.

4.5 Impact of the Prior Copula on the Posterior

4.5.1 Convergence of the Posterior Mode

For an arbitrary model, it may be challenging to correctly specify the dependence structure a priori. As such, we consider how the choice of copula for the prior distribution impacts the posterior. We suppose that two potential priors for $\boldsymbol{\theta}$, denoted $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$, are defined as in (4.4) using the *same* marginal distributions $F_1, \ldots, F_d$ but *different* copula density functions $c_1(\boldsymbol{u})$ and $c_2(\boldsymbol{u})$. We require that both copula density functions are absolutely continuous and twice differentiable with respect to $\boldsymbol{u} = (u_1, \ldots, u_d)$. To ensure the domain of the parameter space is not inadvertently restricted, the priors should be chosen such that $p_1(\boldsymbol{\theta}) > 0$ and $p_2(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

We define posteriors $p_1(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ and $p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ by combining the likelihood $L(\boldsymbol{\theta}; \boldsymbol{y}^{(n)})$ for the model $m(\boldsymbol{y}^{(n)}|\boldsymbol{\theta})$ with $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$, respectively. We summarize each posterior via its posterior mode for $\boldsymbol{\theta}$, denoted by $\tilde{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}} p_k(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ for $k = 1, 2$. In this section, we consider the convergence of the posterior mode to a fixed value $\boldsymbol{\theta}_0$. For a given

sample $\boldsymbol{y}^{(n)}$, we compare the Euclidean distance between $\boldsymbol{\theta}_0$ and each posterior mode, denoted by $\mathcal{D}_k = \|\tilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_0\|_2$ for $k = 1, 2$. In contrast to Michimae and Emura (2022), Theorem 4.2 indicates that we generally do not expect the choice of prior copula to impact whether the posterior mode is closer to $\boldsymbol{\theta}_0$ for large sample sizes n.

Theorem 4.2. *Let $\boldsymbol{Y}^{(n)}$ be generated independently from $m(\cdot|\boldsymbol{\theta}_0)$ such that all conditions for the BvM theorem hold. Let priors $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ be defined as in (4.4) with the same marginals $F_1, \ldots, F_d$ but different copula density functions $c_1(\cdot)$ and $c_2(\cdot)$ that are absolutely continuous and twice differentiable. Suppose $p_1(\boldsymbol{\theta}) > 0$ and $p_2(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Given $\boldsymbol{y}^{(n)}$, define posteriors $p_k(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}) \propto L(\boldsymbol{\theta}; \boldsymbol{y}^{(n)}) p_k(\boldsymbol{\theta})$ with posterior mode $\tilde{\boldsymbol{\theta}}^{(k)}$ for $k = 1, 2$. Let $\boldsymbol{u}_0 = (F_1(\theta_{1,0}), \ldots, F_d(\theta_{d,0}))$ and $\mathcal{D}_k = \|\tilde{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}_0\|_2$ for $k = 1, 2$.*

(a) If $\nabla_{\boldsymbol{u}} \big[\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))\big]_{\boldsymbol{u}=\boldsymbol{u}_0} \neq \boldsymbol{0}$, then $\lim_{n\to\infty} Pr(\mathcal{D}_2 \leq \mathcal{D}_1) = 0.5$.

(b) If $\boldsymbol{u}_0$ is a local maximum of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$, then $\lim_{n\to\infty} Pr(\mathcal{D}_2 \leq \mathcal{D}_1) = 1$.

Theorem 4.2 explains that whether the choice of prior copula gives rise to faster convergence depends on the function $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$. The posterior mode $\tilde{\boldsymbol{\theta}}^{(k)}$ maximizes the logarithm of $p_k(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ for $k = 1, 2$. The log-posteriors $\log[p_1(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})]$ and $\log[p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})]$ differ *only* by their prior copula log-density functions. Differences in $\tilde{\boldsymbol{\theta}}^{(1)}$ and $\tilde{\boldsymbol{\theta}}^{(2)}$ are therefore driven by differences in $\log(c_1(\boldsymbol{u}))$ and $\log(c_2(\boldsymbol{u}))$. Each copula log-density function prompts an additive contribution to the log-posterior. We suppose that $\boldsymbol{u}_0$ is not a stationary point of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$ in part (a). As n increases, the contribution from $\log(c_2(\boldsymbol{u}))$ will not uniformly force its posterior mode closer to (or further from) $\boldsymbol{\theta}_0$ for all samples $\boldsymbol{y}^{(n)}$ than that from $\log(c_1(\boldsymbol{u}))$. When $\boldsymbol{u}_0$ is instead a local maximum of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$, the contribution from $\log(c_2(\boldsymbol{u}))$ will force $\tilde{\boldsymbol{\theta}}^{(2)}$ closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$ for all samples $\boldsymbol{y}^{(n)}$ as n increases. This case is considered in part (b).

While $c_1(\cdot)$ and $c_2(\cdot)$ are functions of $\boldsymbol{\theta}$ because $\boldsymbol{u} = (F_1(\theta_1), \ldots, F_d(\theta_d))$, we consider the partial derivatives of the copula log-density functions with respect to $\boldsymbol{u}$ instead of $\boldsymbol{\theta}$. For $j = 1, \ldots, d$, it follows by the chain rule that

$$\frac{\partial}{\partial \theta_j}\Big[\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))\Big] = f_j(\theta_j)\frac{\partial}{\partial u_j}\Big[\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))\Big]. \tag{4.8}$$

We can factor out the $f_j(\theta_j)$ term because $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ are defined using the same marginals. The priors were also defined such that $p_1(\boldsymbol{\theta}) > 0$ and $p_2(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Therefore, $f_j(\theta_j)$ must be positive, and the partial derivative in (4.8) with respect to θ_j is 0 if and only if the partial derivative with respect to u_j is 0. This correspondence

ensures the results from Theorem 4.2 generalize over different specifications for the marginal distributions.

We prove Theorem 4.2 in Appendix C.3. We note that if $\boldsymbol{u}_0$ is a local minimum of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$, then $\lim_{n\to\infty} Pr(\mathcal{D}_2 \leq \mathcal{D}_1) = 0$. This follows directly from part *(b)* of Theorem 4.2 by switching the labels on $c_1(\cdot)$ and $c_2(\cdot)$. The case where $\boldsymbol{u}_0$ is a saddle point of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$ is excluded from both parts *(a)* and *(b)*. In that case, $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ may converge to a constant that is not 0.5 or 1. We explore the results from Theorem 4.2 via simulation and explain their practical implications in Section 4.5.2.

4.5.2 Practical Implications

Here, we conduct simulations to consider Theorem 4.2 in practice. To do so, we consider an example adapted from Michimae and Emura (2022) since their recommendations are inconsistent with the results of Theorem 4.2. They considered a ridge regression model with three regression coefficients in the presence of multicollinearity, where the parameters of the relevant prior copulas were random variables specified using a hierarchical framework. The simplified example for our numerical study adapts aspects of their model for illustrative purposes. Note that we contrast our results with Michimae and Emura's (2022) findings in Section 4.5.3.

Our simplified example considers the following linear regression model for the outcome y_i and predictors x_{i1} and x_{i2}:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 5)$ independently for $i = 1, \ldots, n$. The assumptions that the linear equation has an intercept of zero and the error terms have known variance reduce the dimensionality of the problem for illustration. That is, $\boldsymbol{\theta} = (\beta_1, \beta_2)$. We specify standard normal marginal priors for both β_1 and β_2.

We join these marginal priors with two prior copulas in this numerical study: $c_1(\boldsymbol{u}) = 1$ for $\boldsymbol{u} \in [0, 1]^2$ corresponds to the independence copula and $c_2(\boldsymbol{u})$ corresponds to a two-dimensional t-copula with $\nu = 4$ degrees of freedom and a diagonal correlation matrix $\boldsymbol{R}$. We select the first copula because it is often assumed that the regression coefficients are independent a priori in Bayesian regression models. The corresponding joint prior for β_1 and β_2 is therefore a standard bivariate normal distribution with diagonal $\boldsymbol{R}$. The choice for the second copula is motivated by $c_2(\boldsymbol{u})$ having a local maximum and saddle points to illustrate the results from Theorem 4.2. Figure 4.4 visualizes the logarithm of this copula

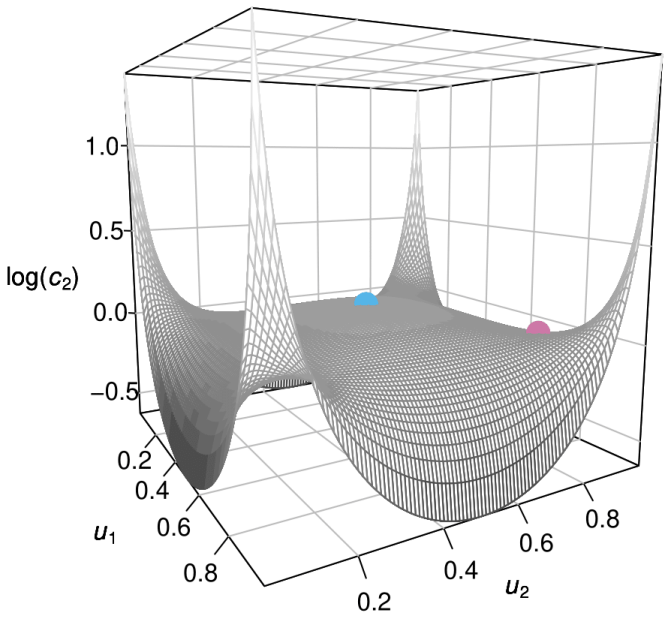


Figure 4.4: The logarithm of the t-copula density function with diagonal $\boldsymbol{R}$ and $\nu = 4$. The local maximum at $\boldsymbol{u} = (0.5, 0.5)$ and saddle point at $\boldsymbol{u} = (0.813, 0.813)$ are given by the blue and pink points, respectively.

density function. The selected t-copula does not accommodate strong negative or positive dependence between β_1 and β_2, but it reflects a greater likelihood of observing extreme values for both β_1 and β_2 relative to their marginal priors. For instance, this might occur if both marginal priors were misspecified.

We consider six values for $\boldsymbol{\theta}_0$ to define the data generation process for $\boldsymbol{Y}^{(n)}$. These six values are meant to illustrate the convergence of the posterior mode in a variety of settings. Because $\boldsymbol{u}_0 = (F_1(\theta_{1,0}), F_2(\theta_{2,0}))$, we can readily convert between $\boldsymbol{\theta}_0$ and $\boldsymbol{u}_0$ given the specified $\mathcal{N}(0, 1)$ marginals. The marginal priors for β_1 and β_2 were chosen to be rather informative so that we observe a range of behaviour for the settings corresponding to part (a) of Theorem 4.2.

For each $\boldsymbol{\theta}_0$ value, we generated 10000 samples of size n for various sample sizes between 5 and 10^5. Each observation was simulated independently as $y_i = \beta_{1,0}x_{i1} + \beta_{2,0}x_{i2} + \varepsilon_i$, where $(x_{i1}, x_{i2}) \sim \mathcal{N}(\boldsymbol{0}, I_2)$ and $\varepsilon_i \sim \mathcal{N}(0, 5)$ for $i = 1, \ldots, n$. For each of these 10000 samples, we

found both posterior modes $\tilde{\boldsymbol{\theta}}^{(1)}$ and $\tilde{\boldsymbol{\theta}}^{(2)}$ using calculus. We then estimated $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ as the proportion of samples for which $\tilde{\boldsymbol{\theta}}^{(2)}$, corresponding to the t-copula, was closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$. To consider the practical impact of Theorem 4.2, we also computed the mean absolute difference between $\mathcal{D}_2$ and $\mathcal{D}_1$ for each sample size n. Figure 4.5 visualizes these results for each of the six scenarios, which we now describe.

We first consider case 1, where $\boldsymbol{u}_0 = (0.5, 0.5)$. This point is a local maximum of $c_2(\boldsymbol{u})$

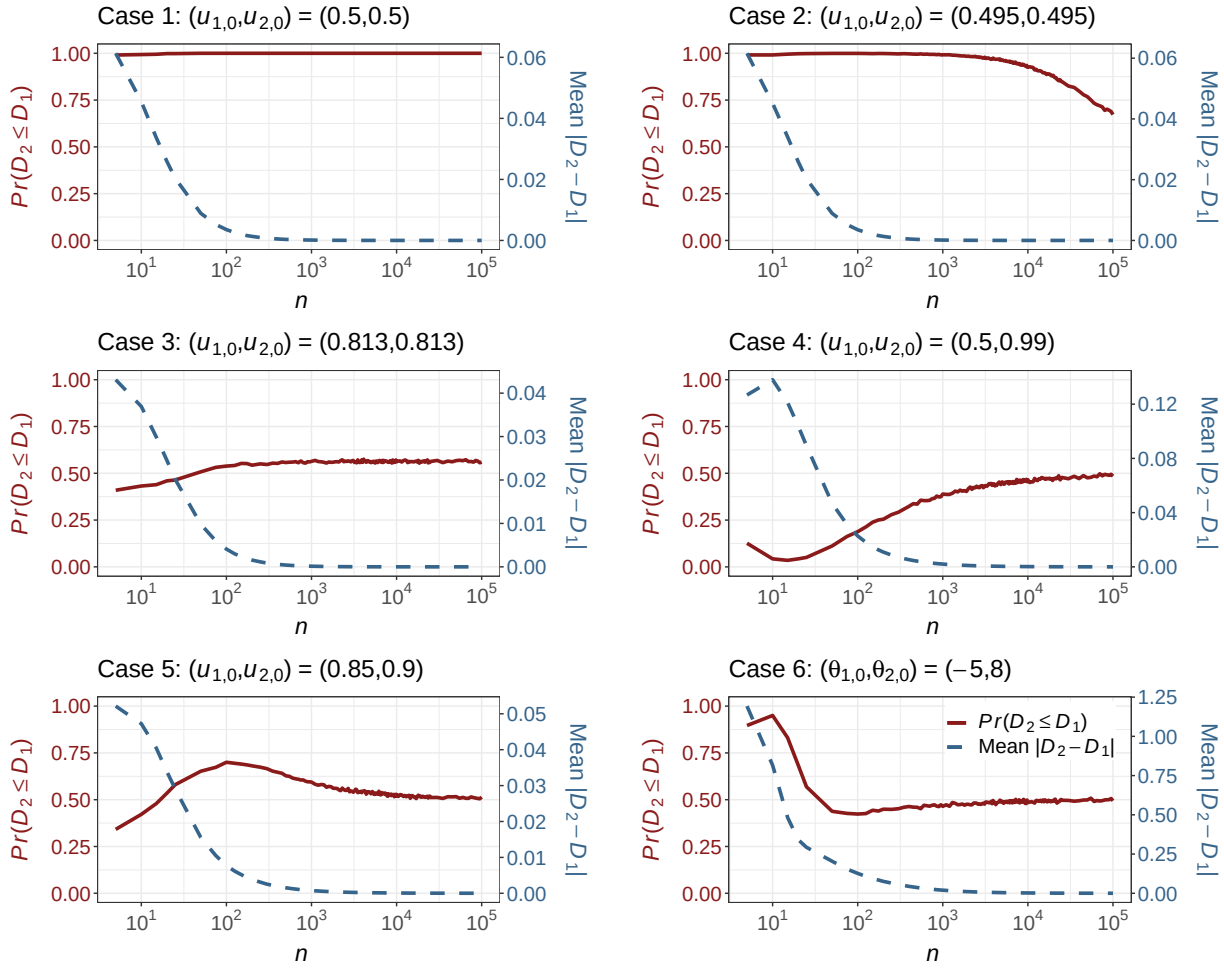


Figure 4.5: Estimated probability that $\tilde{\boldsymbol{\theta}}^{(2)}$ is closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$ (solid red) and mean absolute difference between $\mathcal{D}_2$ and $\mathcal{D}_1$ (dashed blue) as a function of n on the logarithmic scale (base 10) for six $\boldsymbol{\theta}_0$ values.

and therefore also a local maximum of $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u})) = \log(c_2(\boldsymbol{u}))$. As indicated by Theorem 4.2, the probability that $\tilde{\boldsymbol{\theta}}^{(2)}$ is closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$ approaches 1 as $n \to \infty$. Even though $\mathcal{D}_2$ is less than $\mathcal{D}_1$, their absolute difference is practically negligible for large sample sizes n. Case 2 considers a point $\boldsymbol{u}_0 = (0.495, 0.495)$ that is extremely close to the local maximum. The plot for case 2 is similar to the previous one for small n, but the estimate for $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ slowly decreases as $n \to \infty$. The estimated probability is still 0.673 for $n = 10^5$. For $n = 5 \times 10^6$ (not pictured), we estimated $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ to be 0.521. While $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ approaches 0.5 in theory, it may not be 0.5 in practice if the sample size must be prohibitively large for the asymptotic result to hold. Case 3 examines a saddle point at $\boldsymbol{u}_0 = (\tilde{F}(1;4), \tilde{F}(1;4))$, where $\tilde{F}(\cdot;4)$ is the CDF of the Student's t-distribution with $\nu = 4$ and $\tilde{F}(1;4) \approx 0.813$. This setting is noteworthy because $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ does not approach 0.5 or 1. The probability for this scenario instead approaches 0.563, and we confirmed this limiting probability using samples of size $n = 5 \times 10^6$.

Case 4 considers a point $\boldsymbol{u}_0 = (0.5, 0.99)$ where the independence copula performs better for smaller sample sizes in that the estimated probability approaches 0.5 from below. We note that $c_2(\boldsymbol{u}_0) = 0.533 < 1 = c_1(\boldsymbol{u}_0)$, which occurs because this bivariate t-copula deems scenarios where only one of β_1 or β_2 is extreme relative to their marginal priors as more rare than the independence copula. If $c_2(\boldsymbol{u}_0)$ is less (greater) than $c_1(\boldsymbol{u}_0)$, $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ often approaches 0.5 from below (above). However, this behaviour is not guaranteed. Case 5 examines a point $\boldsymbol{u}_0 = (0.85, 0.9)$ at which the estimated probability approaches 0.5 from above. Here, $c_2(\boldsymbol{u}_0) = 1.047 > 1$, but the estimate for $Pr(\mathcal{D}_2 \leq \mathcal{D}_1)$ is less than 0.5 for sample sizes less than $n = 20$. Lastly, case 6 is defined in terms of its value for $\boldsymbol{\theta}_0 = (\beta_{1,0}, \beta_{2,0}) = (-5, 8)$ because the corresponding $\boldsymbol{u}_0$ value of $(2.87 \times 10^{-7}, 1 - 6.22 \times 10^{-16})$ is quite extreme. For this setting, $c_2(\boldsymbol{u}_0) = 5054.68 >> 1$, and $\mathcal{D}_2$ is roughly 0.8 smaller than $\mathcal{D}_1$ on average for $n = 10$. Yet, the estimated probability approaches 0.5 from below for large sample sizes. This likely occurs because $c_2(\boldsymbol{u})$ is volatile near $\boldsymbol{u}_0$.

For each case and prior combination, we also estimated empirical coverage as the proportion of the 10000 posteriors for which the 95% HPD set included the parameter value $\boldsymbol{\theta}_0$ when $n = \{10^1, 10^2, 10^3, 10^4, 10^5\}$. The HPD sets were again estimated using two-dimensional kernel density estimation on posterior draws obtained via sampling-resampling methods. The empirical coverage results are summarized for both prior copulas in Table 4.2. Per Table 4.2, the empirical coverage is similar for both prior copulas at all samples sizes considered in cases 1, 2, 3, and 5. For small sample sizes, the empirical coverage exceeds the nominal value of 95% in cases 1, 2, and 3. Since this trend is observed for both prior copulas, it is caused by the informative nature of the relatively well-specified marginal priors for β_1 and β_2. These marginal priors are misspecified in cases 4 and 6, so the empirical coverage is less than 95% for both prior copulas when the sample size is

Table 4.2: Empirical coverage of 95% HPD sets for $\boldsymbol{\theta} = (\beta_1, \beta_2)$ across 10000 posteriors defined using both prior copulas.

	Case for Independence Copula					
n	1	2	3	4	5	6
10^1	0.9928	0.9916	0.9685	0.8723	0.9465	0.0023
10^2	0.9574	0.9575	0.9558	0.9387	0.9479	0.5698
10^3	0.9510	0.9550	0.9515	0.9509	0.9504	0.9181
10^4	0.9519	0.9525	0.9517	0.9546	0.9542	0.9452
10^5	0.9546	0.9501	0.9507	0.9523	0.9473	0.9508
	Case for t-Copula					
n	1	2	3	4	5	6
10^1	0.9939	0.9938	0.9683	0.7769	0.9465	0.0002
10^2	0.9583	0.9603	0.9570	0.9234	0.9503	0.5148
10^3	0.9509	0.9558	0.9532	0.9481	0.9507	0.9162
10^4	0.9515	0.9532	0.9524	0.9512	0.9553	0.9451
10^5	0.9539	0.9503	0.9518	0.9516	0.9479	0.9524

small. The empirical coverage is better when using the independence copula for small n in both cases – even though $\tilde{\boldsymbol{\theta}}^{(2)}$ was generally closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$ for case 6 when n was small in Figure 4.5.

4.5.3 Connections to Other Work

Our simulations demonstrate the value in considering the copula density functions $c_1(\boldsymbol{u})$ and $c_2(\boldsymbol{u})$ when choosing between two prior dependence structures. In particular, we may want to consider the local optima for the copula densities. This numerical study also suggests that copulas have limited ability to reliably prompt more accurate and concentrated posterior distributions around a fixed parameter value $\boldsymbol{\theta}_0$. The true value of $\boldsymbol{u}_0$ is unknown in practice. And even *if* $\boldsymbol{u}_0$ is such that $c_2(\boldsymbol{u}_0) > c_1(\boldsymbol{u}_0)$, it does not guarantee that $\tilde{\boldsymbol{\theta}}^{(2)}$ will be closer to $\boldsymbol{\theta}_0$ than $\tilde{\boldsymbol{\theta}}^{(1)}$ for small or large sample sizes. Regardless of whether the prior dependence structure is a chronically rejected one, choosing a prior copula to improve posterior concentration does not appear to be a sensible objective.

We now contrast our conclusions with Michimae and Emura's (2022) recommendations. Their numerical studies considered several levels of multicollinearity between the regression

covariates for a *single* $\boldsymbol{\theta}_0$ value. The joint prior for their three regression coefficients leveraged a vine structure with three bivariate copulas. For each bivariate copula, the corresponding $\boldsymbol{u}_0$ value lies directly on the upper Fréchet-Hoeffding bound. Each $\boldsymbol{u}_0$ value does not necessarily correspond to a local maximum of the Clayton or Gumbel copula density functions they considered. However, the Clayton and Gumbel families of copulas accommodate positive prior dependence and their density functions are generally largest near the upper Fréchet-Hoeffding bound. Their single $\boldsymbol{\theta}_0$ value is therefore similar to that used in case 2 of our study. They also considered relatively small sample sizes ranging from $n = 20$ to 200. Given the results for case 2 in Figure 4.5, it is not surprising that their normal posterior medians for the regression coefficients were generally closer to their $\boldsymbol{\theta}_0$ values when using Clayton and Gumbel copulas. Our numerical results suggest that it is pertinent to consider a variety of $\boldsymbol{\theta}_0$ values and sample sizes n before making general statements about the impact of the prior copula on the posterior distribution.

We acknowledge that Michimae and Emura (2022) considered a more complicated model, which leveraged a hierarchical framework to specify the prior copula. Although not the focus of this chapter, Monte Carlo simulation could likely be used to explore the local optima of such prior copula density functions and extend the results from Theorem 4.2 to a hierarchical framework. This extension is one of several possible generalizations that could be made to how the analysis prior in (4.4) is defined.

4.6 Discussion

In this chapter, we proposed a framework to consider whether prior dependence structures can be retained a posteriori. This framework improves transparency when making inference about posterior dependence structures and helps discard chronically rejected dependence structures for a parameter $\boldsymbol{\theta}$ that cannot be retained as data are observed. Discarding such dependence structures simplifies the prior specification process, particularly when practitioners aim to collect large samples. We also discussed small-sample objectives for prior specification to clarify whether the inability to retain the prior dependence structure presents practical issues for a given posterior analysis. This chapter emphasized copula-based priors, but the notion of chronically rejected dependence structures is still applicable to multivariate prior distributions that are not explicitly defined using copulas, such as multivariate normal priors.

Since correctly specifying the dependence structure a priori for an arbitrary model may be difficult, we examined how the choice of prior copula impacts the posterior distribution. We proved asymptotic results regarding how this choice of copula impacts the convergence

of the posterior mode to a fixed parameter value $\boldsymbol{\theta}_0$. While examining the local optima of candidate copula density functions is valuable, our numerical studies showed that prior copulas should not be selected to improve the posterior's ability to recover $\boldsymbol{\theta}_0$. These results contradicted past recommendations that suggested the choice of prior copula could reliably improve posterior concentration.

Future work could extend the results from this chapter to hierarchical settings, in which the hyperparameters of the prior copula are themselves random variables. Moreover, the theoretical results in this chapter are based on the standard BvM theorem. To broaden the applicability of this framework, we could consider nonparametric methods for specifying prior dependence. We could also consider chronically rejected prior dependence structures in the presence of model misspecification (i.e., when $L(\boldsymbol{\theta}; \boldsymbol{y})$ that defines the posterior does not coincide with $m(\cdot|\boldsymbol{\theta})$ used to generate the data). In that case, we require more complex characterizations of prior dependence than $\mathcal{I}(\boldsymbol{\theta}_0)^{-1}$ for large samples.

Chapter 5

Design with Posterior-Based Operating Characteristics

5.1 Preamble

In this chapter, we propose a design framework that boasts several advantages over the methodology for Bayesian hypothesis tests presented in Chapter 3. First, we relax the design assumption that data are generated from statistical models with known, fixed parameter values and can use the guidance for prior elicitation provided in Chapter 4 to specify nondegenerate design priors. Second, we aim to formally incorporate an analogue to type I error into these designs by allowing the critical value γ to vary as a context-specific probabilistic cutoff. The approach from Chapter 3 is aligned with power criteria, but the notion of type I error would need to be explored via additional simulation. Moreover, designs that control posterior-based operating characteristics for both type I and II errors may better comply with requirements of certain regulating bodies. This chapter also considers imbalanced sample size determination in more detail than Chapter 3. Lastly, the work in this chapter accommodates two-group comparisons that account for additional covariates and scenarios where the normal approximation to the sampling distribution of the MLE is poor for moderate sample sizes.

Despite these advantages, the methodology from Chapter 3 is still useful due to its computational efficiency and accessibility to practitioners who have limited experience with Bayesian statistics. While this thesis chapter borrows most notation from Chapter 3, there are a few differences that we highlight here. In this chapter, (fixed) parameter values drawn from a design prior are denoted by $\boldsymbol{\eta}_j^*$ instead of $\boldsymbol{\eta}_{j,0}$ for reasons described later.

Because $\boldsymbol{\eta}^*_{j,n}$ was used to denote approximations to the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n}$ in Chapter 3, these approximations are denoted by $\ddot{\boldsymbol{\eta}}_{j,n}$ in this chapter.

5.2 Background

In recent decades, Bayesian methods for data-driven decision making have become increasingly popular. Two-group comparisons have long been a cornerstone of statistical analysis. Posterior analyses that compare scalar quantities θ_1 and θ_2 are often of interest, where the characteristic θ_j describes a comparison ($j = 1$) or reference ($j = 2$) group. This chapter emphasizes two-group comparisons facilitated via the posterior of $\theta = \theta_1 - \theta_2$, including those made with ratio-based metrics $\theta > 0$ that can be expressed as a difference on the logarithmic scale. For such analyses, interval hypotheses of the form $H_1 : \theta \in (\delta_L, \delta_U)$ are routinely considered, where $-\infty \leq \delta_L < \delta_U \leq \infty$. The interval (δ_L, δ_U) accommodates the context of comparison. Assuming larger θ_j values are preferred, the intervals $(\delta_L, \delta_U) = \{(0, \infty), (-\delta, \delta), (-\delta, \infty)\}$ for an equivalence margin $\delta > 0$ may be used to respectively assess whether θ_1 is superior, practically equivalent, or noninferior to θ_2 (Spiegelhalter et al., 1994, 2004).

Decision-making methods with posterior probabilities have been proposed in a variety of settings (see e.g., Berry et al. (2011); Brutti et al. (2014); Stevens and Hagar (2022)). Given data $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ observed from two groups, the posterior probability $Pr(H_1 \mid \boldsymbol{y}_1, \boldsymbol{y}_2)$ is compared to a critical value $0.5 \leq \gamma < 1$. If that probability is greater than γ, one should conclude $\theta \in (\delta_L, \delta_U)$. When comparing complementary hypotheses $H_1 : \theta \in (\delta_L, \delta_U)$ and $H_0 : \theta \notin (\delta_L, \delta_U)$, decision-making methods with Bayes factors (Jeffreys, 1935; Kass and Raftery, 1995; Morey and Rouder, 2011) were demonstrated to be a special case of those with posterior probabilities in (3.2). This chapter therefore focuses on posterior probabilities, though the methods extend to the use of Bayes factors. Unlike in Chapter 3, hypothesis tests with credible intervals are not addressed with this methodology, but this accommodation could be made in future work.

In clinical trials, regulatory agencies require that Bayesian designs are assessed with respect to frequentist operating characteristics (FDA, 2019). Decision makers in industrial and corporate settings may also want to control the power and type I error rate of Bayesian designs to justify funding studies and using them to draw trustworthy conclusions. Since these design procedures leverage theory from Bayesian and frequentist statistics, they are often called hybrid approaches to sample size determination (Berry et al., 2011). These hybrid approaches involve exploring the sampling distribution of posterior probabilities

under various data generation processes. These sampling distributions are often explored by approximating posteriors corresponding to many hypothetical samples.

In design settings, the data have not been observed and are random variables. In this chapter, data from a random sample are represented by $\boldsymbol{Y}^{(n,q)}$, consisting of observations $\{y_{i1}\}_{i=1}^{n}$ from group 1 and observations $\{y_{i2}\}_{i=1}^{\lfloor qn \rceil}$ from group 2 for some constant $q > 0$. A *design* prior $p_D(\boldsymbol{\eta})$ (De Santis, 2007; Berry et al., 2011; Gubbiotti and De Santis, 2011) models uncertainty regarding the model parameters $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ from each group in pre-experimental settings. As in Chapter 3, the characteristic of interest θ_j for group j is typically specified as a function $g(\cdot)$ of the model parameters: $\theta_j = g(\boldsymbol{\eta}_j)$ for $j = 1, 2$. Since the (informative) design prior is concentrated on θ values that are relevant to the objective of the study, it is usually different from the *analysis* prior used to analyze the observed data. The design prior gives rise to the prior predictive distribution of $\boldsymbol{Y}^{(n,q)}$:

$$p(\boldsymbol{y}^{(n,q)}) = \int \prod_{i=1}^{n} f(y_{i1}; \boldsymbol{\eta}_1) \prod_{i=1}^{\lfloor qn \rceil} f(y_{i2}; \boldsymbol{\eta}_2) p_D(\boldsymbol{\eta}) \, d\boldsymbol{\eta}, \tag{5.1}$$

where $f(y; \boldsymbol{\eta}_j)$ is the model for group $j = 1, 2$. Gubbiotti and De Santis (2011) defined the conditional and predictive approaches for specifying the prior predictive distribution of $\boldsymbol{Y}^{(n,q)}$. The conditional approach assigns all prior weight in $p_D(\boldsymbol{\eta})$ to a design value $\boldsymbol{\eta}^*$, whereas the predictive approach uses a nondegenerate design prior.

Various methods have been proposed to control posterior-based operating characteristics (Berry et al., 2011; Gubbiotti and De Santis, 2011; Brutti et al., 2014). To control both power and the type I error rate for a posterior analysis, it is often necessary to specify two prior predictive distributions for $\boldsymbol{Y}^{(n,q)}$. One defines the power criterion under the assumption that the hypothesis H_1 is true. The selected sample size n ensures the probability of concluding that H_1 is true is at least $1 - \beta$ for some target power $1 - \beta \in (0, 1)$:

$$\mathbb{E}\left[\mathbb{I}\{Pr(H_1 \,|\, \boldsymbol{Y}_1^{(n,q)}) \geq \gamma\}\right] \geq 1 - \beta, \tag{5.2}$$

for some critical value $\gamma \in [0.5, 1)$. The criterion in (5.2) is considered when $\boldsymbol{Y}_1^{(n,q)} \sim p_1(\boldsymbol{y}_1^{(n,q)})$ as defined in (5.1) with a design prior $p_{D_1}(\boldsymbol{\eta})$ such that $p_{D_1}(H_1) = 1$.

In contrast, a criterion for the type I error rate is defined by assuming that H_1 is false. To bound the type I error rate, the selected critical value γ ensures the probability of concluding that H_1 is true is at most α for some significance level $\alpha \in (0, 1)$:

$$\mathbb{E}\left[\mathbb{I}\{Pr(H_1 \,|\, \boldsymbol{Y}_0^{(n,q)}) \geq \gamma\}\right] \leq \alpha. \tag{5.3}$$

The criterion in (5.3) is considered when $\boldsymbol{Y}_0^{(n,q)} \sim p_0(\boldsymbol{y}_0^{(n,q)})$ as defined in (5.1) with a design prior $p_{D_0}(\boldsymbol{\eta})$ such that $p_{D_0}(H_1) = 0$. For design with posterior-based operating characteristics, the choice of critical value $\gamma \in [0.5, 1)$ is not purely dictated by the upper bound α for the type I error rate. With a degenerate design prior $p_{D_0}(\boldsymbol{\eta})$ such that $\theta^* = g(\boldsymbol{\eta}_1^*) - g(\boldsymbol{\eta}_2^*)$ equals δ_L or δ_U, the sampling distribution of $Pr(H_1 | \boldsymbol{Y}_0^{(n,q)})$ converges to a uniform distribution as $n \to \infty$ under weak conditions (Golchi and Willard, 2023). In such cases, choosing $\gamma \approx 1 - \alpha$ will satisfy the criterion in (5.3) for large sample sizes. However, the optimal choice for γ may differ substantially from $1 - \alpha$ for moderate sample sizes or when nondegenerate design priors are used to define the prior predictive distribution of $\boldsymbol{Y}_0^{(n,q)}$. Although not pursued in this chapter, the proposed methodology can be trivially extended to control the false discovery rate (FDR) by taking $\alpha = (1 - \beta) \ / \ (1/\text{FDR} - 1)$.

To support flexible study design, (n, γ) combinations that control posterior-based operating characteristics can be found using simulation. Most simulation-based procedures to evaluate the power criterion in (5.2) with design priors follow a similar process (Wang and Gelfand, 2002). First, an (n, γ) combination is selected. Second, a value $\boldsymbol{\eta}^*$ is drawn from the design prior $p_{D_1}(\boldsymbol{\eta})$.[3] Third, data $\boldsymbol{y}_{*1}^{(n,q)}$ are generated according to the model $f(y; \boldsymbol{\eta}^*)$. Fourth, the posterior of θ given $\boldsymbol{y}_{*1}^{(n,q)}$ is approximated to check if $Pr(H_1 | \boldsymbol{y}_{*1}^{(n,q)}) > \gamma$. This process is repeated many times to determine whether the power criterion is satisfied with probability at least $1 - \beta$ for the selected (n, γ) combination.

A similar process can be repeated to evaluate whether the criterion in (5.3) is satisfied for a given (n, γ) combination with the design prior $p_{D_0}(\boldsymbol{\eta})$, samples $\boldsymbol{y}_{*0}^{(n,q)}$, and significance level α. To find a suitable design, time is wasted considering (n, γ) combinations that are suboptimal. This computational inefficiency is compounded over all combinations of the design inputs that practitioners wish to investigate – including the interval (δ_L, δ_U), design and analysis priors, and values for α, β, and q. A fast framework to determine the (n, γ) combination that minimizes the sample size n while satisfying criteria for both posterior-based operating characteristics would mitigate this issue and expedite collaborative study design.

Recently, several strategies have been employed to reduce the computational burden associated with controlling posterior-based operating characteristics in Bayesian study design. Certain strategies are tailored to specific statistical models. For instance, Shi and Yin (2019) exploited the monotonicity of posterior probabilities as a function of the number of successful Bernoulli trials to find optimal critical values that maintained type I error

[3]The draw $\boldsymbol{\eta}^*$ from the design prior is akin to $\boldsymbol{\eta}_0$ from Chapter 3. However, we do not use zeros to denote fixed parameter values in this chapter since "0" and "1" are used to distinguish between sampling distributions under H_0 and H_1.

rates in sequential designs. Other approaches accommodate a variety of statistical models. One such general strategy imposes parametric assumptions on the sampling distribution of posterior probabilities. Golchi (2022) fit beta distributions to approximate such sampling distributions for various design values $\boldsymbol{\eta}^*$ using Gaussian processes that exploited spatial correlation between similar design inputs. Golchi and Willard (2023) presented an alternative method to fit those beta distributions using asymptotic theory.

An alternative general strategy that we propose here involves exploring segments of sampling distributions of posterior probabilities. In Chapter 3, we proposed such a method for power curve approximation with posterior analyses. That method prioritizes exploring posterior probabilities such that $Pr(H_1|data) \approx \gamma$ without imposing parametric assumptions on the sampling distribution of $Pr(H_1|data)$. That approach is useful but its simplifying assumptions may be impractical in complex design scenarios. First, we only considered the prior predictive distribution of $\boldsymbol{Y}^{(n,q)}$ under degenerate design priors. Second, our method from Chapter 3 did not jointly explore the (n, γ)-space and required complete reimplementation to consider various, user-specified γ values. Finally, we did not consider type I error rates, so the impact of the analysis priors on such rates was not well studied. Here, we overcome these limitations from Chapter 3 by generalizing our design methods with sampling distribution segments to facilitate scalable design with posterior-based operating characteristics.

The remainder of this chapter is structured as follows. We describe an example with genetically modified crops that involves the comparison of ordinal means in Section 5.3. This example is referenced throughout the chapter to illustrate the proposed methods. In Section 5.4, we present a general framework to define nondegenerate design priors under the assumption that H_1 is true or false, and we choose design priors for the illustrative example. In Section 5.5, we propose a method to determine which (n, γ) combination minimizes the sample size n while satisfying the criteria in (5.2) and (5.3). This procedure explores segments of sampling distributions of posterior probabilities using theoretical results that we prove in this chapter. In Section 5.6, we repurpose the posterior probabilities used to find the optimal (n, γ) combination to create contour plots that facilitate the investigation of various n and γ values; this process is illustrated for the example with ordinal data. We conclude with a summary and discussion of extensions to this work in Section 5.7. Additional theoretical results and numerical studies are provided in Appendix D. To streamline the discussion in the main portion of this chapter, our design framework for two-group comparisons with additional covariates is also proposed in that appendix.

5.3 Illustrative Example

Since malnutrition caused by nutrient intake deficiencies is a serious concern in the African country of Malawi, the investigation of genetically modified maize (corn) varieties is prevalent. These varieties contain more provitamin A carotenoids that are converted into vitamin A during the digestion process than standard varieties. Munkhuwa et al. (2022) recently conducted a study at the Lilongwe University of Agriculture and Natural Resources (LUANAR) to compare an existing maize variety (MH43 A, provitamin A level: 9.3 μg/g) with a newer one (MH44 A, provitamin A level: 9.6 μg/g). While the newer maize variety boasts a higher provitamin A level, this increase will not lead to more vitamin A production if there are substantial aversions to the newer variety compared to existing ones.

The LUANAR study characterized how much children between 6 and 24 months of age enjoyed a porridge sample made with one of the two maize varieties using a Likert scale (Likert, 1932) with $w = 5$ categories. Scores of 1 and 5 respectively indicated that the child was very dissatisfied and very satisfied with the porridge sample, prompting an observation $y_{ij} \in \{1, 2, 3, 4, 5\}$ for each child $i = 1, ..., n_j,\ j = 1, 2$. In total, $n_1 = 108$ and $n_2 = 137$ children were given porridge samples made with the MH44 A ($j = 1$) and MH43 A ($j = 2$) varieties, respectively. For group j, the multinomial distribution assumes that each participant is assigned Likert response v with probability $0 < p_{jv} < 1$ for $v = 1, ..., w$ such that $\sum_{v=1}^{w} p_{jv} = 1$. Our metric of interest is $\theta_j = \mathbb{E}(y_{ij}) = \sum_{v=1}^{w} v p_{jv}$ for group $j = 1, 2$. We consider the new maize variety to be noninferior to the existing one when $\theta = \theta_1 - \theta_2 \in (-0.5, \infty)$, where $\delta_L = -0.5$ was chosen for illustration to reflect half of the distance between consecutive categories on this ordinal scale.

The observed sample means for the Likert data are $\hat{\theta}_1 = 4.18$ and $\hat{\theta}_2 = 4.38$. The distribution of Likert responses for each maize variety is visualized in the left plot of Figure 5.1. We assign uninformative Dirichlet DIR$(0.8, 0.8, 0.8, 0.8, 0.8)$ priors to $\boldsymbol{\eta}_j = \boldsymbol{p}_j = (p_{j1}, p_{j2}, ..., p_{j5})$ for $j = 1, 2$. We obtain 10^5 posterior draws for $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ using MCMC methods to approximate the posterior of $\theta = \theta_1 - \theta_2$ given $\boldsymbol{y}_1 = \{y_{i1}\}_{i=1}^{n_1}$ and $\boldsymbol{y}_2 = \{y_{i2}\}_{i=1}^{n_2}$, which is illustrated in the right plot of Figure 5.1. The posterior probability $Pr(\theta > -0.5 \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = 0.9877$ is larger than most conventional critical values $\gamma \in [0.5, 1)$, suggesting that the new maize variety is noninferior to the old. Nevertheless, design methods that prescribe γ to control posterior-based operating characteristics prior to observing data provide a valuable framework to draw informed conclusions based on such posterior probabilities.

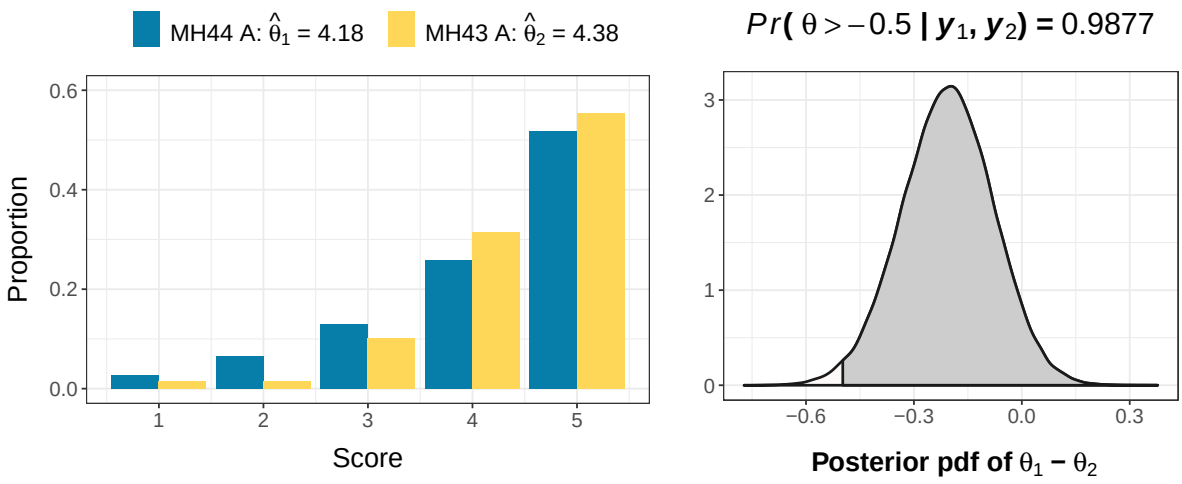


Figure 5.1: Left: Distribution of Likert data for each maize variety. Right: Visualization of the posterior for the difference between the ordinal means.

5.4 A Framework for Design Prior Specification

5.4.1 Design Prior Specification and Segmentation

Our framework for prior specification directly elicits priors for the model parameters $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ and indirectly induces priors on the characteristics θ_1, θ_2, and θ. Prior specification for the model parameters corresponding to a single group of data was considered in Chapter 4. For two-group comparisons, existing knowledge of the reference group ($j = 2$) can often be used to choose a prior $p_D(\boldsymbol{\eta}_2)$ and induce a prior $p_D(\theta_2)$ on $\theta_2 = g(\boldsymbol{\eta}_2)$. Interactive graphical interfaces are commonly used to facilitate iterative prior elicitation procedures (Chaloner, 1996; Williams et al., 2021). These interfaces provide instant feedback regarding how changes to the directly specified prior $p_D(\boldsymbol{\eta}_j)$ impact the induced prior $p_D(\theta_j)$. In Section 5.4.2, we demonstrate the utility of such procedures for the illustrative example. Prior specification for $\boldsymbol{\eta}_1$ and θ_1 in the comparison group ($j = 1$) is typically more difficult. However, we can often use visualization techniques along with the anticipated effect size for $\theta = \theta_1 - \theta_2$ and the prior $p_D(\boldsymbol{\eta}_2)$ to ensure the priors $p_D(\boldsymbol{\eta}_1)$ and $p_D(\theta_1)$ are suitable.

We aim to reduce the cognitive burden associated with specifying separate design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ for the criteria in (5.2) and (5.3), respectively. To do so, we specify a design prior $p_D(\boldsymbol{\eta})$ for $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ that is diffuse enough to ensure that the induced prior $p_D(\theta)$ assigns nonneglible prior weight to the interval (δ_L, δ_U) *and* its neighbouring regions.

We respectively define $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ for the power and type I error rate criteria by segmenting the prior $p_D(\boldsymbol{\eta})$. For design purposes, we define two regions in the θ-space to carry out this segmentation. The first region $\mathcal{G} = \{(G_L, G_U) : \delta_L \leq G_L < G_U \leq \delta_U\}$ pertains to the power criterion in (5.2), where $\theta \in \mathcal{G}$ ensures that $H_1 : \theta \in (\delta_U, \delta_L)$ is true. The second region $\mathcal{R} = \{(R_{1L}, R_{1U}] \cup [R_{2L}, R_{2U}) : R_{1L} < R_{1U} \leq \delta_L < \delta_U \leq R_{2L} < R_{2U}\}$ pertains to the criterion for the type I error rate in (5.3). The region $\mathcal{R}$ is defined to accommodate interval hypotheses based on practical equivalence where both $\theta < \delta_L$ and $\theta > \delta_U$ are undesirable outcomes. When either δ_L or δ_U is not finite, the region $\mathcal{R}$ can be expressed as a single region $\mathcal{R} = \{(R_L, R_U] : R_L < R_U \leq \delta_L \vee [R_L, R_U) : \delta_U \leq R_L < R_U\}$. For either scenario, the hypothesis H_1 is false when $\theta \in \mathcal{R}$. In this work, we refer to the regions $\mathcal{G}$ and $\mathcal{R}$ as the *green* and *red* regions of the θ-space, respectively. Further details concerning the definition of these regions are provided in Section 5.4.2.

One method to define the design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ involves truncating the more diffuse prior $p_D(\boldsymbol{\eta})$ according to the regions $\mathcal{G}$ and $\mathcal{R}$ such that $p_{D_1}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta})\mathbb{I}(\theta \in \mathcal{G})$ and $p_{D_0}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta})\mathbb{I}(\theta \in \mathcal{R})$. Since we can readily sample from $p_D(\boldsymbol{\eta})$, rejection sampling methods (Casella et al., 2004) allow us to obtain samples from $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$. Alternatively, we could define design priors such that $p_{D_1}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta}|\theta \sim \mathcal{U}(\mathcal{G}))$ and $p_{D_0}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta}|\theta \sim \mathcal{U}(\mathcal{R}))$, where $\mathcal{U}(\cdot)$ indicates that θ is uniformly distributed over that region. These design priors provide a mechanism for obtaining parameter values $\boldsymbol{\eta}$ corresponding to particular regions of the θ-space that de-emphasizes the shape of the induced prior $p_D(\theta)$. For these design priors, we can use sampling-resampling methods (Rubin, 1988; Smith and Gelfand, 1992) to obtain a sample from $p_{D_1}(\boldsymbol{\eta})$ or $p_{D_0}(\boldsymbol{\eta})$ given a sample from $p_D(\boldsymbol{\eta})$. This sampling-resampling approach is the one we employ in this chapter, but a host of other methods could also be used to choose the design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$. We recommend consulting the literature on prior elicitation if alternative prior specification methods are required (Chaloner, 1996; Garthwaite et al., 2005).

With respect to the objectives for prior specification from Chapter 4, it typically does not matter whether the design prior dependence structure of $p_{D_1}(\boldsymbol{\eta})$ or $p_{D_0}(\boldsymbol{\eta})$ can be retained a posteriori. The design priors will not be combined with the relevant likelihood functions, so it could be worthwhile to invest time eliciting a complicated dependence structure that would not be retained upon collecting enough data. We generally advocate for simplifying design prior specification where possible to ensure this process is feasible, but the extensiveness of this process should be dictated by the stakeholders of a study. Lastly, we note that our framework can be simplified to accommodate the conditional approach to define prior predictive distributions for $\boldsymbol{Y}_1^{(n,q)}$ and $\boldsymbol{Y}_0^{(n,q)}$ given a pair of design values $\boldsymbol{\eta}_{\mathcal{G}}^*$ and $\boldsymbol{\eta}_{\mathcal{R}}^*$ corresponding to θ values in $\mathcal{G}$ and $\mathcal{R}$, respectively.

5.4.2 Design Priors for the Illustrative Example

For the multinomial model used in Section 5.3, it is not trivial to choose an informative prior for $\boldsymbol{\eta}_j = \boldsymbol{p}_j$ that enforces the unit-sum constraint $\sum_{v=1}^{w} p_{jv} = 1$. We instead assign a joint prior to variables obtained with the invertible transformation from Elfadaly and Garthwaite (2017) discussed in Chapter 4:

$$Z_{j1} = p_{j1}, \quad Z_{jv} = \frac{p_{jv}}{1 - \sum_{t=1}^{v-1} p_{jt}} \text{ for } v = 2, ..., w-1, \text{ and } Z_{jw} = 1, \tag{5.4}$$

for groups $j = 1, 2$. This transformation was previously considered in Section 4.2.3 for a single group of data. The variable Z_{jv} represents the probability that an observation from group j is assigned to category v given that it has not been assigned to categories $1, ..., v-1$. We assign marginal BETA$(\alpha_{jv}, \beta_{jv})$ priors to Z_{jv}, $v = 1, ..., w-1$ to induce a joint prior on $\boldsymbol{p}_j$ that satisfies the unit-sum constraint. We join the marginal beta priors with an independence copula for illustration. If $\{Z_{jv}\}_{v=1}^{w-1}$ are independent, $\boldsymbol{p}_j$ follows a standard Dirichlet distribution when $\beta_v = \alpha_{v+1} + \beta_{v+1}$ for $v = 1, ..., w-2$, and a generalized Dirichlet distribution (Connor and Mosimann, 1969) otherwise.

The following interactive graphical interface was developed and used to specify design priors for the illustrative example: `https://luke-hagar.shinyapps.io/Ordinal_Priors/`. To use this interface to specify $p_D(\boldsymbol{\eta}_j)$, practitioners input point estimates $\hat{p}_{j1}, ..., \hat{p}_{jw}$ such that $\sum_{v=1}^{w} \hat{p}_{jv} = 1$. For the reference group ($j = 2$), we used point estimates informed by the Likert data: $(\hat{p}_{21}, ..., \hat{p}_{25}) = (0.015, 0.015, 0.102, 0.314, 0.554)$. Initial estimates for the medians of $\{Z_{jv}\}_{v=1}^{w-1}$ are populated via (4.2). Progressing from $v = 1$ to $w-1$, practitioners consider the ξ-quantiles of each Z_{jv} variable for some $0 < \xi \neq 0.5 < 1$. These estimates for the median and ξ-quantile uniquely define marginal beta priors for each Z_{jv} variable. We used this process with $\xi = 0.95$ to specify the following marginal priors: BETA$(2.20, 123.29)$ for Z_{21}, BETA$(2.15, 118.50)$ for Z_{22}, BETA$(3.43, 29.87)$ for Z_{23}, and BETA$(6.67, 12.16)$ for Z_{24}. These priors jointly induce a design prior on θ_2 that is visualized in the center plot of Figure 5.2. The prior median of 4.38 coincides with the observed ordinal mean for the reference group.

To specify $p_D(\boldsymbol{\eta}_1)$ for the comparison group, we consider the reference data and observed effect size of -0.2, which serves as an anticipated effect size for this illustration. The point estimates $(\hat{p}_{11}, ..., \hat{p}_{15}) = (0.029, 0.040, 0.138, 0.305, 0.488)$ were obtained by systematically shifting probability mass to lower ordinal categories until the point estimate for the ordinal mean was 4.18. We repeated the process detailed above to specify the following marginal priors: BETA$(1.99, 56.22)$ for Z_{11}, BETA$(3.16, 66.19)$ for Z_{12}, BETA$(5.61, 34.18)$ for Z_{13}, and BETA$(11.66, 19.45)$ for Z_{14}. The induced design prior on θ_1 with prior median of

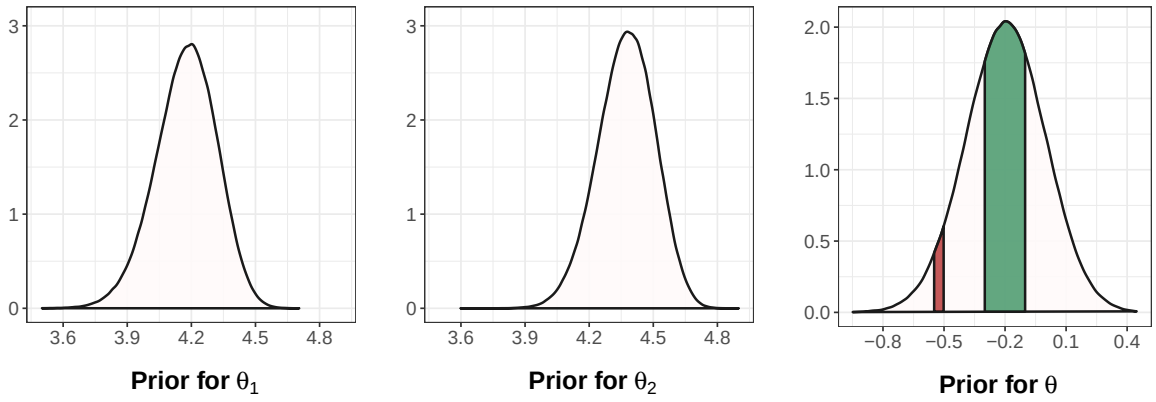


Figure 5.2: Induced design priors for θ_1 (left), θ_2 (center), and θ (right). The green and red regions of the θ-space are visualized on the right plot.

4.18 is visualized in left plot of Figure 5.2. Under the assumption that $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are independent, the induced prior on θ is depicted in the right plot of Figure 5.2.

We now provide general guidance for choosing the regions $\mathcal{G}$ and $\mathcal{R}$, and we overview how the recommended (n, γ) combination depends on these choices. We focus on the case where one of δ_L or δ_U is not finite, but this guidance can be extended to settings where $-\infty < \delta_L < \delta_U < \infty$. First, we advise that $\mathcal{G}$ and $\mathcal{R}$ be chosen as noncontiguous regions so that the study aims to detect meaningful effects. If $G_L = R_U + \epsilon$ or $R_L = G_U + \epsilon$ for some small $\epsilon > 0$, then impractically large sample sizes n may be required to discern miniscule differences between $\theta \in \mathcal{R}$ and $\theta \in \mathcal{G}$. We recommend choosing $\mathcal{R}$ to be contiguous with the interval (δ_L, δ_U). We recommend centering $\mathcal{G}$ around an anticipated or meaningful value for $\theta \in (\delta_L, \delta_U)$ such that discerning differences between $\theta \in \mathcal{G}$ and $\theta \notin (\delta_L, \delta_U)$ is important, where there is sufficient distance between the endpoints of $\mathcal{G}$ and (δ_L, δ_U).

If $R_L << R_U < \delta_L$ or $R_U >> R_L > \delta_U$, the optimal critical value γ typically approaches 0.5 as the sample size n increases because $Pr(H_1 | \boldsymbol{Y}_1^{(n,q)}) \to 1$ and $Pr(H_1 | \boldsymbol{Y}_0^{(n,q)}) \to 0$. With smaller values for $\gamma \in [0.5, 1)$, we require less evidence to support H_1; however, specifying a wide interval for $\mathcal{R}$ will lead to an inflated type I error rate if we are truly only concerned with controlling type I error for θ values that are *just* outside the interval (δ_L, δ_U). As such, we generally recommend specifying $\mathcal{R}$ to be a narrow interval that is contiguous with (δ_L, δ_U). These recommendations are applied with the illustrative example in Section 5.3: the region $\mathcal{G} = (-0.3, -0.1)$ is centered at the anticipated effect size of -0.2, and $\mathcal{R} = (-0.55, -0.5)$ is contiguous with the interval $(\delta_L, \delta_U) = (-0.5, \infty)$ defined previously. These red and green regions are depicted on the prior for θ in the right plot of Figure 5.2.

5.5 Design with Multiple Operating Characteristics

5.5.1 Mapping the Sampling Distribution of Posterior Probabilities to Low-Dimensional Hypercubes

Design methods with posterior-based operating characteristics typically require that we estimate sampling distributions of posterior probabilities for various sample sizes n. Here, we extend methods from Chapter 3 to improve computational complexity by mapping these sampling distributions to low-dimensional hypercubes $[0,1]^{2d}$, where the model $f(y;\boldsymbol{\eta}_j)$ is parameterized by $\boldsymbol{\eta}_j \in \mathbb{R}^d$. Given parameter values $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$, each observation in $\boldsymbol{y}^{(n,q)}$ is typically simulated using CDF inversion with one coordinate of the point $\boldsymbol{u} \in [0,1]^{n_1+n_2}$. We typically have that $2d << n_1 + n_2$, and this dimension reduction allows us to estimate posterior-based operating characteristics using only a subspace of $[0,1]^{2d}$ in Section 5.5.2.

Our design framework assumes that data $\{y_{i1}\}_{i=1}^{n_1}$ and $\{y_{i2}\}_{i=1}^{n_2}$ are to be collected independently, where the data generation process for samples of size $n_1 = n$ and $n_2 = \lfloor qn \rceil$ is characterized by the procedure detailed in Section 5.2. That is, data from group j are generated from the model $f(y;\boldsymbol{\eta}_j^*)$, where the parameter values $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)$ are drawn from the relevant design prior. These parameter values specify anticipated values $\theta_j^* = g(\boldsymbol{\eta}_j^*)$ for the characteristics of interest and their difference $\theta^* = \theta_1^* - \theta_2^*$.

To broadly map posterior probabilities to low-dimensional hypercubes, we can generate maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ instead of data $\boldsymbol{y}^{(n,q)}$. For sufficiently large sample sizes, the MLEs $\hat{\boldsymbol{\eta}}_{j,n_j}$ approximately and independently follow $\mathcal{N}(\boldsymbol{\eta}_j^*, \mathcal{I}^{-1}(\boldsymbol{\eta}_j^*)/n_j)$ distributions, $j = 1, 2$. We require a sequence of m points $\{\boldsymbol{u}_r\}_{r=1}^m \in [0,1]^{2d}$ to simulate from the joint limiting distribution of $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$, where each point corresponds to a simulation repetition. We can use these maximum likelihood estimates for posterior approximation when the models $f(y;\boldsymbol{\eta}_1)$ and $f(y;\boldsymbol{\eta}_2)$ belong to the exponential family (Lehmann and Casella, 1998). We accommodate posterior approximation for models that are not in the exponential family in Appendix D.4.

For models $f(y;\boldsymbol{\eta}_j)$ in the exponential family, the first derivative of the log-likelihood with respect to $\eta_{j,k}$, the k^{th} component of $\boldsymbol{\eta}_j$, takes the form

$$\frac{\partial}{\partial \eta_{j,k}} \log \left[\prod_{i=1}^{n_j} f(y_{ij};\boldsymbol{\eta}_j) \right] = -n \frac{\partial}{\partial \eta_{j,k}} A(\boldsymbol{\eta}_j) + \sum_{s=1}^{d} \frac{\partial}{\partial \eta_{j,k}} C_s(\boldsymbol{\eta}_j) \sum_{i=1}^{n_j} T_s(y_{ij}), \tag{5.5}$$

where $A(\boldsymbol{\eta}_j)$, $C_s(\boldsymbol{\eta}_j)$, and $T_s(y)$ are known functions for $s = 1, ..., d$. The result in (5.5) is the same result as that stated in (3.7) of Chapter 3. As in Chapter 3, $T_{j^\dagger}(\boldsymbol{y}^{(n,q)}) =$

$(\sum_{i=1}^{n_j} T_1(y_{ij}), ..., \sum_{i=1}^{n_j} T_d(y_{ij}))$ are sufficient statistics for group j that provide as much information about the parameter $\boldsymbol{\eta}_j$ as the entire sample $\boldsymbol{y}^{(n,q)}$. Given a maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{j,n_j}$, all components of $T_{j\dagger}(\boldsymbol{y}^{(n,q)})$ can be recovered by substituting $\hat{\boldsymbol{\eta}}_{j,n_j}$ into the linear system that arises from equating all components of (5.5) to 0.

Given analysis priors $p_1(\boldsymbol{\eta}_1)$ and $p_2(\boldsymbol{\eta}_2)$, we use the Laplace approximation (Gelman et al., 2020) to the posteriors of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ along with the multivariate delta method to obtain the following large-sample approximation to the posterior of θ:

$$\mathcal{N}\left(g(\tilde{\boldsymbol{\eta}}_{1,n_1}) - g(\tilde{\boldsymbol{\eta}}_{2,n_2}), \sum_{j=1}^{2}\left[\frac{\partial g}{\partial \boldsymbol{\eta}}^T \mathcal{J}_j(\boldsymbol{\eta})^{-1}\frac{\partial g}{\partial \boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}_{j,n_j}}\right) \text{ for } \mathcal{J}_j(\boldsymbol{\eta}) = -\frac{\partial^2}{\partial \boldsymbol{\eta}^2}\log(p_j(\boldsymbol{\eta}|data)), \tag{5.6}$$

where $\tilde{\boldsymbol{\eta}}_{j,n_j} = \arg\max_{\boldsymbol{\eta}_j} p_j(\boldsymbol{\eta}_j|data)$ is the posterior mode for groups $j = 1$ and 2. As in Chapter 3, we henceforth consider the relevant posteriors conditional on the general vector or matrix *data* instead of conditioning on $\boldsymbol{y}^{(n,q)}$, $\boldsymbol{y}_1$, or $\boldsymbol{y}_2$. This choice again allows us to consider posterior probabilities prompted by conduits for the data and those produced by generating samples $\boldsymbol{y}^{(n,q)}$ using unified notation. The approximation to the posterior of θ in (5.6) is the same as that from (3.6) if the function $h(\theta_1, \theta_2)$ is $\theta = \theta_1 - \theta_2$. In this chapter, we emphasize such two-group comparisons based on differences.

Algorithm 5.1 details how we map a single point $\boldsymbol{u} \in [0,1]^{2d}$ to the posterior approximation in (5.6), where $\hat{\eta}_{j,n_j,k}$ and $\eta^*_{j,k}$ denote the k^{th} component of their vectors. As in Chapter 3, we require that the conditions for the BvM theorem hold to ensure the process to generate maximum likelihood estimates in Algorithm 5.1 is valid. The four necessary assumptions to invoke the BvM theorem (van der Vaart, 1998) are detailed in Appendix B.1.1. The first three assumptions are weaker than the regularity conditions for the asymptotic normality of the MLE (Lehmann and Casella, 1998), which are listed in Appendix B.1.2. The final assumption for the BvM theorem requires that the prior distribution of $\boldsymbol{\eta}_j$ be continuous in a neighbourhood of $\boldsymbol{\eta}^*_j$ with positive density for $j = 1, 2$.

Algorithm 5.1 is effectively the same mapping process as Algorithm 3.2, where we now accommodate imbalanced sample size determination. In Appendix D.3.1, we propose an alternative method to obtain $T_{1\dagger}(\boldsymbol{y}^{(n,q)})$ and $T_{2\dagger}(\boldsymbol{y}^{(n,q)})$ from $\boldsymbol{u} \in [0,1]^{2d}$ that is useful when the distributions of $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ are not approximately normal for moderate sample sizes. That method is based on linear approximations to the CDF of $T_{j\dagger}(\boldsymbol{y}^{(n,q)})$ for discrete models in the exponential family. As such, the illustrative example from Section 5.3 is useful since it allows us to explore and overcome departures from asymptotic normality with finite sample sizes.

For concision, we let $\mathcal{N}(\underline{\theta}_r^{(n,q)}, \underline{\tau}_r^{(n,q)})$ denote the approximation to the posterior of θ in

Algorithm 5.1 Mapping Posteriors to $[0,1]^{2d}$ with Imbalanced Sample Sizes

procedure MAPIMBALANCED($f(y;\boldsymbol{\eta}_1^*)$, $f(y;\boldsymbol{\eta}_2^*)$, $g(\cdot)$, n, q, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$)
  for j in 1:2 **do**
    for k in 1:d **do**
      Let $\hat{\eta}_{j,n_j,k}(\boldsymbol{u})$ be the $u_{(j-1)d+k}$-quantile of the conditional normal CDF of $\hat{\eta}_{j,n_j,k}(\boldsymbol{u}) \mid \{\hat{\eta}_{j,n_j,l}(\boldsymbol{u})\}_{l=0}^{k-1}$ where $\hat{\boldsymbol{\eta}}_{j,n_j}(\boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{\eta}_j^*, \mathcal{I}(\boldsymbol{\eta}_j^*)^{-1}/n_j)$.
    Equate the system of equations in (5.5) to 0 with $\boldsymbol{\eta}_j = \hat{\boldsymbol{\eta}}_{j,n_j}(\boldsymbol{u})$ to solve for $T_{j\dagger}(\boldsymbol{y}^{(n,q)})$.
    Use $T_{j\dagger}(\boldsymbol{y}^{(n,q)})$ to obtain the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n_j}$ via optimization.
  Use $\tilde{\boldsymbol{\eta}}_{1,n_1}(\boldsymbol{u})$, $\tilde{\boldsymbol{\eta}}_{2,n_2}(\boldsymbol{u})$, $T_{1\dagger}(\boldsymbol{y}^{(n,q)})$, $T_{2\dagger}(\boldsymbol{y}^{(n,q)})$, and $g(\cdot)$ to obtain (5.6).

(5.6) corresponding to the point $\boldsymbol{u}_r \in [0,1]^{2d}$ and sample sizes $n_1 = n$ and $n_2 = \lfloor qn \rceil$ for $r = 1, ..., m$. The mean $\underline{\theta}_r^{(n,q)}$ and variance $\underline{\tau}_r^{(n,q)}$ of this approximation are implicit functions of n. These posteriors can be mapped to posterior probabilities $Pr(\theta < \delta \mid data)$ as follows:

$$p^{\delta}_{n,q,\boldsymbol{u}_r} = \Phi\left(\frac{\delta - \underline{\theta}_r^{(n,q)}}{\sqrt{\underline{\tau}_r^{(n,q)}}}\right), \tag{5.7}$$

where $\Phi(\cdot)$ is the the CDF of the standard normal distribution. The estimates from (5.7) comprise the sampling distributions of posterior probabilities after mapping to $[0,1]^{2d}$. These distributions should accurately approximate the exact sampling distribution of posterior probabilities; Theorem 5.1 demonstrates that the total variation distance between the sampling distribution of posterior probabilities induced by (5.6) with data $\boldsymbol{Y}^{(n,q)}$ and that prompted by Algorithm 5.1 with pseudorandom sequences converges to 0 as $n \to \infty$. We emphasize that Theorem 5.1 applies to the sampling distribution of posterior probabilities when $\boldsymbol{\eta}^* \sim p_{D_1}(\boldsymbol{\eta})$ as in (5.2) or when $\boldsymbol{\eta}^* \sim p_{D_0}(\boldsymbol{\eta})$ as in (5.3).

Theorem 5.1. *Let $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*) \sim p_D(\boldsymbol{\eta})$ for some design prior $p_D(\boldsymbol{\eta})$ such that the following conditions hold for all $\boldsymbol{\eta}^*$ with $p_D(\boldsymbol{\eta}^*) > 0$. Let $f(y;\boldsymbol{\eta}_1^*)$ and $f(y;\boldsymbol{\eta}_2^*)$ satisfy the regularity conditions from Appendix B.1.2. Let the prior $p_j(\boldsymbol{\eta}_j)$ be continuous in a neighbourhood of $\boldsymbol{\eta}_j^*$ with positive density for $j = 1, 2$. Let $g(\boldsymbol{\eta})$ be differentiable at $\boldsymbol{\eta}_j^*$ for $j = 1, 2$ with nonzero derivatives. Let $\boldsymbol{U} \overset{i.i.d.}{\sim} \mathcal{U}([0,1]^{2d})$ and $\boldsymbol{Y}^{(n,q)}$ be generated independently from $f(y;\boldsymbol{\eta}_1^*)$ and $f(y;\boldsymbol{\eta}_2^*)$. Let $\mathcal{P}^{\delta}_{n,q,\Pi,\zeta}$ denote the sampling distribution of posterior probabilities for $Pr(\theta < \delta \mid data)$ given sample sizes $n_1 = n$ and $n_2 = \lfloor qn \rceil$ produced using input Π with method ζ. Let $\|Q_1 - Q_2\|_{TV}$ be the total variation distance between two probability measures Q_1 and Q_2. Then, $\|\mathcal{P}^{\delta}_{n,q,\boldsymbol{Y}^{(n,q)},(5.6)} - \mathcal{P}^{\delta}_{n,q,\boldsymbol{U},Alg.5.1}\|_{TV} \xrightarrow{P} 0$.*

The proof of Theorem 5.1 is given in Appendix D.1.1. Strictly speaking, the exact sampling distribution of posterior probabilities $\mathcal{P}^{\delta}_{n,q,\boldsymbol{Y}^{(n,q)},(5.6)}$ converges in distribution to a sampling distribution of posterior probabilities based on pseudorandom sequences and an analogue to *Algorithm 3.1* that accounts for imbalanced sample sizes. Algorithm 5.1 is an analogue to Algorithm 3.2 that is based on the Laplace approximation instead of the approximation prompted by the BvM theorem. In Chapter 3, we showed that $\|\mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.2} - \mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}\|_{TV} \xrightarrow{P} 0$ as defined in Theorem 3.1. Since this result holds true and we recommend using Algorithm 3.2 over Algorithm 3.1 in Chapter 3, we only focus on the Laplace approximation to the posterior in this chapter for simplicity.

We can improve upon this procedure by using randomized low-discrepancy sequences instead of pseudorandom ones (as suggested in Theorem 5.1) to estimate power and the type I error rate more precisely. We use randomized Sobol' sequences (Sobol', 1967) for this purpose as in Chapter 3. Based on the discussion in Section 1.2.3, randomized Sobol' sequences prompt consistent estimators for power and the type I error rate:

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^{m}\Psi(\boldsymbol{U}_r)\right) = \int_{[0,1]^{2d}}\Psi(\boldsymbol{u})d\boldsymbol{u}, \tag{5.8}$$

for the function $\Psi(\cdot)$ defined in Corollary 5.1. Due to the negative dependence between the points in randomized low-discrepancy sequences, the variance of the estimator in (5.8) is typically reduced compared to estimators informed by pseudorandom sequences. With Sobol' sequences, we can therefore use fewer simulation repetitions m to estimate posterior-based operating characteristics as detailed in Corollary 5.1. This corollary follows directly from Theorem 5.1 and (5.8).

Corollary 5.1. *Let $p^{\delta}_{n,q,\boldsymbol{u}_r}$ from (5.7) be the estimate for $Pr(\theta < \delta \mid data)$ corresponding to sample sizes $n_1 = n$ and $n_2 = \lfloor qn \rceil$ and point $\boldsymbol{u}_r \in [0,1]^{2d}$. Let $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r} = p^{\delta_U}_{n,q,\boldsymbol{u}_r} - p^{\delta_L}_{n,q,\boldsymbol{u}_r}$ for $\delta_L < \delta_U$. Under the conditions for Theorem 5.1 as $n \to \infty$, power in (5.2) and the type I error rate in (5.3) are consistently estimated by*

$$\frac{1}{m}\sum_{r=1}^{m}\mathbb{I}(p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r} \geq \gamma),$$

when $\{\boldsymbol{U}_r\}_{r=1}^{m}$ are generated using pseudorandom or randomized Sobol' sequences.

Corollary 5.1 ensures that Algorithm 5.1 with randomized Sobol' sequences gives rise to consistent estimators for power and the type I error rate as $n \to \infty$ when the hypercube $[0,1]^{2d}$ is thoroughly explored with all points from such sequences. However, it does not

guarantee that these estimators are unbiased for finite n. As in Chapter 3, we may require fewer observations for the approximate distributions of the MLEs to be approximately normal if we consider some transformation of $\boldsymbol{\eta}_j$, particularly if any of its parameters do not have support on $\mathbb{R}$. Similarly, the posterior of a monotonic transformation of θ may need to be considered for the normal approximation in (5.6) to be suitable for moderate n. Rather than introduce new notation for these untransformed and transformed variables, we assume that $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, and θ are specified to improve the quality of the normal approximation in (5.6).

5.5.2 Estimating Operating Characteristics with Sampling Distribution Segments

We now propose a method to explore segments of sampling distributions of posterior probabilities. The novel method presented in Algorithm 5.2 allows us to estimate posterior-based operating characteristics without the need to explore points from throughout $[0,1]^{2d}$ to estimate *entire* sampling distributions of posterior probabilities. This method is the main contribution of this chapter, and we use it to consistently explore (n, γ) combinations with only a subset of the points $\boldsymbol{u}_r \in [0,1]^{2d+1},\ r = 1, ..., m$. As described later in Algorithm 5.2, we add an extra dimension to the hypercube so that we can sample $\boldsymbol{\eta}^*$ from the relevant design prior $p_D(\boldsymbol{\eta})$. It is using only a subset of such points – and corresponding sampling distribution segments – to explore most designs that greatly enhances the scalability of our method.

The mapped posterior probabilities $p^{\delta}_{n,q,\boldsymbol{u}_r}$ depend on the models $f(y;\boldsymbol{\eta}_1^*)$ and $f(y;\boldsymbol{\eta}_2^*)$, the sample size n, and the Sobol' sequence point $\boldsymbol{u}_r,\ r = 1, ..., m$. Standard practice fixes the sample size n and varies the point $\boldsymbol{u}_r \in [0,1]^{2d+1}$ to estimate power and the type I error rate. We now fix the point $\boldsymbol{u}_r$ and let the sample size n vary. When the point $\boldsymbol{u}_r$ and models $f(y;\boldsymbol{\eta}_j^*)$ are fixed, $p^{\delta}_{n,q,\boldsymbol{u}_r}$ is a deterministic function of n. Lemma 5.1 motivates our approach to explore $[0,1]^{2d+1}$ in a targeted manner for each sample size n considered. This lemma is written more generally than Lemma 3.1 from Chapter 3; this generality allows us to accommodate two-group comparisons with additional covariates in Appendix D.5.

Lemma 5.1. *Let the conditions for Theorem 5.1 be satisfied and define* $\text{logit}(x) = \log(x) - \log(1-x)$. *For a given point* $\boldsymbol{u}_r = (u_1, ..., u_{2d+1}) \in [0,1]^{2d+1}$, *conditional generation of normal conduits for the data* $\hat{\boldsymbol{\eta}}_n = (\hat{\boldsymbol{\eta}}_{1,n}, \hat{\boldsymbol{\eta}}_{2,qn})$ *using the* $\boldsymbol{u}_r$*-quantiles prompts the following results:*

(a) $\hat{\eta}_{n,k}(\boldsymbol{u}_r) = \eta_k^* + \dfrac{\omega_k(u_1, ..., u_k)}{\sqrt{n}}$ *for* $k = 1, ..., 2d$, *where* $\omega_k(\cdot)$ *are functions that do not depend on* n.

(b) $\hat{\theta}_n = g_*(\hat{\boldsymbol{\eta}}_n(\boldsymbol{u}_r)) \approx g_*(\boldsymbol{\eta}^*) + \dfrac{\omega_\dagger(u_1, ..., u_{2d})}{\sqrt{n}}$ *for sufficiently large* n*, where* $g_*(\cdot)$ *and* $\omega_\dagger(\cdot)$ *are functions that do not depend on* n*.*

(c) $p^{\delta}_{n,q,\boldsymbol{u}_r} \approx \Phi\left(a(\delta, \theta^*)\sqrt{n} + b(\boldsymbol{u}_r)\right)$ *for sufficiently large* n*, where* $\theta^* = g_*(\boldsymbol{\eta}^*)$ *and* $a(\cdot)$ *and* $b(\cdot)$ *are functions that do not depend on* n*.*

(d) $\lim_{n\to\infty} \dfrac{d}{dn} \text{logit}\left[\Phi\left(a(\delta_U, \theta^*)\sqrt{n} + b(\boldsymbol{u}_r)\right) - \Phi\left(a(\delta_L, \theta^*)\sqrt{n} + b(\boldsymbol{u}_r)\right)\right]$ *is* $\min\{a(\delta_U, \theta^*)^2, a(\delta_L, \theta^*)^2\}/2$ *when* $\theta^* \in [\delta_L, \delta_U]$ *and* $-\min\{a(\delta_U, \theta^*)^2, a(\delta_L, \theta^*)^2\}/2$ *otherwise.*

We prove Lemma 5.1 in Appendix D.2. Here, we describe how the more general notation in this lemma maps to the notation from Algorithm 5.1. In Algorithm 5.1, the maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ serve as conduits for the data. While Algorithm 5.1 generates $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ independently, this process is equivalent to generating $\hat{\boldsymbol{\eta}}_n = (\hat{\boldsymbol{\eta}}_{1,n_1}, \hat{\boldsymbol{\eta}}_{2,n_2}) \in \mathbb{R}^{2d}$ from a normal distribution where the covariance matrix has a 2×2 block structure with $\mathbf{0}$ matrices on the off-diagonals. For part (b) of Lemma 5.1, the difference between the characteristics θ can be expressed as a function g_* of $\hat{\boldsymbol{\eta}}_n$: $\theta = g_*(\hat{\boldsymbol{\eta}}_n) = g(\hat{\boldsymbol{\eta}}_{1,n_1}) - g(\hat{\boldsymbol{\eta}}_{2,n_2})$. The remaining notation in Lemma 5.1 aligns with that in Algorithm 5.1, and further modifications to the notation for posterior analyses with additional covariates are detailed in Appendix D.5.

We now consider the practical implications of Lemma 5.1. This lemma suggests that the linear approximation to $\text{logit}(p^{\delta_U - \delta_L}_{n,q,\boldsymbol{u}_r})$ as a function of n is a good global approximation for sufficiently large sample sizes. Moreover, this linear approximation should be locally suitable for a range of sample sizes. These linear approximations disclose which points $\boldsymbol{u}_r \in [0,1]^{2d+1}$ correspond to posterior probabilities that are in a neighbourhood of the β-quantile of the sampling distribution for the design prior $p_{D_1}(\boldsymbol{\eta})$ specified in (5.2). Likewise, those approximations reveal which points $\boldsymbol{u}_r \in [0,1]^{2d+1}$ correspond to posterior probabilities that are in a neighbourhood of the $(1-\alpha)$-quantile of the sampling distribution for $p_{D_0}(\boldsymbol{\eta})$ specified in (5.3). This knowledge allows us to explore segments of the sampling distributions of posterior probabilities in a targeted manner. Lemma 5.1 is original to this chapter. In Chapter 3, we proved that $p^{\delta_U - \delta_L}_{n,q,\boldsymbol{u}_r}$ was increasing for sufficiently large n when $\theta^* \in (\delta_L, \delta_U)$. We used that fact to prioritize exploring posterior probabilities such that $Pr(H_1 | data) \approx \gamma$. However, that approach to select sampling distribution segments based on root-finding algorithms is of limited use when the critical value γ is not predetermined.

Algorithm 5.2 allows users to *jointly* explore the (n, γ)-space in a targeted manner to find the (n, γ) combination that minimizes the sample size while satisfying the criteria in (5.2) and (5.3). This flexibility is crucial for design with posterior-based operating

characteristics. Our approach generates independent Sobol' sequences $\{\boldsymbol{u}_r^{(h)}\}_{r=1}^m$ for each hypothesis H_h, $h = 0, 1$. To implement our approach, we must choose a Sobol' sequence length m and a constant $m_0 << m$. We use $m = 8192$ and $m_0 = 512$ to balance the computational efficiency and precision of the estimates for the operating characteristics. Algorithm 5.2 leverages order statistics, and we herein abbreviate the term order statistic as OS. Our approach involves thorough exploration of sampling distributions of posterior probabilities at three sample sizes – $n^{(0)}$, $n^{(1)}$, and $n^{(2)}$ – and exploration of sampling distribution segments for all other values of n.

Algorithm 5.2 Procedure to Determine Optimal Sample Size and Critical Value

1: **procedure** OPTIMIZE($f(y;\boldsymbol{\eta})$, $g(\cdot)$, $p_j(\boldsymbol{\eta}_j)$, $p_{D_1}(\boldsymbol{\eta})$, $p_{D_0}(\boldsymbol{\eta})$, (δ_L, δ_U), q, β, α, m, m_0)
2: **for** h in 0:1 **do**
3: Generate a sample $\boldsymbol{\eta}^{*(h)} \sim p_{D_h}(\boldsymbol{\eta})$ of size m and Sobol' sequence $\{\boldsymbol{u}_r^{(h)}\}_{r=1}^m$
4: Reorder $\boldsymbol{\eta}^{*(h)}$ so its r^{th} realization prompts the $\lceil m u_{r,1}^{(h)} \rceil^{\text{th}}$ OS of $g(\boldsymbol{\eta}_1^{*(h)}) - g(\boldsymbol{\eta}_2^{*(h)})$
5: Use Algorithm 5.1 with $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^{m_0}$ and $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^{m_0}$ to obtain the smallest $n^{(0)}$ such that the $\lfloor m_0\beta \rfloor^{\text{th}}$ OS of $p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(1)}} \geq$ the $\lceil m_0(1-\alpha) \rceil^{\text{th}}$ OS of $p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(0)}}$ via binary search
6: Compute $p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(1)}}$ for $\{\boldsymbol{u}_r^{(1)}\}_{r=m_0}^m$ and $p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(0)}}$ for $\{\boldsymbol{u}_r^{(0)}\}_{r=m_0}^m$ via Algorithm 5.1
7: $n^{(1)} \leftarrow \left\lfloor 0.9n^{(0)} + 0.2n^{(0)}\, \mathbb{I}\left(\lfloor m\beta \rfloor^{\text{th}} \text{ OS of } p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(1)}} \geq \lceil m(1-\alpha) \rceil^{\text{th}} \text{ OS of } p^{\delta_U-\delta_L}_{n^{(0)},q,\boldsymbol{u}_r^{(0)}} \right) \right\rfloor$
8: **for** r in 1:m **do**
9: **for** h in 0:1 **do**
10: Compute $p^{\delta_U-\delta_L}_{n^{(1)},q,\boldsymbol{u}_r^{(h)}}$ to approximate $\text{logit}(p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r^{(h)}})$ as a linear function of n
11: Find the smallest $n^{(2)}$ such that the $\lfloor m\beta \rfloor^{\text{th}}$ OS of $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(1)}} \geq$ the $\lceil m(1-\alpha) \rceil^{\text{th}}$ OS of $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(0)}}$ via binary search, where each n value is considered with only the m_0 points from $\{\boldsymbol{u}_r^{(h)}\}_{r=1}^m$ such that $\hat{p}^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r^{(h)}}$ from Line 10 is nearest to the relevant OS
12: **for** r in 1:m **do**
13: Compute $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(1)}}$ and $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(0)}}$ via Algorithm 5.1 if not computed in Line 11
14: **return** $n^{(2)}$ as recommended n and the $\lceil m(1-\alpha) \rceil^{\text{th}}$ OS of $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(0)}}$ as γ

We now elaborate on several of the steps in Algorithm 5.2. In Line 4, we use the first coordinate of each point $u_{r,1}^{(h)}$ to reorder the draws in $\boldsymbol{\eta}^{*(0)}$ and $\boldsymbol{\eta}^{*(1)}$ with respect to the anticipated value for $\theta = g(\boldsymbol{\eta}_1) - g(\boldsymbol{\eta}_2)$. This reordering is beneficial because we utilize only the first m_0 points from $\{\boldsymbol{u}_r^{(h)}\}_{r=1}^m$, $h = 0, 1$ to find an initial sample size $n^{(0)}$ in Line 5. Because subsequences of the Sobol' sequence are also low discrepancy, this reordering

guarantees that the anticipated values for θ corresponding to the first m_0 points are evenly distributed over $\mathcal{G}$ and $\mathcal{R}$. For the r^{th} point in the sequence for hypothesis h, we obtain the posterior approximation for (5.7) via Algorithm 5.1, where the inputs $\boldsymbol{\eta}^*$ and $\boldsymbol{u} \in [0,1]^{2d}$ are respectively the r^{th} realization of $\boldsymbol{\eta}^{*(h)}$ and the final $2d$ components of $\boldsymbol{u}_r^{(h)}$. The inequality for the order statistics of the two sampling distributions in Line 5 must hold true for there to exist a critical value γ such that the criteria in both (5.2) and (5.3) are satisfied.

We approximate $\text{logit}(p_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L})$ as a linear function of n to efficiently explore sample sizes. To obtain this approximation for finite n, we do not use the first derivatives from part (d) of Lemma 5.1 prompted by limiting results. We instead construct this approximation by estimating the probabilities that correspond to $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$ for the sample size $n^{(1)}$. This integer sample size is chosen to be larger or smaller than $n^{(0)}$ depending on the indicator function in Line 7. The linear approximations are obtained in Lines 8 to 10 as the lines that respectively pass through $\text{logit}(p_{n^{(0)},q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L})$ and $\text{logit}(p_{n^{(1)},q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L})$ at the sample sizes $n^{(0)}$ and $n^{(1)}$. Once these linear approximations are obtained for each point $\boldsymbol{u}_r^{(h)}$, we find the optimal (n,γ) combination in Line 11. We find the optimal design by exploring sample sizes with binary search. However, we leverage Lemma 5.1 to determine whether a value for n is suitable using a subset of m_0 points from each of $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$.

Unlike in Line 5, we choose these points in a targeted way from each sequence to correspond to sampling distribution segments. For each sample size n we consider, we estimate each posterior probability (and their order statistic) using the linear approximations on the logit scale: $\hat{p}_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L}$ for $r = 1, ..., m$ and $h = 0, 1$. We use Algorithm 5.1 to approximate $p_{n,q,\boldsymbol{u}_r^{(1)}}^{\delta_U-\delta_L}$ for points in $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$ that correspond to relevant order statistics of $\hat{p}_{n,q,\boldsymbol{u}_r^{(1)}}^{\delta_U-\delta_L}$ near $\lfloor m\beta \rfloor$. If $2m\beta < m_0$, these order statistics are the smallest m_0 ones; otherwise, we consider the order statistics ranging from $\lfloor m\beta \rfloor - m_0/2 + 1$ to $\lfloor m\beta \rfloor + m_0/2$. Similarly, we only approximate $p_{n,q,\boldsymbol{u}_r^{(0)}}^{\delta_U-\delta_L}$ for points in $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$ that correspond to relevant order statistics of $\hat{p}_{n,q,\boldsymbol{u}_r^{(0)}}^{\delta_U-\delta_L}$ near $\lceil m(1-\alpha) \rceil$. If $2m\alpha < m_0$, these order statistics are the largest m_0 ones; otherwise, we consider the order statistics ranging from $\lfloor m(1-\alpha) \rfloor - m_0/2 + 1$ to $\lfloor m(1-\alpha) \rfloor + m_0/2$. When computing power and the type I error rate, we assume the $p_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L}$ values for the remaining points do not differ enough from their estimates $\hat{p}_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L}$ to impact the order statistics in Line 11. This approach allows us to accommodate minor discrepancies between $p_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L}$ and $\hat{p}_{n,q,\boldsymbol{u}_r^{(h)}}^{\delta_U-\delta_L}$ without exploring all points $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$.

Algorithm 5.2, however, allows us to obtain the same level of simulation precision as if

we were to use all m points from each sequence to explore every sample size considered. Our method requires that we consider these $2m$ points for only three sample sizes: $n^{(0)}, n^{(1)}$, and $n^{(2)}$. In Lines 12 and 13, we consistently estimate power and the type I error rate at the optimal sample size $n^{(2)}$. The optimal critical value is the $\lceil m(1-\alpha)\rceil^{\text{th}}$ order statistic of $p^{\delta_U-\delta_L}_{n^{(2)},q,\boldsymbol{u}_r^{(0)}}$. We investigate the performance and computational efficiency of this approach when considering study design for the illustrative example in Section 5.5.3.

5.5.3 Scalable Design for the Illustrative Example

We made most choices required to design a study for the illustrative example in previous sections. In Section 5.4.2, we specified design priors $p_D(\boldsymbol{\eta}_1)$ and $p_D(\boldsymbol{\eta}_2)$ for the conditional multinomial probabilities defined in (5.4) to obtain a prior $p_D(\boldsymbol{\eta}) = p_D(\boldsymbol{\eta}_1) \times p_D(\boldsymbol{\eta}_2)$ for $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. We also chose the interval $(\delta_L, \delta_U) = (-0.5, \infty)$ in Section 5.3 along with DIR$(0.8, 0.8, 0.8, 0.8, 0.8)$ analysis priors for $\boldsymbol{p}_j,\ j = 1, 2$. This Dirichlet prior assumes that the variables $\{Z_{jv}\}_{v=1}^{w-1}$ defined in (5.4) are independent a priori. Based on Chapter 4, we know the posterior of $\{Z_{jv}\}_{v=1}^{w-1}$ cannot retain positive or negative dependence given enough data, so this choice is sensible.

Furthermore, the regions $\mathcal{G} = (-0.3, -0.1)$ and $\mathcal{R} = (-0.55, -0.5)$ were selected in Section 5.4.2 in recognition of the interval (δ_L, δ_U) and the anticipated effect size for the study. We define design priors for Algorithm 5.2 of $p_{D_1}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta}|\theta \sim \mathcal{U}(\mathcal{G}))$ and $p_{D_0}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta}|\theta \sim \mathcal{U}(\mathcal{R}))$. We described how to sample from those priors in Section 5.4.1. We use $m = 8192$ and $m_0 = 512$ as recommended in Section 5.5.2. To define operating characteristics, we choose $\alpha = 0.05$ and $\beta = 0.2$ for illustration. We consider $q = 1.25$ to reflect the reference group ($j = 2$) having roughly 25% more observations than the comparison group ($j = 1$) in Section 5.3. With Algorithm 5.2, we used the modified version of Algorithm 5.1 presented in Appendix D.3.1 that accommodates departures from the approximate normality of $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$.

For these inputs, Algorithm 5.2 returned an optimal design characterized by $(n, \gamma) = (111, 0.9341)$. Figure 5.3 visualizes the distributions of the $m = 8192$ logits of $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r^{(1)}}$ and $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r^{(0)}}$ respectively used to compute confirmatory estimates for power (green) and the type I error rate (red) at $n = 111$. We visualize the distributions of the logits of the posterior probabilities since many of the posterior probabilities from the green region are very close to 1. We note that the green and red curves do not precisely estimate sampling distributions of posterior probabilities since randomized Sobol' sequences induce negative dependence between points in the unit hypercube. However, we can use the

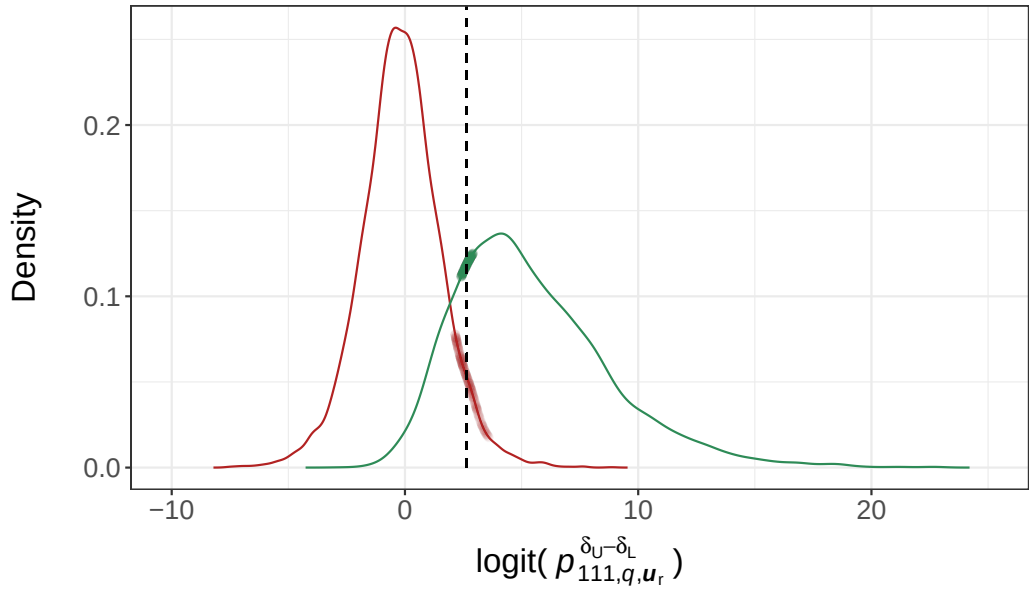


Figure 5.3: Distributions of the $m = 8192$ logits of $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r}$ for the confirmatory estimates of power (green) and the type I error rate (red) at $n = 111$. The $m_0 = 512$ logits of $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r}$ used to assess $n = 111$ are plotted with jitter. The dotted line denotes the logit of the recommended $\gamma = 0.9341$.

posterior probabilities depicted in Figure 5.3 to compute power and the type I error rate as discussed in Section 1.2.3. The $m_0 = 512$ logits of $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}^{(1)}_r}$ and $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}^{(0)}_r}$ used to assess the operating characteristics for $n = 111$ are depicted as jittered points in Figure 5.3. These points visualize the segmented nature of how we explore sampling distributions of posterior probabilities in Algorithm 5.2.

For validation purposes, we repeated the sample size calculation from the previous paragraph with a modified version of Algorithm 5.2. This modified version uses binary search to explore entire sampling distributions for *each* sample size n considered with all points from the same Sobol' sequences used in the previous calculation. We obtained the *same* optimal design when thoroughly exploring the sampling distributions of posterior probabilities using that nontargeted approach. We repeated the process of determining the optimal design for the illustrative example with both methods 1000 times using different Sobol' sequences $\{\boldsymbol{u}^{(1)}_r\}^m_{r=1}$ and $\{\boldsymbol{u}^{(0)}_r\}^m_{r=1}$. We obtained the exact same optimal design using both methods in each of the 1000 repetitions.

Algorithm 5.2 took roughly 30 seconds with one core on a standard laptop to return

an optimal design for the illustrative example. The modified version of Algorithm 5.2 that explored entire sampling distributions of posterior probabilities took approximately 95 seconds to obtain the same results. The discrepancy in runtime between design methods with sampling distribution segments and those that consider entire sampling distributions increases with the recommended sample size n. To consider a range of B consecutive sample sizes with standard binary search, we need to thoroughly explore the sampling distributions of posterior probabilities at $\mathcal{O}(\log_2 B)$ values of n. Regardless of the magnitude of the sample size recommendation, Algorithm 5.2 only requires that we thoroughly explore the sampling distributions at three samples sizes – the final of which is used to obtain confirmatory estimates for power and the type I error rate. In Appendix D.3.2, we also illustrate that using Sobol' sequences instead of pseudorandom ones allows us to estimate the optimal (n, γ) combination with the same precision using three times fewer simulation repetitions.

5.6 Contour Plots for Design Criteria Exploration

Although Algorithm 5.2 returns the (n, γ) combination that minimizes the sample size n while satisfying the criteria in (5.2) and (5.3), practitioners may want to explore multiple designs that are similar to the optimal one. The sampling distributions of posterior probabilities corresponding to $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ are thoroughly explored at three sample sizes in Algorithm 5.2: $n^{(0)}, n^{(1)}$, and $n^{(2)}$. These sample sizes can be ordered such that $n^{((0))} < n^{((1))} < n^{((2))}$. We approximate the sampling distributions for sample sizes less than $n^{((1))}$ using the linear approximations to $\text{logit}(p^{\delta_U - \delta_L}_{n,q,\boldsymbol{u}_r^{(h)}})$ informed by the posterior probabilities estimated at the sample sizes $n^{((0))}$ and $n^{((1))}$. For sample sizes greater than $n^{((1))}$, we use linear approximations informed by the estimated probabilities at $n^{((2))}$ instead of $n^{((0))}$. We use contour plots to synthesize these approximations to the sampling distributions. These plots visualize how changes to n and the critical value γ impact power and the type I error rate. If it is not feasible to collect a sample of the recommended size, we can use these plots to inform the choice of an alternative (n, γ) combination.

The left column of Figure 5.4 illustrates the contour plots with respect to the type I error rate and power for the sample size calculation in Section 5.5.3. These contour plots are available with a single application of our methodology. To assist with interpretation, the green contour corresponding to power of $1 - \beta$ and the red contour corresponding to a type I error rate of α are depicted on both plots. The criteria in (5.2) and (5.3) are respectively satisfied for the regions of the (n, γ)-space that are below the green contour and above the red contour. The optimal design characterized by $(n, \gamma) = (111, 0.9341)$ is

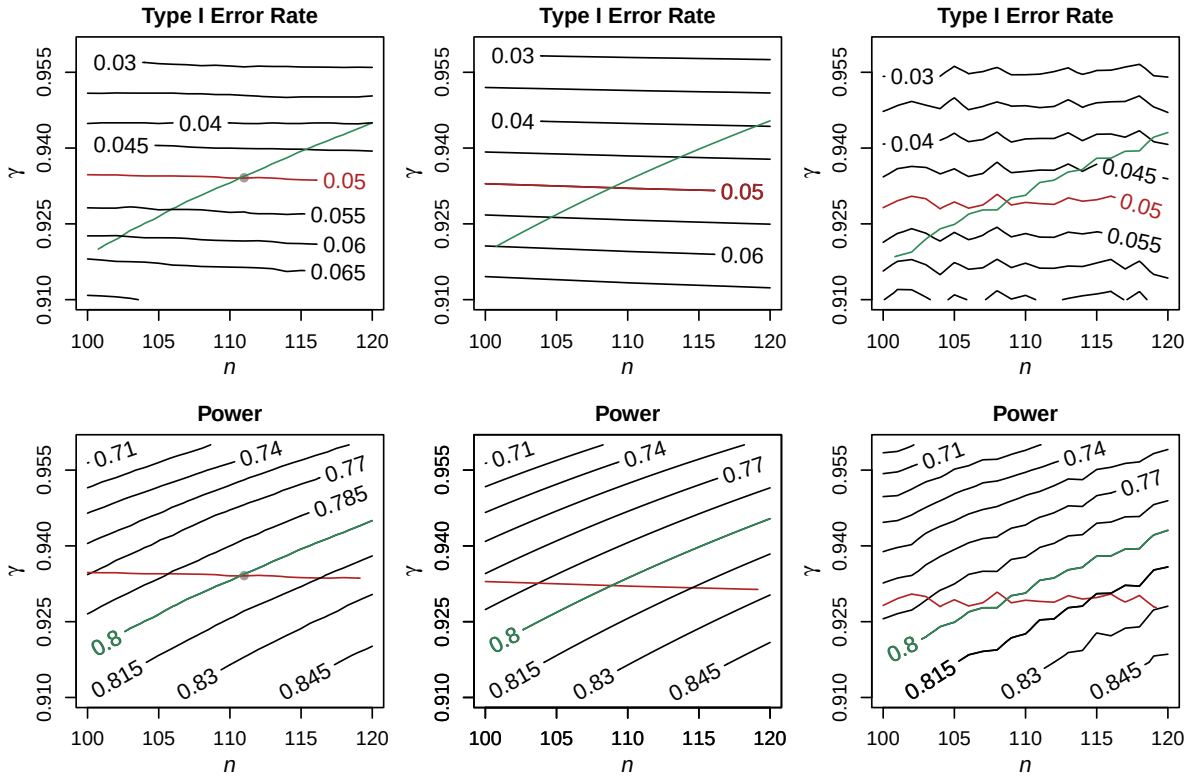


Figure 5.4: Left: Contour plots for the type I error rate and power for one sample size calculation with the optimal (n, γ) combination in grey. Center: Averaged contour plots from 1000 sample size calculations. Right: Contour plots estimated by simulating data.

depicted by the grey point. The optimal sample size of $n = 111$ is the smallest $n \in \mathbb{Z}^+$ that is to the right of the intersection of the red and green contours.

To gain insight into how our method performs under repeated simulation, we averaged contour plots corresponding to the 1000 repetitions of the sample size calculation from Section 5.5.3. These plots are given in the center column of Figure 5.4, but they take 1000 times as long to generate as the left plots and are not feasible to create in practice. Based on these plots, the smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours is 109. This discrepancy between $n = 111$ and 109 highlights that the optimal design differs slightly for each simulation repetition. The contour plots in the right column of Figure 5.4 were created by simulating $m = 81920$ samples from the prior predictive distributions for $n = \{100, 101, ..., 120\}$ following the process detailed in Section 5.2. The contours in the right plots are jagged since $q = 1.25 \notin \mathbb{Z}$. The linear approximations to

$\text{logit}(p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r^{(h)}})$ used to create the first two columns of plots in Figure 5.4 do not ensure that $n_2 \in \mathbb{Z}$ when $n \notin \{n^{((0))}, n^{((1))}, n^{((2))}\}$. Nevertheless, the plots in the center and right columns are similar, which is a direct consequence of Lemma 5.1 and the consistency of the power and type I error rate estimates at $n^{((0))}, n^{((1))}$, and $n^{((2))}$ via Corollary 5.1. The smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours in the right plots is also $n = 109$. Moreover, the fact that the center and right columns of Figure 5.4 do not differ much from the left column builds confidence in the single-application contour plots. In Appendix D.3.3, we illustrate why the modified version of Algorithm 5.1 used here yields superior performance for the illustrative example.

5.7 Discussion

In this chapter, we developed a framework for scalable design with posterior-based operating characteristics – namely power and the type I error rate – that determines optimal sample sizes and decision criteria. The scalability of this framework stems from mapping posterior probabilities to low-dimensional hypercubes and using this mapping to explore segments of sampling distributions of posterior probabilities at most sample sizes considered. That targeted exploration approach substantially reduces the number of simulation repetitions required to design posterior analyses, making them much more attractive and accessible to practitioners who want to control type I and II error. The posterior probabilities used to determine the optimal sample size and decision criteria can also be repurposed to efficiently and helpfully investigate various sample sizes and decision criteria using contour plots.

Our proposed methods are broadly applicable and could radically reframe how (conduits for) data are simulated in efficient study design. They could be extended in many aspects to accommodate more complex designs, including generalizations of the design framework with additional covariates presented in Appendix D.5. Furthermore, future work could consider design methods with sampling distribution segments that account for sequential analyses allowing for early termination or the multiple comparisons problem more generally. It may also be of interest to use these methods to design studies based on the precision of an interval estimate (as overviewed in Section 6.2.3) or maximizing the expectation of a utility function. In any of these settings, it would be pertinent to determine whether low-discrepancy sequences could be combined with targeted exploration approaches to prompt scalable design methods. Work on targeted exploration of the unit hypercube could even be applied to make simulation-based methods more accessible in non-design settings.

Chapter 6

Discussion

6.1 Summary

The purpose of this thesis was to develop a broad suite of methods for the scalable design of two-group comparisons. Two-group comparisons are routinely conducted in many settings including clinical, manufacturing, and corporate contexts. These comparisons are carried out using both frequentist and Bayesian statistical methods. The design methods proposed in this thesis accommodate comparisons that assess superiority, noninferiority, and practical equivalence. These comparisons can be based on differences or ratios, and our proposed methods can be simplified to design studies with a single group of data.

The design methods proposed in this thesis are unified by their efficient use of sampling distribution segments. The novelty and computational efficiency of design with sampling distribution segments is practically important. Standard design methods with flexible statistical models use simulation to assess the operating characteristics of hypothesis tests by estimating *entire* sampling distributions. However, we need only estimate particular quantiles of sampling distributions to determine whether a given sample size prompts a suitable study power and type I error rate. The research in this thesis was directed toward two objectives. First, we needed to automate the process to select sampling distribution segments for arbitrary designs. Second, we needed to ensure that the operating characteristics for given sample sizes were reliably assessed using sampling distribution segments. We achieved these objectives in several contexts, each of which required tailored methodology.

In Chapter 2, we applied design with sampling distribution segments in frequentist settings. Such sampling distribution segments expedited power analysis for designs where

exact pivotal quantities do not exist, with an emphasis on two-group equivalence tests with unequal variances. The goal of the work in this chapter was to provide a straightforward context in which sufficient statistics could be mapped to the unit hypercube. By leveraging root-finding algorithms for points in the unit hypercube, we automated the process of selecting sampling distribution segments. The rejection regions for these hypothesis tests bound the type I error rate. Formal investigation of the test statistics as a function of the sample size justified that power could be reliably assessed with sampling distribution segments – prompting unbiased sample size recommendations.

Chapter 3 extended design with sampling distribution segments to Bayesian settings in a preliminary context that mirrored frequentist power analysis. In that context, it was assumed that data were generated from statistical models with known, fixed parameter values for design purposes. The purpose of this work was to provide realistic scenarios where it was sensible to map maximum likelihood estimates that may or may not be sufficient statistics to the unit hypercube. Root-finding algorithms were again used to select sampling distribution segments. Since the posterior probabilities prompted by our mappings were generally increasing functions of the sample size, we could reliably assess power with sampling distribution segments. Our sample size recommendations were therefore consistent under the conditions for the BvM theorem, and our method for power curve approximation was orders of magnitude faster than conventional power curve estimation for Bayesian hypothesis tests. The type I error rate was not considered in this work, but this limitation was addressed with the methodology in Chapter 5.

In Chapter 4, we considered how prior dependence structures impact posterior distributions. We proved that under broad conditions, the posterior cannot retain many flexible prior dependence structures that arise from leveraging copula models in large-sample settings. We also clarified how the choice of prior copula can and cannot impact the posterior distribution in terms of asymptotic convergence of the posterior mode. The objective of this work was to help practitioners determine whether eliciting complicated prior dependence structures aligns with their objectives for posterior analysis. Many Bayesian design settings require that we specify different priors to generate and analyze data, so the objective of this work is relevant when applying design with sampling distribution segments. To achieve this objective, we contextualized our theoretical results by discussing various goals for prior specification.

Chapter 5 generalized Bayesian design with sampling distribution segments to more comprehensive settings. The goal of this work was to account for uncertainty in the parameter values used to generate the data while formally considering type I error rates. This framework considered two sampling distributions of posterior probabilities to select a sample size and critical value for the hypothesis test. This design framework is more com-

prehensive than the one from Chapter 3 but requires that practitioners make additional choices. We selected sampling distribution segments using linear approximations to the logits of the posterior probabilities prompted by our mappings to the unit hypercube. We proved that these linear approximations are theoretically valid in large-sample settings, so power and the type I error rate can be reliably assessed using the resulting sampling distribution segments. We repurposed the posterior probabilities computed in that approach to efficiently investigate various sample sizes and decision criteria using contour plots.

6.2 Extensions

6.2.1 Design of Sequential Analyses

The work in Chapters 2, 3, and 5 can be extended to accommodate sequential experiments that allow for early termination. For example, stopping for superiority or futility is common in adaptive designs for clinical trials. Sequential designs consider the overall power to reject H_0 across all analyses: interim or final. To obtain overall power, we require the joint sampling distribution of the test statistics across all analyses to aggregate their marginal powers. For power analysis in Chapters 2 and 3, each point in the unit hypercube was mapped to sufficient statistics or maximum likelihood estimates via CDF inversion. Since the CDFs depend on n, we exploited this mapping to obtain the smallest sample size n at which H_0 was rejected using root-finding algorithms. We could multiply the dimension of each point in the sequence by the maximum number of planned analyses to recommend a sample size for each potential analysis.

Here, we provide a sketch of how such an approach could be implemented for the two-group equivalence tests with unequal variances considered in Chapter 2. We suppose that the study is to consist of one interim and one final analysis for illustration. We use points from a hypercube of dimension $d = 8$. We could simulate sample means and variances for the interim analysis using CDF inversion and the first four coordinates of the hypercube: $\bar{y}_{1r,1}^{(n,q)}$, $\bar{y}_{2r,1}^{(n,q)}$, $s_{1r,1}^{2\ (n,q)}$, $s_{2r,1}^{2\ (n,q)}$. The first component of the subscript denotes the group number, the second component denotes the point $\{\boldsymbol{u}_r\}_{r=1}^m$, and the third component denotes the first (interim) analysis. In this case, we would need to generate both components of $\bar{d}_{r,1}^{(n,q)} = \bar{y}_{1r,1}^{(n,q)} - \bar{y}_{2r,1}^{(n,q)}$. We could use CDF inversion and the last four coordinates of the hypercube to simulate the sample statistics corresponding to the period between the interim and final analysis: $\bar{y}_{1r,2}^{(n,q)}$, $\bar{y}_{2r,2}^{(n,q)}$, $s_{1r,2}^{2\ (n,q)}$, $s_{2r,2}^{2\ (n,q)}$.

We suppose that we plan to collect the same number of observations in both periods of the study. With that information, we can calculate summary statistics for the final

analysis that maintain the desired level of dependence with the interim analysis. The mean difference $\bar{d}_{r,F}^{(n,q)}$ is $0.5\bar{d}_{r,1}^{(n,q)} + 0.5\bar{d}_{r,2}^{(n,q)}$, where the subscript F denotes that this statistic corresponds to the final analysis. The sample variance for the final analysis in each group $j = 1$ and 2 is

$$s_{jr,F}^{2\ (n,q)} = \frac{0.5n_j s_{jr,1}^{2\ (n,q)} + 0.5n_j s_{jr,2}^{2\ (n,q)}}{n_j - 1} + \frac{0.25n_j^2(\bar{y}_{jr,1}^{(n,q)} - \bar{y}_{jr,2}^{(n,q)})^2}{n_j(n_j - 1)}, \tag{6.1}$$

where $n_1 = n$ and $n_2 = qn$. The variance in (6.1) follows from standard results involving conditional normal distributions.

Given $\bar{d}_{r,1}^{(n,q)}$, $s_{1r,1}^{2\ (n,q)}$, and $s_{2r,1}^{2\ (n,q)}$, we could propose a process similar to Algorithm 2.2 to obtain a sample size recommendation for the interim analysis corresponding to each point $\{\boldsymbol{u}_r\}_{r=1}^m$. We could similarly leverage $\bar{d}_{r,F}^{(n,q)}$, $s_{1r,F}^{2\ (n,q)}$, and $s_{2r,F}^{2\ (n,q)}$ to obtain sample size recommendations for the final analysis. We note that the triangular rejection regions for the two analyses may differ substantially if the significance level α_1 for the interim analysis differs from α_F used for the final one. We suppose that the m sample size recommendations for each analysis are stored in the vectors `sampInterim` and `sampFinal`. Preliminary simulations suggest that the empirical CDF of `pmax(2*sampInterim, sampFinal)` accurately estimates the power curve as a function of the sample size n for the final analysis.

More formal investigation is still required to generalize this process to sequential experiments with more than one interim analysis and analyses where fewer observations are collected in earlier stages of the experiment. Moreover, we must also consider how to extend power curve approximation for the Bayesian hypothesis tests from Chapter 3 where low-dimensional sufficient statistics cannot be generated for each phase of a sequential analysis. More sophisticated methods to map posterior probabilities to the unit hypercube for models that do not belong to the exponential family in the presence of prior misspecification are also discussed in Appendix D.4.3. The formalization of such methods could help extend the design framework with additional covariates from Appendix D.5 to accommodate Bayesian generalized linear models.

The design of sequential experiments requires that we consider sampling distributions for each planned analysis. The computational savings associated with using sampling distribution segments are therefore compounded in these settings. However, it may be inappropriate to explore the unit hypercube using high-dimensional low-discrepancy sequences if we plan to conduct a large number of interim analyses. When we must simulate many sufficient statistics or maximum likelihood estimates to design such studies, we may want to consider simulation based on pseudorandom sequences.

The work in Chapter 5 considers both power and the type I error rate. In that chapter,

each point $\boldsymbol{u}_r$ coincides with a posterior probability that dictates whether to reject H_0. To accommodate sequential testing, we could approximate the logit of the posterior probability for each planned analysis as a linear function of the sample size. This process would prompt a collection of linear functions corresponding to each point $\boldsymbol{u}_r$. Future work could use these collections of linear functions to assess the operating characteristics of sequential designs using sampling distribution segments.

6.2.2 Design in a Nonparametric Framework

Current design methods for nonparametric hypothesis tests are computationally intensive because they rely on naïve simulation instead of leveraging the structure of the tests. Analytical power analysis for nonparametric tests in frequentist settings tends to leverage large-sample normal approximations (Shieh et al., 2007); however, such tests are commonly used when asymptotic results are unsuitable. Exploring subspaces of the unit hypercube that correspond to sampling distribution segments could yield fast design methods for rank-sum tests that account for the exact distributions of the test statistics.

We cannot generate sufficient statistics or maximum likelihood estimates in a nonparametric framework, but we could generate data conditional on low-discrepancy sequences of sample totals. This data generation process would be fast for the gamma distribution. We illustrate how to use such a process to obtain a sample $\{Y_i\}_{i=1}^n$ from the GAMMA(α, λ) model conditional on $\sum_{i=1}^n Y_i = \sum_{i=1}^n y_i$ in Algorithm 6.1, where Lines 3 to 5 leverage results from Devroye (2006).

Algorithm 6.1 Procedure to Generate Gamma Data Conditional on Sample Totals

1: **procedure** GENERATEGAMMA(n, α, λ)
2: Generate $\sum_{i=1}^n y_i \sim$ GAMMA$(n\alpha, \lambda)$.
3: Generate $x_i \sim$ GAMMA$(\alpha, 1)$ for $i = 1, ..., n$.
4: **for** i in 1:n **do**
5: Let $y_i = x_i \dfrac{\sum_{j=1}^n y_j}{\sum_{j=1}^n x_j}$.
6: **return** $\{y_i\}_{i=1}^n$ as a sample from GAMMA(α, λ) conditional on $\sum_{i=1}^n Y_i = \sum_{i=1}^n y_i$.

The fast procedure described in Algorithm 6.1 extends to mixture gamma models, which accommodate multimodality and skewness inherent to nonparametric settings. The sample totals simulated in Line 2 over repeated implementation of Algorithm 6.1 would exhibit negative dependence if generated using low-discrepancy sequences. These sample totals are

insufficient statistics based on CDF inversion with the points $\{\boldsymbol{u}_r\}_{r=1}^m$. Preliminary simulations suggest that this negative dependence is partially retained in test statistics based on rank sums, reducing the value of m required to estimate power. To make these methods more scalable, we could investigate approaches to select subspaces of the unit hypercube – and corresponding sampling distribution segments – in nonparametric contexts.

6.2.3 Design with Precision Criteria

Statistically insignificant study results are scientifically valuable if the confidence or credible interval for $\theta_1 - \theta_2$ is sufficiently narrow. Precision criteria aim to select the smallest sample size n such that the length of a credible or confidence interval is at most l with probability at least Γ (Joseph and Belisle, 1997; De Santis and Pacifico, 2004). Precision criteria for interval estimates do not incentivize researchers to repurpose their data until they obtain a statistically significant result. These criteria are underused, and the American Statistical Association's recent official statement (Wasserstein and Lazar, 2016) underscores their promotion of sound scientific practice. In past work, we defined an analogue to the power curve for the length criterion, where the Γ-quantile of this length curve is the smallest sample size n such that the interval estimate for $\theta_1 - \theta_2$ has length of at most l with probability at least Γ (Stevens and Hagar, 2022). We observed this length curve to be approximately normal for large n via simulation in that work.

In Bayesian settings, we have formally proven that the length curve is approximately normal using the BvM theorem in an unpublished manuscript (Hagar and Stevens, 2023). This paper was included in the proposal for this thesis, but we decided to focus on design methods with power and the type I error rate in the final version. We note that the length curve is also useful for frequentist design. As such, we believe that this work would best be repurposed as a paper that presents unified methodology for design with precision criteria in Bayesian and frequentist settings.

Future research on this topic is required because the length curve is not guaranteed to be a strictly nondecreasing function of n (Hagar and Stevens, 2023). This decreasing behaviour suggests that we may require simulation to precisely estimate the length curve. Instead of estimating entire sampling distributions of confidence or credible interval lengths at each sample size n considered, it would be efficient to leverage sampling distribution segments. The decreasing behaviour of the length curve also complicates selecting subspaces from the unit hypercube and sampling distribution segments, but its approximate normality in large-sample settings could mitigate this complication.

6.2.4 Design with Computational Posterior Approximation

The work in Chapters 3 and 5 rely on large-sample approximations to the posterior (see e.g., Gelman et al. (2020)). While monotonic transformations of the relevant parameters can improve the quality of these approximations for moderate n, these approximations are not reliable for small sample sizes. Future research could combine the techniques to select sampling distribution segments from this thesis that are based on analytical posterior approximation with computational approximation methods. For instance, we may not want to assess posterior-based operating characteristics using analytical posterior approximations, but we might still use these approximations to *select* sampling distribution segments. In that case, we could combine the sufficient statistics or maximum likelihood estimates mapped to the analytical posteriors with MCMC methods to assess power and the type I error rate. We could also investigate how the segmentation of sampling distributions enhances methods that leverage computational posterior approximation to fit parametric models to sampling distributions of posterior probabilities (Golchi, 2022; Golchi and Willard, 2023).

6.2.5 Software Development

We developed the `dent` package in R to implement the methods for empirical power analysis and power curve approximation proposed in Chapter 2 (Hagar and Stevens, 2024a). Moreover, the R code to reproduce the numerical studies in this thesis has been made available on Github at the following links.

- Chapter 2: `https://github.com/lmhagar/BioDesignSegments`
- Chapter 3: `https://github.com/lmhagar/BayesianPower`
- Chapter 4: `https://github.com/lmhagar/PosteriorRamifications`
- Chapter 5: `https://github.com/lmhagar/PosteriorBasedOCs`

However, we have not formally published any R packages or R Shiny apps to make the Bayesian design methods with sampling distribution segments proposed in this thesis more accessible. While we proposed computationally efficient methods in this thesis, our methods require substantial mathematical overhead to approximate posteriors analytically and implement the delta method. Practitioners would also require considerable programming skills to implement our methods. It would therefore be valuable to develop software that

automates the implementation of the Bayesian design methods proposed in this thesis, particularly for common statistical models that are members of the exponential family.

Moreover, we developed a method to leverage – and improve upon – Gibbs sampling with the `rjags` package in R (Plummer, 2019). That method can be applied with statistical models that belong to the exponential family. We used that method in Section 3.6.1 when comparing our power curves based on analytical posterior approximation to those obtained by simulating data and using MCMC methods. As such, we did not thoroughly discuss that method to improve upon the `rjags` package in this thesis. Nevertheless, we could develop that method into a formal R package in future work. This extension would be relevant to this thesis if implementing the computational extensions for design with sampling distribution segments detailed in Section 6.2.4.

References

Balas, E. and R. Jeroslow (1972). Canonical cuts on the unit hypercube. *SIAM Journal on Applied Mathematics 23*(1), 61–69.

Barone, R. and L. Dalla Valle (2023). Bayesian nonparametric modeling of conditional multidimensional dependence structures. *Journal of Computational and Graphical Statistics 32*, 1–10.

Bartlett, M. S. (1934). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh 53*, 260–283.

Bedford, T. and R. M. Cooke (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031 – 1068.

Berry, S. M., B. P. Carlin, J. J. Lee, and P. Muller (2011). *Bayesian Adaptive Methods for Clinical Trials*. CRC press.

Borchers, H. W. (2021). *pracma: Practical Numerical Math Functions*. R package version 2.3.3.

Braaten, E. and G. Weller (1979). An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration. *Journal of Computational Physics 33*(2), 249–258.

Brent, R. P. (1973). An algorithm with guaranteed convergence for finding the minimum of a function of one variable. *Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, NJ*, 61–80.

Brutti, P., F. De Santis, and S. Gubbiotti (2014). Bayesian-frequentist sample size determination: A game of two priors. *Metron 72*(2), 133–151.

Carlin, B. P. and T. A. Louis (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (2 ed.). Springer.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician 39*(2), 83–87.

Casella, G., C. P. Robert, and M. T. Wells (2004). Generalized accept-reject sampling schemes. *Lecture Notes-Monograph Series*, 342–347.

Casement, C. J. and D. J. Kahle (2018). Graphical prior elicitation in univariate models. *Communications in Statistics - Simulation and Computation 47*(10), 2906–2924.

Casement, C. J. and D. J. Kahle (2023). The phoropter method: A stochastic graphical procedure for prior elicitation in univariate data models. *Journal of the Korean Statistical Society 52*(1), 60–82.

Chaloner, K. (1996). Elicitation of prior distributions. In *Bayesian Biostatistics*, pp. 141–156. Marcel Dekker, New York.

Chaloner, K., T. Church, T. A. Louis, and J. P. Matts (1993). Graphical elicitation of a prior distribution for a clinical trial. *Journal of the Royal Statistical Society: Series D (The Statistician) 42*(4), 341–353.

Chow, S. C. and J. P. Liu (2008). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Chapman and Hall/CRC.

Chow, S. C., J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research*. Chapman & Hall/CRC.

Clemen, R. T. and T. Reilly (1999). Correlations and copulas for decision and risk analysis. *Management Science 45*(2), 208–224.

Connor, R. J. and J. E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association 64*(325), 194–206.

Cranley, R. and T. N. L. Patterson (1976). Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis 13*(6), 904–914.

Crispino, M. and I. Antoniano-Villalobos (2023). Informative priors for the consensus ranking in the Bayesian Mallows model. *Bayesian Analysis 18*(2), 391–414.

Dannenberg, O., H. Dette, and A. Munk (1994). An extension of Welch's approximate t-solution to comparative bioequivalence trials. *Biometrika 81*(1), 91–101.

De Santis, F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 170*(1), 95–113.

De Santis, F. and M. P. Pacifico (2004). Two experimental settings in clinical trials: Predictive criteria for choosing the sample size in interval estimation. In *Applied Bayesian Statistical Studies in Biology and Medicine*, pp. 109–130. Springer.

Demarta, S. and A. J. McNeil (2005). The t copula and related copulas. *International statistical review 73*(1), 111–129.

Devroye, L. (2006). Nonuniform random variate generation. *Handbooks in Operations Research and Management Science 13*, 83–121.

Eberly, L. E. and B. P. Carlin (2000). Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in medicine 19*(17-18), 2279–2294.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM.

Elfadaly, F. G. and P. H. Garthwaite (2017). Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Statistics and Computing 27*(2), 449–467.

Entacher, K. (1998). Bad subsequences of well-known linear congruential pseudorandom number generators. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 8*(1), 61–70.

Faure, H. (1982). Discrépance de suites associées à un système de numération (en dimension s). *Acta arithmetica 41*(4), 337–351.

FDA (2003). Guidance on bioavailability and bioequivalence studies for orally administered drug products — General considerations. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD.

FDA (2006). Guidance for industry - Bioequivalence guidance. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD.

FDA (2019). Adaptive designs for clinical trials of drugs and biologics — Guidance for industry. Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Rockville, MD.

Feroze, N. and M. Aslam (2021). Comparison of improved class of priors for the analysis of the Burr type VII model under doubly censored samples. *Hacettepe Journal of Mathematics and Statistics 50*(5), 1509–1533.

Fisher, R. A. (1934). *Statistical Methods for Research Workers* (5th ed.). Oliver and Boyd, Edinburgh and London.

Fishman, G. S. and L. R. Moore, III (1986). An exhaustive analysis of multiplicative congruential random number generators with modulus 2ˆ31-1. *SIAM Journal on Scientific and Statistical Computing 7*(1), 24–45.

Fox, B. L. (1986). Algorithm 647: Implementation and relative efficiency of quasirandom sequence generators. *ACM Transactions on Mathematical Software (TOMS) 12*(4), 362–376.

Freedman, L. and D. Spiegelhalter (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Journal of the Royal Statistical Society: Series D (The Statistician) 32*(1-2), 153–160.

Fúquene, J., M. Steel, and D. Rossell (2019). On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society Series B: Statistical Methodology 81*(5), 809–837.

Gao, C., A. W. van der Vaart, and H. H. Zhou (2020). A general framework for Bayes structured linear models. *The Annals of Statistics 48*(5), 2848 – 2878.

Garthwaite, P. H., S. A. Al-Awadhi, F. G. Elfadaly, and D. J. Jenkinson (2013). Prior distribution elicitation for generalized linear and piecewise-linear models. *Journal of Applied Statistics 40*(1), 59–75.

Garthwaite, P. H., J. B. Kadane, and A. O'Hagan (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association 100*(470), 680–701.

Gatti, D. M., P. Simecek, L. Somes, C. T. Jeffrey, M. J. Vincent, K. Choi, X. Chen, G. A. Churchill, and K. L. Svenson (2017). The effects of sex and diet on physiology and liver gene expression in diversity outbred mice. *bioRxiv*.

Gelfand, A. E. and S. K. Sahu (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association 94*(445), 247–253.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2020). *Bayesian Data Analysis*. CRC Press.

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

Gentle, J. E. (2003). *Random Number Generation and Monte Carlo Methods* (2 ed.). Springer.

Ghosal, S., J. K. Ghosh, and T. Samanta (1995). On convergence of posterior distributions. *The Annals of Statistics 23*(6), 2145–2152.

Golchi, S. (2022). Estimating design operating characteristics in Bayesian adaptive clinical trials. *Canadian Journal of Statistics 50*(2), 417–436.

Golchi, S. and J. Willard (2023). Estimating the sampling distribution of test-statistics in Bayesian clinical trials. *arXiv preprint arXiv:2306.09151*.

Gruman, J. A., R. Cribbie, and C. A. Arpin-Cribbie (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods 6*(1), 133–140.

Gubbiotti, S. and F. De Santis (2011). A Bayesian method for the choice of the sample size in equivalence trials. *Australian & New Zealand Journal of Statistics 53*(4), 443–460.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science 20*(2), 111–140.

Gustafson, P. (2012). On the behaviour of Bayesian credible intervals in partially identified models. *Electronic Journal of Statistics 6*, 2107 – 2124.

Hagar, L. and N. T. Stevens (2023). Fast sample size determination for Bayesian equivalence tests. https://arxiv.org/abs/2306.09476.

Hagar, L. and N. T. Stevens (2024a). dent: Design of equivalence and noninferiority tests. R package version 0.0.1. https://github.com/lmhagar/dent.

Hagar, L. and N. T. Stevens (2024b). Posterior ramifications of prior dependence structures. Under review at *Statistical Science*.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik 2*, 84–90.

Hickernell, F. J., H. S. Hong, P. L'Écuyer, and C. Lemieux (2000). Extensible lattice sequences for quasi-Monte Carlo quadrature. *SIAM Journal on Scientific Computing 22*(3), 1117–1138.

Hickernell, F. J. and H. Niederreiter (2003). The existence of good extensible rank-1 lattices. *Journal of Complexity 19*(3), 286–300.

Hofert, M. and C. Lemieux (2020). *qrng: (Randomized) Quasi-Random Number Generators*. R package version 0.0-8.

Instituto Nacional de Estadística, Geografía e Informática [National Institute of Statistics, Geography, and Informatics] (INEGI) (2019). Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2018 Nueva serie [National Survey of Household Income and Expenses. New edition 2018]. www.inegi.org.mx/programas/enigh/nc/2018/#Datos_abiertos.

Instituto Nacional de Estadística, Geografía e Informática [National Institute of Statistics, Geography, and Informatics] (INEGI) (2021). Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020 Nueva serie [National Survey of Household Income and Expenses. New edition 2020]. www.inegi.org.mx/programas/enigh/nc/2020/#Datos_abiertos.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Irizarry, R. A. and A. Gill (2023). *dslabs: Data Science Labs*. R package version 0.7.6.

Jan, S.-L. and G. Shieh (2017). Optimal sample size determinations for the heteroscedastic two one-sided tests of mean equivalence: Design schemes and software implementations. *Journal of Educational and Behavioral Statistics 42*(2), 145–165.

Jan, S.-L. and G. Shieh (2020). On the extended Welch test for assessing equivalence of standardized means. *Statistics in Biopharmaceutical Research 12*(3), 344–351.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 31, No. 2, pp. 203–222. Cambridge University Press.

Joe, H. (1996). Families of m-variate distributions with given margins and m $(m$-1)/2 bivariate dependence parameters. *Lecture Notes - Monograph Series*, 120–141.

Joe, H. and D. Kurowicka (2011). *Dependence modeling: vine copula handbook*. World Scientific.

Johnson, S. R., G. A. Tomlinson, G. A. Hawker, J. T. Granton, and B. M. Feldman (2010). Methods to elicit beliefs for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology 63*(4), 355–369.

Jones, G. and W. O. Johnson (2014). Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician 68*(1), 42–51.

Joseph, L. and P. Belisle (1997). Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician) 46*(2), 209–226.

Kadane, J. B. (1986). Progress toward a more ethical method for clinical trials. *The Journal of Medicine and Philosophy 11*(4), 385–404.

Kadane, J. B., J. M. Dickey, R. L. Winkler, W. S. Smith, and S. C. Peters (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association 75*(372), 845–854.

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

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

Knuth, D. E. (2014). *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley Professional.

Koch, K.-R. (2007). *Introduction to Bayesian Statistics*. Springer Science & Business Media.

Kocis, L. and W. J. Whiten (1997). Computational investigations of low-discrepancy sequences. *ACM Transactions on Mathematical Software (TOMS) 23*(2), 266–294.

Korobov, A. (1959). The approximate computation of multiple integrals. In *Dokl. Akad. Nauk SSSR*, Volume 124, pp. 1207–1210.

Kroese, D. P., T. Taimre, and Z. I. Botev (2013). *Handbook of Monte Carlo Methods*. John Wiley & Sons.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science 1*(2), 270–280.

Kurowicka, D. and R. Cooke (2005). Distribution-free continuous Bayesian belief. *Modern Statistical and Mathematical Methods in Reliability 10*, 309.

Le Cam, L. and G. L. Yang (2000). *Asymptotics in Statistics: Some Basic Concepts*. Springer Science & Business Media.

L'Écuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators. *Operations Research 47*(1), 159–164.

L'Écuyer, P. (2001). Software for uniform random number generation: Distinguishing the good and the bad. In *Proceeding of the 2001 Winter Simulation Conference (Cat. No. 01CH37304)*, Volume 1, pp. 95–105. IEEE.

L'Écuyer, P. and C. Lemieux (2000). Variance reduction via lattice rules. *Management Science 46*(9), 1214–1235.

L'Écuyer, P., R. Simard, and S. Wegenkittl (2002). Sparse serial tests of uniformity for random number generators. *SIAM Journal on Scientific Computing 24*(2), 652–668.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation*. Springer Science & Business Media.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer.

Libby, P. and J. S. Pober (2001). Chronic rejection. *Immunity 14*(4), 387–397.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology 140*, 1–55.

Lui, K.-J. (2016). *Crossover Designs: Testing, Estimation, and Sample Size*. John Wiley & Sons.

Matsumoto, M. and T. Nishimura (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 8*(1), 3–30.

Metropolis, N. and S. Ulam (1949). The Monte Carlo method. *Journal of the American Statistical Association 44*(247), 335–341.

Michimae, H. and T. Emura (2022). Bayesian ridge estimators based on copula-based joint prior distributions for regression coefficients. *Computational Statistics 37*(5), 2741–2769.

Morey, R. D. and J. N. Rouder (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods 16*(4), 406–419.

Munkhuwa, V., K. Masamba, and W. Kasapila (2022). Beta carotene apparent retention dataset, Likert scale dataset, and preference ranking scale dataset. https://doi.org/10.5281/zenodo.7180722.

NCSS, LLC. (2023). Chapter 529 - Two-sample t-tests for equivalence allowing unequal variance. *Power Analysis and Sample Size (PASS) 2023 Documentation*. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Two-Sample_T-Tests_for_Equivalence_Allowing_Unequal_Variance.pdf.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Science & Business Media.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM.

Ning, S. and N. Shephard (2018). A nonparametric Bayesian approach to copula estimation. *Journal of Statistical Computation and Simulation 88*(6), 1081–1105.

Nocedal, J. and S. J. Wright (2006). *Numerical Optimization* (2 ed.). Springer.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society 4*(1), 119–130.

Plummer, M. (2019). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10.

Reimherr, M., X.-L. Meng, and D. L. Nicolae (2021). Prior sample size extensions for assessing prior impact and prior-likelihood discordance. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(3), 413–437.

Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association 82*(398), 543–546.

Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3*, 395–402.

Rusticus, S. A. and C. Y. Lovato (2014). Impact of sample size and variability on the power and type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation 19*(1), 11.

Santos, J. D. and J. M. Costa (2019). An algorithm for prior elicitation in dynamic Bayesian models for proportions with the logit link function. *Methodology and Computing in Applied Probability 21*, 169–183.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2*(6), 110–114.

Schillings, C., B. Sprungk, and P. Wacker (2020). On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numerische Mathematik 145*, 915–971.

Schindler, W. (2009). Random number generators for cryptographic applications. In *Cryptographic Engineering*, pp. 5–23. Springer.

Schuirmann, D. J. (1981). On hypothesis-testing to determine if the mean of a normal-distribution is contained in a known interval [Abstract]. *Biometrics 37*(3), 617.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics 15*, 657–680.

Schweizer, B. and A. Sklar (2011). *Probabilistic Metric Spaces*. Courier Corporation.

Seo, J. I. and Y. Kim (2022). Nonparametric prior elicitation for a binomial proportion. *Communications in Statistics-Simulation and Computation 51*(6), 2809–2821.

Shao, J. (2003). *Mathematical statistics*. Springer Science & Business Media.

Shi, H. and G. Yin (2019). Control of type I error rates in Bayesian sequential designs. *Bayesian Analysis 14*(2), 399–425.

Shieh, G., S.-L. Jan, and C.-S. Leu (2022). Exact properties of some heteroscedastic TOST alternatives for bioequivalence. *Statistics in Biopharmaceutical Research 14*(4), 651–660.

Shieh, G., S.-L. Jan, and R. H. Randles (2007). Power and sample size determinations for the Wilcoxon signed-rank test. *Journal of Statistical Computation and Simulation 77*(8), 717–724.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Smith, A. F. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician 46*(2), 84–88.

Snedecor, G. W. and W. G. Cochran (1989). *Statistical Methods* (8 ed.). Iowa State University Press.

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki 7*(4), 784–802.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Volume 13. John Wiley & Sons.

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

Srivastava, R., S. Upadhyay, and V. Shukla (2019). Subjective elicitation of hyperparameters of a conjugate Dirichlet prior and the corresponding Bayes analysis. *Communications in Statistics - Theory and Methods 48*(19), 4874–4887.

Stevens, N. T. and L. Hagar (2022). Comparative probability metrics: Using posterior probabilities to account for practical equivalence in A/B tests. *The American Statistician 76*(3), 224–237.

Student (1908). The probable error of a mean. *Biometrika 6*(1), 1–25.

Süli, E. and D. F. Mayers (2003). *An Introduction to Numerical Analysis*. Cambridge University Press.

Tartakovsky, A., I. Nikiforov, and M. Basseville (2015). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC press.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vos, R., B. Vanaudenaerde, S. E. Verleden, S. De Vleeschauwer, A. Willems-Widyastuti, D. Van Raemdonck, A. Schoonis, T. Nawrot, L. Dupont, and G. Verleden (2011). A randomised controlled trial of azithromycin to prevent chronic rejection after lung transplantation. *European Respiratory Journal 37*(1), 164–172.

Walker, E. and A. S. Nowacki (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine 26*(2), 192–196.

Wand, M. P. and M. C. Jones (1994). *Kernel Smoothing*. CRC press.

Wang, F. and A. E. Gelfand (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science 17*(2), 193–208.

Wasserstein, R. L. and N. A. Lazar (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p $<$0.05". *The American Statistician 73*(sup1), 1–19.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population varlances are involved. *Biometrika 34*(1-2), 28–35.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC.

Wiart, J., C. Lemieux, and G. Y. Dong (2021). On the dependence structure and quality of scrambled (t, m, s)-nets. *Monte Carlo Methods and Applications 27*(1), 1–26.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pp. 196–202. Springer.

Wilding, J. P., R. L. Batterham, S. Calanna, M. Davies, L. F. Van Gaal, I. Lingvay, B. M. McGowan, J. Rosenstock, M. T. Tran, T. A. Wadden, et al. (2021). Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine 384*(11), 989–1002.

Williams, C. J., K. J. Wilson, and N. Wilson (2021). A comparison of prior elicitation aggregation using the classical method and SHELF. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(3), 920–940.

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press.

Wilson, K. J. (2018). Specification of informative prior distributions for multinomial models using vine copulas. *Bayesian Analysis 13*(3), 749–766.

Wilson, K. J., F. G. Elfadaly, P. H. Garthwaite, and J. E. Oakley (2021). Recent advances in the elicitation of uncertainty distributions from experts for multinomial probabilities. *Expert Judgement in Risk and Decision Analysis*, 19–51.

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association 62*(319), 776–800.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association 66*(336), 675–685.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 32–52.

Wong, W. H. and L. Ma (2010). Optional Pólya tree and Bayesian inference. *The Annals of Statistics 38*(3), 1433 – 1459.

Wu, J., X. Wang, and S. G. Walker (2015). Bayesian nonparametric estimation of a copula. *Journal of Statistical Computation and Simulation 85*(1), 103–116.

Ye, K., Z. Han, Y. Duan, and T. Bai (2022). Normalized power prior Bayesian analysis. *Journal of Statistical Planning and Inference 216*, 29–50.

APPENDICES

Appendix A

Additional Material for Chapter 2

A.1 Further Justification for Using Root-Finding Algorithms

A.1.1 The Potential for Multiple Intersections

Here, we conduct more extensive numerical studies to justify using root-finding algorithms in our efficient approach to power curve approximation. We reconsider the illustrative example from Section 2.3.2, originally adapted from *PASS 2023* documentation (NCSS, LLC., 2023). In this example, data were generated independently and identically for groups $j = 1$ and 2 according to $\mathcal{N}(\mu_1 = 92, \sigma_1^2 = 18^2)$ and $\mathcal{N}(\mu_2 = 96, \sigma_2^2 = 15^2)$ distributions, respectively. The interval endpoints were $(\delta_L, \delta_U) = (-19.2, 19.2)$. The significance level for the test was $\alpha = 0.05$. We now extend this example to admit three scenarios. These three scenarios are defined by $(\sigma_1, \sigma_2) \in \{(16.5, 16.5), (18, 15), (19.5, 13)\}$. We considered each scenario with $q = \{1, \sigma_2/\sigma_1, \sigma_1/\sigma_2\}$.

For each scenario and q combination, we now consider sample sizes $n_1 = \{2, 3, ..., 100\}$. We require that $n_1, n_2 \geq 2$ to estimate the standard deviation for each group. For the example from Section 2.3.2, $\theta_1 - \theta_2 = \mu_1 - \mu_2 = -4$. We also consider the illustrative example where $\theta_1 - \theta_2 = \{0, -8, -12, -16\}$ with maximum n_1 values of $\{100, 200, 500, 2500\}$. As $\theta_1 - \theta_2$ approaches $\delta_L = -19.2$, we must consider larger sample sizes to approximate the entire power curve for those settings. Given values for $\theta_1 - \theta_2$, q, σ_1, and σ_2, we generated a Sobol' sequence $\boldsymbol{u}_r = (u_{1r}, u_{2r}, u_{3r}) \in [0, 1]^3$ for $r = 1, ..., m$. We used $m = 1024$ for this study. For each Sobol' sequence point, we computed $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ at

all $(n_1, n_2) = (n, \lfloor qn \rceil)$ pairs considered with the relevant $\theta_1 - \theta_2$ specification. We repeated this process 1000 times for each $\theta_1 - \theta_2$, q, σ_1, and σ_2 combination. The results from this numerical study are detailed in Table A.1. This numerical study allows us to consider (i) scenarios where $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersect more than once for a given Sobol' sequence point $\boldsymbol{u}_r$ and (ii) the nondecreasing behaviour of $se_r^{(n,q)}$ as a function of n.

The center section of Table A.1 concerns scenarios where $se_r^{(n,q)} = \Lambda_r^{(n,q)}$ have nonunique solutions. The prevalence column indicates the mean percentage of the $m = 1024$ Sobol' sequence points that had multiple solutions for $se_r^{(n,q)} = \Lambda_r^{(n,q)}$. This percentage is very low, particularly when $\theta_1 - \theta_2$ is close to the center of the equivalence region $0.5(\delta_U - \delta_L) = 0$. The prevalence of multiple intersections increases as $\theta_1 - \theta_2$ approaches $\delta_L = -19.2$, but $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ intersect only once for roughly 99% of the Sobol' sequence points when $\theta_1 - \theta_2 = -16$. For the Sobol' sequence points with nonunique solutions, the departure column details the mean value of n such that $se_r^{(n-1,q)} < \Lambda_r^{(n-1,q)}$ but $se_r^{(n,q)} > \Lambda_r^{(n,q)}$. That is, this column summarizes the mean value at which this Sobol' sequence point leaves the rejection region for the TOST procedure. This sample size is very small for all scenarios considered. In the vast majority of situations, this departure occurs at a sample size of 3 (i.e., $\boldsymbol{u}_r$ prompts a sample that is in the rejection region when $n = 2$ but not when $n = 3$).

The duration column summarizes the mean value for the smallest $\zeta \in \mathbb{Z}^+$ such that $se_r^{(n+\zeta,q)} < \Lambda_r^{(n+\zeta,q)}$ for the departing sample size n (i.e., the number of sample sizes before the sample corresponding to $\boldsymbol{u}_r$ returns to the TOST rejection region). The mean duration of these departures increases as $\theta_1 - \theta_2$ approaches $\delta_L = -19.2$ but so do the sample sizes n_1 and n_2 required to achieve the desired target power. For instance, we require n between roughly 200 and 450 to obtain 80% power for the settings where $\theta_1 - \theta_2 = -16$. Therefore, the mean duration of these departures is small with respect to the recommended sample sizes.

The right section of Table A.1 concerns the nondecreasing behaviour of $se_r^{(n,q)}$, which does not depend on the value for $\theta_1 - \theta_2$. The mean column indicates the average sample size n at which $se_r^{(n,q)}$ peaks over all simulation repetitions. Because a minimum n value of 2 is required to estimate σ_1 and σ_2, $se_r^{(n,q)}$ is a generally decreasing function of n for the majority of Sobol' sequence points. As indicated in the two rightmost columns of Table A.1, it is uncommon for $se_r^{(n,q)}$ to peak at sample sizes $n > 5$, and nondecreasing behaviour of $se_r^{(n,q)}$ is incredibly rare for $n > 10$. These results are encouraging because the nondecreasing behaviour of $se_r^{(n,q)}$ drives many of the multiple intersections between $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$. Given the results in Table A.1, we conclude that root-finding algorithms are a suitable mechanism to select sampling distribution segments.

$\theta_1 - \theta_2 = 0$		Nonunique $se_r^{(n,q)} = \Lambda_r^{(n,q)}$			argmax n for $se_r^{(n,q)}$		
Scenario	q	Prevalence	Departure	Duration	Mean	$n > 5$	$n > 10$
1	1	0.03%	3.00	1.04	2.54	2.25%	0.03%
2	1	0.03%	3.00	1.04	2.55	2.31%	0.03%
	1.2^{-1}	0.09%	3.07	1.12	2.89	3.58%	0.07%
	1.2	0.01%	3.00	1.00	2.47	1.83%	0.02%
3	1	0.04%	3.00	1.06	2.58	2.58%	0.04%
	1.5^{-1}	0.10%	3.01	1.11	2.94	5.84%	0.15%
	1.5	0.09%	3.00	1.01	2.48	2.22%	0.04%
$\theta_1 - \theta_2 = -4$		Prevalence	Departure	Duration	Mean	$n > 5$	$n > 10$
1	1	0.06%	3.00	1.76	2.54	2.25%	0.03%
2	1	0.06%	3.01	1.81	2.55	2.31%	0.03%
	1.2^{-1}	0.13%	3.31	1.61	2.89	3.58%	0.07%
	1.2	0.05%	3.00	1.71	2.47	1.82%	0.02%
3	1	0.07%	3.01	1.78	2.58	2.58%	0.04%
	1.5^{-1}	0.14%	3.31	1.65	2.94	5.84%	0.15%
	1.5	0.16%	3.00	1.34	2.48	2.21%	0.04%
$\theta_1 - \theta_2 = -8$		Prevalence	Departure	Duration	Mean	$n > 5$	$n > 10$
1	1	0.16%	3.14	3.35	2.54	2.25%	0.03%
2	1	0.17%	3.13	3.42	2.55	2.31%	0.04%
	1.2^{-1}	0.23%	3.76	3.18	2.89	3.58%	0.07%
	1.2	0.15%	3.12	3.27	2.47	1.82%	0.02%
3	1	0.18%	3.14	3.42	2.58	2.56%	0.04%
	1.5^{-1}	0.31%	4.08	3.07	2.94	5.84%	0.16%
	1.5	0.33%	3.05	2.42	2.48	2.22%	0.03%
$\theta_1 - \theta_2 = -12$		Prevalence	Departure	Duration	Mean	$n > 5$	$n > 10$
1	1	0.36%	3.46	8.03	2.54	2.25%	0.03%
2	1	0.38%	3.46	8.15	2.55	2.32%	0.03%
	1.2^{-1}	0.49%	4.47	7.78	2.89	3.58%	0.07%
	1.2	0.38%	3.39	7.34	2.47	1.82%	0.02%
3	1	0.41%	3.48	8.02	2.58	2.57%	0.04%
	1.5^{-1}	0.65%	5.06	6.88	2.94	5.84%	0.15%
	1.5	0.60%	3.26	5.89	2.48	2.22%	0.04%
$\theta_1 - \theta_2 = -16$		Prevalence	Departure	Duration	Mean	$n > 5$	$n > 10$
1	1	0.77%	4.33	35.72	2.54	2.25%	0.04%
2	1	0.79%	4.37	35.38	2.55	2.31%	0.04%
	1.2^{-1}	1.01%	5.79	33.17	2.89	3.58%	0.07%
	1.2	0.84%	4.24	31.50	2.47	1.82%	0.02%
3	1	0.86%	4.36	35.54	2.58	2.57%	0.05%
	1.5^{-1}	1.24%	6.66	29.49	2.94	5.85%	0.15%
	1.5	1.20%	4.22	25.87	2.48	2.22%	0.04%

Table A.1: Simulation results for 1000 repetitions of all scenario and q combinations for five $\theta_1 - \theta_2$ values with $m = 1024$. The center section of the table concerns nonunique solutions to $se_r^{(n,q)} = \Lambda_r^{(n,q)}$. The right section concerns nondecreasing behaviour of $se_r^{(n,q)}$.

A.1.2 The Impact of Multiple Intersections on Power Curve Approximation

As visualized in Section 2.4.4, Algorithm 2.2 approximates power curves without estimating the entire sampling distribution of test statistics for all sample sizes. The segments of the relevant sampling distributions are selected using root-finding algorithms under the assumption that the functions $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$ have a unique solution for each point $\boldsymbol{u}_r, r = 1, ..., m$. Reinitializing the root-finding algorithm in Lines 7 to 13 of Algorithm 2.2 allows us to obtain an unbiased sample size recommendation in the presence of multiple intersections. In Section 2.4.3, we conducted a numerical study with 8000 power curves for the illustrative example in which the root-finding algorithm never needed to be reinitialized. We now extend that numerical study to the expanded set of scenarios defined in Section A.1.1.

We considered the 35 scenarios from Table A.1. These scenarios detail five values for $\theta_1 - \theta_2$: $\{0, -4, -8, -12, -16\}$. For each $\theta_1 - \theta_2$ value, seven (σ_1, σ_2, q) combinations explored unequal variances and imbalanced sample sizes: $\{1 = (16.5, 16.5, 1),\ 2 = (18, 15, 1),\ 3 = (18, 15, 1.2^{-1}),\ 4 = (18, 15, 1.2),\ 5 = (19.5, 13, 1),\ 6 = (19.5, 13, 1.5^{-1}),\ 7 = (19.5, 13, 1.5)\}$. For each of these 35 scenarios, we considered eight values for the target power $1 - \beta = \{0.2, 0.3, \ldots, 0.9\}$. The remaining inputs for Algorithm 2.2 are $\alpha = 0.05$, $\delta_L = -19.2$, $\delta_U = 19.2$, and $m = 1024$ as used in Section 2.4.3. For each of the $35 \times 8 = 280$ scenario and target power combinations, we approximated 100 power curves using Algorithm 2.2.

We only needed to reinitialize the root-finding algorithm in Lines 7 to 13 of Algorithm 2.2 for four of the 2.867×10^7 points used to generate these 28000 curves. Those four points were used for scenarios where $1 - \beta = 0.2$ and $\theta_1 - \theta_2 = -12$; two of those points were used with the first (σ_1, σ_2, q) combination, and one of those points was used with each of the third and fifth (σ_1, σ_2, q) combinations. Those four points prompted multiple intersections between $se_r^{(n,q)}$ and $\Lambda_r^{(n,q)}$, one of which occurred for $n \leq n_*$ in Line 6 of Algorithm 2.2 and the other of which occurred for $n > n_*$. We needed to choose the other intersection to obtain an unbiased power estimate – even though $\lceil n_* \rceil$ and $\lceil n^* \rceil$ from Algorithm 2.2 were the same for the four power curves created using these points. We therefore very rarely need to adjust for multiple intersections, especially for high-powered studies since the root-finding algorithm never needed to be reimplemented for $1 - \beta \geq 0.3$. However, we cannot guarantee that it is unnecessary to adjust for multiple intersections with an arbitrary design, so that is why we incorporated Lines 7 to 13 into Algorithm 2.2 to ensure unbiased sample size recommendations.

A.2 Competing Methods for Power Analysis

In this section, we consider alternative methods for power analysis with the Welch-based TOST procedure. We illustrate that the consistency of the power estimates produced by such methods depends on the numerical integration settings. For the Welch-based TOST procedure, Jan and Shieh (2017) proposed an analytical method to compute power given sample sizes n_1 and n_2. They accounted for the degrees of freedom being unknown a priori by expressing the test statistic in terms of simpler normal, chi-square, and beta random variables. The sum of the sample variances for the two groups is related to a chi-square distribution, and the proportion of total variability arising from the first group is related to a beta distribution. Power was computed by integrating with respect to the expectation of these (independent) chi-square and beta random variables. Shieh et al. (2022) provided R code to implement exact power calculations for the Welch-based TOST procedure using two-dimensional quadrature with Simpson's rule (Süli and Mayers, 2003).

We computed power estimates using Jan and Shieh's (2017) method for the illustrative example from Section 2.3.2 with $n = \{3, 5, 8, 10, 15, 20, 30, 40, 50, 60\}$. When the numerical integration parameters for Simpson's rule are properly tuned, these power estimates coincided with those obtained by Algorithm 2.1 in Table 2.1 to four decimal places. With $n = 2$, Jan and Shieh's (2017) method provided a power estimate of 2.0838 using the recommended quadrature settings with 5×10^5 points. When using the default settings with 5×10^6, 5×10^7, and 2.5×10^8 points, their method respectively estimated power to be 0.2296, 0.0443, and 0.0278. The final estimate took roughly 56 seconds to compute on a standard laptop. For $n = 2$, we estimated power 100 times using Algorithm 2.1 with $m = 65536$ as done in Table 2.1. This gave rise to an empirical power estimate of 0.0238 and a 95% confidence interval of $(0.0236, 0.0240)$ created using the percentile bootstrap method (Efron, 1982); this confidence interval does not contain the final estimate returned by Jan and Shieh's (2017) method of 0.0278.

When fewer than 5×10^7 points are used with their default settings, power is not a nondecreasing function of the sample size n for the illustrative example. This occurs because the quadrature rule has not converged. This lack of convergence is problematic – even for the smallest possible sample size of $n = 2$. To find a suitable sample size, power is often computed successively for $n = \{2, 3, 4, ...\}$ until the target power $1 - \beta$ for the study is achieved. If sample sizes are explored using a bisection method, this process is initialized by computing power at lower and upper bounds for n. This lower bound is typically $n = 2$. Depending on the chosen quadrature settings, using either approach for this example could lead us to incorrectly conclude that $n = 2$ is sufficient for a high-powered study.

Alternative numerical integration techniques may also yield unstable results when com-

bined with Jan and Shieh's (2017) method. We modify Shieh et al.'s (2022) code to compute power using two-dimensional numerical integration techniques in R. This requires us to integrate over a beta variable with domain $(0, 1)$ and a chi-square variable with domain $\mathbb{R}^+$. In practice, we often need to choose a finite upper bound of integration for the chi-square variable. Figure A.1 illustrates that the estimated power for the illustrative example at various sample sizes n is sensitive to this choice of upper bound when implementing numerical integration via R's `pracma` package (Borchers, 2021).

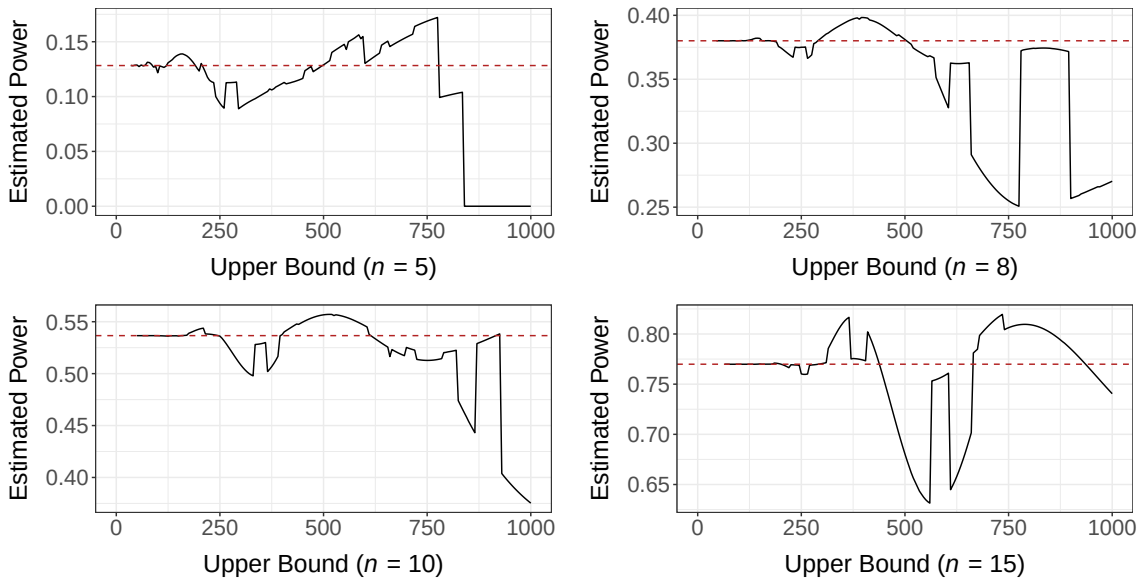


Figure A.1: Estimated power (black) for the illustrative example with $n = \{5, 8, 10, 15\}$ and various upper bounds of integration for the chi-square variable. Actual power for these designs is visualized in red.

Thus, the consistency of the power estimates returned by competing methods depends on the chosen integration bounds or point grid for the quadrature rule. These issues with consistency can be diagnosed when considering various values for the numerical integration parameters; however, diagnosing and correcting these issues may be outside the comfort zone of some practitioners. Jan and Shieh's (2017) method computes power for fixed sample sizes n_1 and n_2, so it is comparable to Algorithm 2.1. Algorithm 2.1's equivalent of the numerical integration parameters is the length m of the Sobol' sequence. We emphasize that this choice for m only impacts the precision – and not the consistency – of the power estimates. With the `dent` package, our methods are therefore easily applied and robust to these tuning issues.

A.3 Power Analysis for Crossover Designs

In crossover designs, each subject receives a different formulation of a drug (or nonclinical treatment) during different periods (Chow and Liu, 2008). Each group (or block) of subjects receives a different sequence of formulations. Crossover designs possess several advantages over their parallel counterparts. First, each subject can serve as their own control, which facilitates within-subject comparison between drug formulations. Crossover designs also remove inter-subject variability from between-formulation comparisons.

Although crossover designs often require fewer subjects to obtain desired power for the equivalence test, they take longer to implement than parallel designs because each subject is analyzed over multiple treatment periods. Moreover, there are typically rest periods between consecutive treatment periods so that the effect of the formulation administered in one treatment period does not persist in the next. These rest periods are called *washout* periods, and they should be long enough for the effect of one formulation to wear off so that there is no *carryover* effect in the next treatment period. If the washout period length is too short relative to the persistence of the formulation effects, we must distinguish between the effect of the drug being administered in a given period (direct drug effect) and the carryover effect. First-order carryover effects are those that last a single treatment period. Generally, higher-order carryover effects that last multiple treatment periods are not considered in bioequivalence studies.

There are many crossover designs that assess average bioequivalence, the most common of which is the two-sequence, two-period (2×2) crossover design. In the 2×2 crossover design, two drug formulations are compared: an established reference (R) drug and a new test (T) formulation. Moreover, subjects are assigned to sequence 1 (RT) or 2 (TR) in this type of design. The acronyms in parentheses denote which order the subjects in that sequence receive the test and reference formulations. There is a washout period between the two treatment periods. We consider the statistical model for the 2×2 crossover design described in Chow and Liu (2008). We let y_{ijk} be the response from the i^{th} subject in the k^{th} sequence at the j^{th} period such that

$$y_{ijk} = \mu + S_{ik} + P_j + F_{(j,k)} + C_{(j-1,k)} + e_{ijk}, \tag{A.1}$$

where $i = 1, ..., n_k$, $j = 1, 2$, and $k = 1, 2$. Here, n_k is the number of subjects in the k^{th} sequence, and μ is the overall mean. S_{ik} is the random effect for the i^{th} subject in the k^{th} sequence, and we assume that these terms are i.i.d. according to a normal distribution with mean 0 and variance σ_S^2. P_j is the fixed effect of the j^{th} period such that $P_1 + P_2 = 0$. $F_{(j,k)}$ is the direct fixed effect of the formulation administered to subjects in sequence k

during the j^{th} period. For the 2×2 crossover design, we have that $F_{(j,k)} = F_R$ if $j = k$ and F_T otherwise. We assume that $F_T + F_R = 0$. $C_{(j-1,k)}$ is the fixed first-order carryover effect of the formulation administered in the $(j-1)^{\text{th}}$ period of sequence k, where $C_{(0,k)} = 0$ for $k = 1, 2$. Furthermore, we have that $C_{(1,1)} = C_R$, $C_{(1,2)} = C_T$, and $C_T + C_R = 0$. Finally, e_{ijk} is the within-subject random error, where these terms are assumed to be i.i.d. normal with mean 0 and variance σ_T^2 or σ_R^2 depending on the formulation administered. We further assume that the S_{ik} and e_{ijk} terms are mutually independent.

The 2×2 crossover design allows us to consider the presence of carryover effects by testing the hypothesis that $C_T - C_R = 0$ (Chow and Liu, 2008). However, we cannot uniquely estimate model (A.1) using a 2×2 crossover design if carryover effects are present. In such scenarios, we also cannot obtain an unbiased estimator for the direct drug effect $F = F_T - F_R$ based on data from both periods. If carryover effects are present, only the data from the first period is typically used. This effectively reverts the design into a parallel study, and the methods from Chapter 2 can be applied. Here, we assume the absence of carryover effects (i.e., $C_T = C_R = 0$) and consider power analysis with carryover effects in the `dent` package. We define period differences for each subject within each sequence as

$$D_{ik} = \frac{1}{2}(y_{i2k} - y_{i1k}), \tag{A.2}$$

for $i = 1, 2, ..., n_k$ and $k = 1, 2$. In the absence of carryover effects, an unbiased estimator for the direct drug effect is $\hat{F} = \bar{D}_{\cdot 1} - \bar{D}_{\cdot 2}$. Under model (A.1), $\sigma_{Dk}^2 = Var(D_{ik}) = (\sigma_T^2 + \sigma_R^2)/4$ for both sequences $k = 1, 2$, and the equal variance assumption is theoretically sound. However, this assumption may be inappropriate if we allow the within-subject random errors to vary by treatment and sequence (i.e., $Var(e_{ijk})$ is σ_{Tk}^2 or σ_{Rk}^2 depending on the formulation administered in period j of sequence k). The equal variance assumption may also be inappropriate for crossover designs that account for carryover effects, such as two-sequence dual designs (Chow and Liu, 2008).

Thus, it is useful to have Welch-based design methods for crossover studies that allow for unequal variances. Algorithms 2.1 and 2.2 can readily be extended to serve this purpose. For parallel designs, an anticipated value for the effect $\theta_1 - \theta_2$ is chosen. In the 2×2 crossover design, a similar input for the sample size calculation is specified for $F = F_T - F_R$. Instead of hypothesizing values for inter-subject standard deviations σ_1 and σ_2, practitioners guess values for the standard deviations of the intra-subject differences in sequences 1 and 2: σ_{D1} and σ_{D2}. Algorithms 2.1 and 2.2 can be directly applied with the 2×2 crossover design by substituting $\theta_1 - \theta_2$ with F, σ_1 with $\sigma_{D1}/2$, and σ_2 with $\sigma_{D2}/2$. The intra-subject standard deviations are divided by two due to the factor of $1/2$ in (A.2).

We illustrate the value of this approach using an example from Chow et al. (2008). This example concerns a clinical trial that compares a test and a reference formulation of a drug using log-transformed area under the curve (AUC), which measures total drug exposure across time in pharmacokinetics contexts. The AUC data are assumed to be normal after this logarithmic transformation. The mean difference of AUC is assumed to be $F = 0.05$. The interval endpoints are chosen to be $\delta_U = -\delta_L = 0.223$ to comply with FDA requirements (FDA, 2003). Balanced samples are to be collected ($n = n_1 = n_2$), and past studies give rise to anticipated intra-subject standard deviations of $\sigma_{D1} = \sigma_{D2} = \sigma_D = 0.4$. The investigator wants to find the sample size n that achieves $100 \times (1-\beta)\% = 80\%$ power at the significance level of $\alpha = 0.05$. Chow et al. (2008) recommended conservatively choosing the smallest sample size n that satisfies

$$n \geq \frac{(t_{\alpha,2n-2} + t_{\beta/2,2n-2})^2 \sigma_D^2}{2(\delta_U - |F|)^2}. \tag{A.3}$$

Because (A.3) does not admit an explicit solution for the sample size per sequence, the desired n must be found numerically. As such, tables populated with n values corresponding to various F, β, and σ_D combinations are often used to select sample sizes. Table 10.2.1 on page 262 of Chow et al. (2008) recommended a sample size of $n = 24$ per sequence with this example. We first implemented Algorithm 2.2 for unequal variances with $\sigma_{D1} = \sigma_{D2} = 0.4$. For comparison, we also implemented an equal variance version of this approach that uses two-dimensional Sobol' sequences and the TOST procedure with Student's t-tests. Both the equal and unequal variance versions of our approach returned a recommended sample size of $n = 18$. Given that $n_1 = n_2$ and $\sigma_{D1} = \sigma_{D2}$, it is not surprising that both approaches recommend the same sample size based on numerical studies from Gruman et al. (2007) and Rusticus and Lovato (2014). The recommended sample size of $n = 24$ from (A.3) is 33% larger than $n = 18$, and a crossover study with $2 \times 24 = 48$ total subjects takes more resources to conduct than one with 36 subjects.

Chow et al. (2008) acknowledged that (A.3) returns a conservative sample size, but the degree of conservatism is not transparent. Their sample size recommendation is conservative when $\delta_U - |F| < |\delta_L|$ as in this example. For this example, using (A.3) to choose a sample size effectively changes the lower interval endpoint to $\delta_L = F - (\delta_U - F) = -0.123$. Both versions of Algorithm 2.2 with equal and unequal variances align with (A.3) and recommend $n = 24$ when $(\delta_L, \delta_U) = (-0.123, 0.223)$. Our approaches therefore better accommodate scenarios where $F \neq 0.5(\delta_L + \delta_U)$ than certain design methods that leverage static tables or analytical formulas – even when the equal variance assumption is appropriate. Since our methods leverage sampling distribution segments, this better performance does not come with a substantial computational cost.

Appendix B

Additional Material for Chapter 3

B.1 Additional Content for Theorem 3.1

B.1.1 Conditions for the Bernstein-von Mises Theorem

Theorem 3.1 requires that the conditions for the BvM theorem are satisfied. These conditions are described in more detail in van der Vaart (1998), starting on page 140. Conditions (B0), (B1), and (B2) concern the likelihood component of the posterior distribution for a parameter θ. (B3) concerns the prior specifications for θ. The fixed parameter value for θ is denoted as θ_0 in Chapter 3 and as θ^* in Chapter 5.

(B0) The observations are drawn independently and identically from a distribution P_{θ_0} for some fixed, nonrandom θ_0.

(B1) The parametric statistical model from which the data are generated is differentiable in quadratic mean.

(B2) There exists a sequence of uniformly consistent tests for testing $H_0 : \theta = \theta_0$ against $H_1 : \|\theta - \theta_0\| \geq \varepsilon$ for every $\varepsilon > 0$.

(B3) Let the prior distribution for θ be absolutely continuous in a neighbourhood of θ_0 with continuous positive density at θ_0.

B.1.2 Conditions for the Asymptotic Normality of the Maximum Likelihood Estimator

Theorem 3.1 also requires that the design distributions $f(y; \boldsymbol{\eta}_{1,0})$ and $f(y; \boldsymbol{\eta}_{2,0})$ satisfy the regularity conditions for the asymptotic normality of the maximum likelihood estimator. These conditions are detailed in Lehmann and Casella (1998); they consider a family of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, where Ω is the parameter space. Lehmann and Casella (1998) use θ as the unknown parameter with true fixed value θ_0, so we state the conditions using this notation. However, we use $\theta = \theta_1 - \theta_2$ or $\theta = \theta_1/\theta_2$ to compare two characteristics in our framework. For our purposes, the conditions in Lehmann and Casella (1998) must hold for the design distributions in Chapter 3 (with unknown parameters $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ and true values $\boldsymbol{\eta}_{1,0}$ and $\boldsymbol{\eta}_{2,0}$). In Chapter 5, these conditions must hold for the distributions parameterized by $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ with values $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$ drawn from a design prior.

Lehmann and Casella (1998) detail nine conditions that guarantee the asymptotic normality of the maximum likelihood estimator. We provide the following guidance on where to find more information about these conditions in their text. The first four conditions – (R0), (R1), (R2), and (R3) – are described on pages 443 and 444 of their text. (R4) is mentioned as part of Theorem 3.7 on page 447. (R5), (R6), and (R7) are described in Theorem 2.6 on pages 440 and 441. (R8) is mentioned in Theorem 3.10 on page 449.

(R0) The distributions P_θ of the observations are distinct.

(R1) The distributions P_θ have common support.

(R2) The observations are $\mathbf{X} = (X_1, ..., X_n)$, where the X_i are identically and independently distributed with probability density function $f(x_i|\theta)$ with respect to a σ-finite measure μ.

(R3) The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

(R4) For almost all x, $f(x|\theta)$ is differentiable with respect to θ in ω, with derivative $f'(x|\theta)$.

(R5) For every x in the set $\{x : f(x|\theta) > 0\}$, the density $f(x|\theta)$ is differentiable up to order 3 with respect to θ, and the third derivative is continuous in θ.

(R6) The integral $\int f(x|\theta)d\mu(x)$ can be differentiated three times under the integral sign.

(R7) The Fisher information $\mathcal{I}(\theta)$ satisfies $0 < \mathcal{I}(\theta) < \infty$.

(R8) For any given $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that $|\partial^3 \log f(x|\theta)/\partial\theta^3| \leq M(x)$ for all $\{x : f(x|\theta) > 0\}$, $\theta_0 - c < \theta < \theta_0 + c$, and $\mathbb{E}[M(X)] < \infty$.

B.1.3 Proof of Theorem 3.1

We first prove part (a) of Theorem 3.1. We extend the notation from (3.10) to estimate the posterior probabilities that comprise $\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.5)}$ and $\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.6)}$ when data $\boldsymbol{Y}^{(n)}$ are generated. For $\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.5)}$, the fraction inside the standard normal CDF of (3.10) converges to the following normal distribution:

$$\sqrt{n}\left(\frac{\delta - \theta_0}{\sqrt{\mathcal{I}(\hat{\theta}_n)^{-1}}} - \frac{\hat{\theta}_n - \theta_0}{\sqrt{\mathcal{I}(\hat{\theta}_n)^{-1}}}\right) \xrightarrow{d} \mathcal{N}\left(\frac{\delta - \theta_0}{\sqrt{\mathcal{I}(\theta_0)^{-1}}}, 1\right). \tag{B.1}$$

This result follows by the asymptotic normality of the MLEs $\hat{\boldsymbol{\eta}}_{1,n}$ and $\hat{\boldsymbol{\eta}}_{2,n}$, the continuous mapping theorem because $g(\cdot)$ and $h(\cdot)$ are differentiable at the design values, and Slutsky's theorem since $\mathcal{I}(\hat{\theta}_n)^{-1} \xrightarrow{P} \mathcal{I}(\theta_0)^{-1}$. When pseudorandom sequences $\boldsymbol{U} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1]^{2d})$ are input into Algorithm 3.1, the left side of (B.1) for $\mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}$ follows the normal distribution on the right side exactly. The CDFs of the sampling distributions $\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.5)}$ and $\mathcal{P}^{\delta}_{n,\boldsymbol{U},Alg.3.1}$ then converge pointwise as $n \to \infty$ by a second application of the continuous mapping theorem with the function $\Phi(\cdot)$. We obtain the result in part (a) regarding the total variation distance by Scheffé's lemma (Williams, 1991).

To prove part (b) for $\mathcal{P}^{\delta}_{n,\boldsymbol{Y}^{(n)},(3.6)}$, we note that the approximations in (3.5) and (3.6) are virtually the same as $n \to \infty$. Under the conditions for Theorem 3.1, the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n}$ converges in probability to $\boldsymbol{\eta}_{j,0}$ for $j = 1, 2$. The following result also holds for $\mathcal{J}_j(\tilde{\boldsymbol{\eta}}_{j,n})/n$ in (3.6):

$$\frac{1}{n}\mathcal{J}_j(\tilde{\boldsymbol{\eta}}_{j,n}) = \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\eta}_j^2}\log(f(y_{ij};\boldsymbol{\eta}_j)) - \frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\eta}_j^2}\log(p_j(\boldsymbol{\eta}_j))\right]_{\boldsymbol{\eta}_j=\tilde{\boldsymbol{\eta}}_{j,n}} \xrightarrow{P} \mathcal{I}(\boldsymbol{\eta}_{j,0}). \tag{B.2}$$

Because $\tilde{\boldsymbol{\eta}}_{j,n} - \hat{\boldsymbol{\eta}}_{j,n} \xrightarrow{P} 0$, the mean and variance of the normal distribution in (3.6) respectively approximate $\hat{\theta}_n$ and $\mathcal{I}(\hat{\theta}_n)^{-1}/n$ in (3.5) for large sample sizes n by the continuous mapping theorem. The result in part (b) then follows from part (a). □

B.2 Proof of Lemma 3.1

To prove part (a) of Lemma 3.1, we only present the proof for group 1 since the proof for group 2 follows the same process. We use induction on the dimension d of $\boldsymbol{\eta}_1$ for this proof. We show the base case corresponding to a model with $d = 2$. To simplify notation, we let

$$\mathcal{I}(\boldsymbol{\eta}_{1,0})^{-1} = \begin{bmatrix} \sigma_{11}^2 & \rho_{12}\sigma_{11}\sigma_{22} \\ \rho_{12}\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{bmatrix}.$$

By properties of the bivariate conditional normal distribution, it follows that

$$\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(1)} = \boldsymbol{\eta}_{1,0}^{(1)} + \frac{1}{\sqrt{n}}\Phi^{-1}(u_1)\sigma_{11} \quad \text{and} \tag{B.3}$$

$$\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(2)} = \boldsymbol{\eta}_{1,0}^{(2)} + \frac{1}{\sqrt{n}}\sigma_{22}\left[\Phi^{-1}(u_1)\rho_{12} + \Phi^{-1}(u_2)\sqrt{1-\rho_{12}^2}\right]. \tag{B.4}$$

The result in part (a) therefore holds true when $d = 2$, where $\omega_1(u_1)$ and $\omega_2(u_1, u_2)$ are given by the expressions to the right of the $1/\sqrt{n}$ terms in (B.3) and (B.4), respectively.

For the inductive hypothesis, we assume that the result in part (a) of Lemma 3.1 holds true for a model with $d = l$ parameters. For the inductive conclusion, we show that this implies the result also holds for a model with $d = l + 1$ parameters. Because $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(1)}, ..., \hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(l)}$ only depend on the components with smaller indices, we just need to prove that the result in part (a) holds for $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(l+1)}$. That result in conjunction with the inductive hypothesis proves the inductive conclusion. To prove the inductive conclusion, we introduce the following block matrix notation:

$$\mathcal{I}(\boldsymbol{\eta}_{1,0})^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{l,l} & \boldsymbol{\Sigma}_{l,1} \\ \boldsymbol{\Sigma}_{1,l} & \Sigma_{l+1,l+1} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{l,l}$ is a $l \times l$ matrix, $\boldsymbol{\Sigma}_{l,1}$ is a $l \times 1$ matrix, $\boldsymbol{\Sigma}_{1,l} = \boldsymbol{\Sigma}_{l,1}^T$, and $\Sigma_{l+1,l+1}$ is scalar.

The marginal distribution of $\hat{\boldsymbol{\eta}}_{1,n}^{(l+1)}$ conditional on the already-generated $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(k)}$ for $k = 1, ..., l$ is

$$\mathcal{N}\left(\boldsymbol{\eta}_{1,0}^{(l+1)} + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_{1,l}\boldsymbol{\Sigma}_{l,l}^{-1}\begin{pmatrix} \omega_1(u_1) \\ \vdots \\ \omega_l(u_1, ..., u_l) \end{pmatrix}, \frac{1}{n}\left[\Sigma_{l+1,l+1} - \boldsymbol{\Sigma}_{1,l}\boldsymbol{\Sigma}_{l,l}^{-1}\boldsymbol{\Sigma}_{l,1}\right]\right).$$

Therefore, we have that

$$\begin{aligned}\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(l+1)} = \boldsymbol{\eta}_{1,0}^{(l+1)} &+ \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_{1,l}\boldsymbol{\Sigma}_{l,l}^{-1}\begin{pmatrix}\omega_1(u_1)\\ \vdots \\ \omega_l(u_1,...,u_l)\end{pmatrix} \\ &+ \frac{1}{\sqrt{n}}\Phi^{-1}(u_{l+1})\left[\Sigma_{l+1,l+1} - \boldsymbol{\Sigma}_{1,l}\boldsymbol{\Sigma}_{l,l}^{-1}\boldsymbol{\Sigma}_{l,1}\right]^{1/2}.\end{aligned} \tag{B.5}$$

The result from part (a) of Lemma 3.1 holds for $\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)^{(l+1)}$ if we take $\omega_{l+1}(u_1,...,u_{l+1})$ as the sum of the two components to the right of the $1/\sqrt{n}$ terms in (B.5). By mathematical induction, part (a) of Lemma 3.1 is true for an arbitrary model with d parameters.

Part (b) of Lemma 3.1 follows from the first-order Taylor expansion of $h(g(\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)), g(\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u}_r)))$ around $(\boldsymbol{\eta}_{1,0}, \boldsymbol{\eta}_{2,0})$. We have that

$$\begin{aligned}&h(g(\hat{\boldsymbol{\eta}}_{1,n}(\boldsymbol{u}_r)), g(\hat{\boldsymbol{\eta}}_{2,n}(\boldsymbol{u}_r))) - h(g(\boldsymbol{\eta}_{1,0}), g(\boldsymbol{\eta}_{2,0})) \\ &\approx \sum_{j=1}^{2}\sum_{k=1}^{d}\frac{\partial h}{\partial g_j}\frac{\partial g_j}{\partial \boldsymbol{\eta}_j^{(k)}}\bigg|_{(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)=(\boldsymbol{\eta}_{1,0},\boldsymbol{\eta}_{2,0})}\left[\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u}_r)^{(k)} - \boldsymbol{\eta}_{j,0}^{(k)}\right] \\ &\approx \frac{1}{\sqrt{n}}\left[\sum_{j=1}^{2}\sum_{k=1}^{d}\frac{\partial h}{\partial g_j}\frac{\partial g_j}{\partial \boldsymbol{\eta}_j^{(k)}}\bigg|_{(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)=(\boldsymbol{\eta}_{1,0},\boldsymbol{\eta}_{2,0})}\omega_{(j-1)d+k}(u_{(j-1)d+1},...,u_{(j-1)d+k})\right].\end{aligned} \tag{B.6}$$

Part (b) of Lemma 3.1 follows if we let $\omega_\dagger(\cdot)$ be the sum on the right side of the $1/\sqrt{n}$ term in (B.6). However, this Taylor series expansion may only be suitable for large sample sizes n, i.e., when $\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u}_r)$ is sufficiently near $\boldsymbol{\eta}_{j,0}$ for $j = 1, 2$. Our simulation procedure for $\hat{\boldsymbol{\eta}}_{j,n}(\boldsymbol{u}_r)$ ensures this convergence occurs for large n, and the conditions for Theorem 3.1 guarantee that $\hat{\boldsymbol{\eta}}_{j,n}$ based on the data $\boldsymbol{y}^{(n)}$ also converges in probability to $\boldsymbol{\eta}_{j,0}$.

To prove part (c), we note that the following result holds for sufficiently large n:

$$\begin{aligned}\frac{\delta - \underline{\theta}_r^{(n)}}{\sqrt{\underline{\mathcal{I}}_r^{(n)}}} &\approx \frac{\delta - (h(g(\boldsymbol{\eta}_{1,0}), g(\boldsymbol{\eta}_{2,0})) + \omega_\dagger(u_1,...,u_{2d})/\sqrt{n})}{\sqrt{\mathcal{I}(\theta_0)^{-1}/n}} \\ &= \frac{\delta - \theta_0}{\sqrt{\mathcal{I}(\theta_0)^{-1}}}\sqrt{n} - \frac{\omega_\dagger(u_1,...,u_{2d})}{\sqrt{\mathcal{I}(\theta_0)^{-1}}}.\end{aligned} \tag{B.7}$$

The approximate equivalence of the numerators in the first line of (B.7) follows from part (b) of Lemma 3.1 and the fact that $\tilde{\boldsymbol{\eta}}_{j,n} - \hat{\boldsymbol{\eta}}_{j,n}$ and $\boldsymbol{\eta}_{j,n}^* - \hat{\boldsymbol{\eta}}_{j,n}$ converge in probability to 0. Moreover, $\underline{\mathcal{I}}_r^{(n)} \approx \mathcal{I}(\theta_0)^{-1}/n$ for sufficiently large n by the continuous mapping theorem

for Algorithm 3.1, (B.2) for Algorithm 3.2, and similar logic to (B.2) for Algorithm 3.3. The second line of (B.7) holds because $\theta_0 = h(g(\boldsymbol{\eta}_{1,0}), g(\boldsymbol{\eta}_{2,0}))$. The expression in (B.7) takes the form $a(\delta, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)$ since neither fraction in the second line depends on n. Part (c) of Lemma 3.1 follows by using the normal CDF as in (3.6). We note that the function $a(\delta, \theta_0)$, which is the fraction to the left of the $\sqrt{n}$ term in the second line of (B.7), must incorporate monotonic transformations applied to the posterior of θ to improve the suitability of its normal approximation.

Part (d) of Lemma 3.1 follows from taking the derivative of the approximation to $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta}$ prompted by part (c) with respect to the sample size n:

$$\begin{aligned} &\frac{d}{dn}\left[\, p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta}\right] \\ &\approx \frac{d}{dn}\left[\, \Phi\left(a(\delta_U, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)\right) - \Phi\left(a(\delta_L, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)\right)\right] \\ &= \frac{a(\delta_U, \theta_0)\phi\left(a(\delta_U, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)\right) - a(\delta_L, \theta_0)\phi\left(a(\delta_L, \theta_0)\sqrt{n} + b(\boldsymbol{u}_r)\right)}{2\sqrt{n}}, \end{aligned} \tag{B.8}$$

where $\phi(\cdot)$ is the probability density function (PDF) of the standard normal distribution. This derivative must be positive for sufficiently large n where the approximation from part (c) of Lemma 3.1 holds. When $\theta_0 \in (\delta_L, \delta_U)$, $a(\delta_U, \theta_0)$ is positive and $a(\delta_L, \theta_0)$ is negative. Because the normal distribution takes support over the entire real line, $\phi(\cdot)$ returns a positive value for any real input. If δ_L or δ_U is not finite, then its component of the difference in the numerator of (B.8) is zero. The remaining component of the numerator is still positive and so is the derivative in (B.8). Hence, part (d) of Lemma 3.1 is true, and $p^{\delta_U}_{n,\boldsymbol{u}_r,\zeta} - p^{\delta_L}_{n,\boldsymbol{u}_r,\zeta}$ is an increasing function for sufficiently large n. □

B.3 Additional Numerical Studies

B.3.1 Numerical Studies with Bayes Factors

We now compare the performance of our power curve approximation procedure for several posterior analyses with Bayes factors. To do so, we modify Settings 1a and 2a from Section 3.6.1. We do not change the target power, interval (δ_L, δ_U), or the prior distributions. For Setting 1a with the uninformative priors, $Pr(H_1) = 0.0128$ for $H_1 : \theta_1/\theta_2 \in (\delta_L, \delta_U)$. We consider a threshold for the nonoverlapping-hypotheses Bayes factor of $K = 100$ for illustration. By (3.2), this corresponds to a critical value of $\gamma = 0.5652$. For Setting

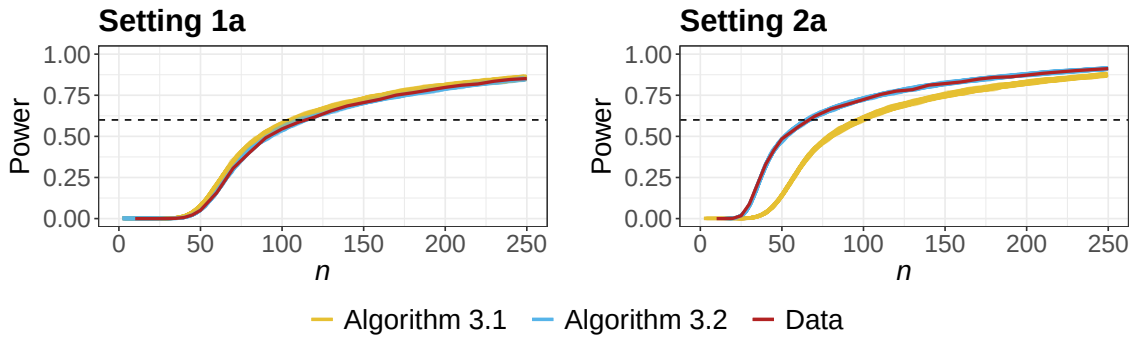


Figure B.1: 100 power curves obtained via Algorithms 3.1 (yellow) and 3.2 (blue), power curve estimated via simulated data (red), and target power $1-\beta$ (dotted line) for Settings 1a and 2a with hypothesis tests facilitated via NOH Bayes factors.

1b with the informative priors, $Pr(H_1) = 0.2835$. We consider a threshold for the NOH Bayes factor of $K = 3$ for illustration, which corresponds to a critical value of $\gamma = 0.5428$. This example illustrates the importance of considering the impact of the prior for θ when choosing a threshold K.

The numerical study in this subsection was otherwise carried out using the same process as described in Section 3.6.1. For each setting, we obtained 100 approximations to the power curve using Algorithm 3.4 with $\zeta = \{\text{Alg. 3.1}, \text{Alg. 3.2}\}$. The results for Settings 1a (left) and 2a (right) are depicted in Figure B.1. This numerical study supports similar conclusions as those drawn regarding Settings 1a and 2a for hypothesis tests with posterior probabilities in Figure 3.4.

B.3.2 Numerical Studies with Imbalanced Sample Sizes

In Section 3.7, we acknowledge that our framework as presented in Chapter 3 does not support imbalanced sample size determination (i.e., where $n_2 = qn_1$ for some constant $q > 0$). In this subsection, we describe how to extend our methods to allow for imbalanced sample size determination. This procedure requires practitioners to choose the constant q a priori. When $n_1 \neq n_2$, we use the following limiting posteriors for each group: $\mathcal{N}(\boldsymbol{\eta}_{1,0}, \mathcal{I}(\boldsymbol{\eta}_{1,0})^{-1}/n)$ and $\mathcal{N}(\boldsymbol{\eta}_{2,0}, \mathcal{I}(\boldsymbol{\eta}_{2,0})^{-1}/(qn))$. To apply the multivariate delta method to obtain the limiting posterior of $\theta = h(g(\boldsymbol{\eta}_1), g(\boldsymbol{\eta}_2))$, both the limiting variances of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ must be functions of n. We therefore treat $\mathcal{I}(\boldsymbol{\eta}_{2,0})^{-1}/q$ as the inverse Fisher information for $\boldsymbol{\eta}_2$ evaluated at the design value $\boldsymbol{\eta}_{2,0}$. This modification is also incorporated into the process to simulate maximum likelihood estimates for $\boldsymbol{\eta}_2$ in Algorithms 3.1, 3.2,

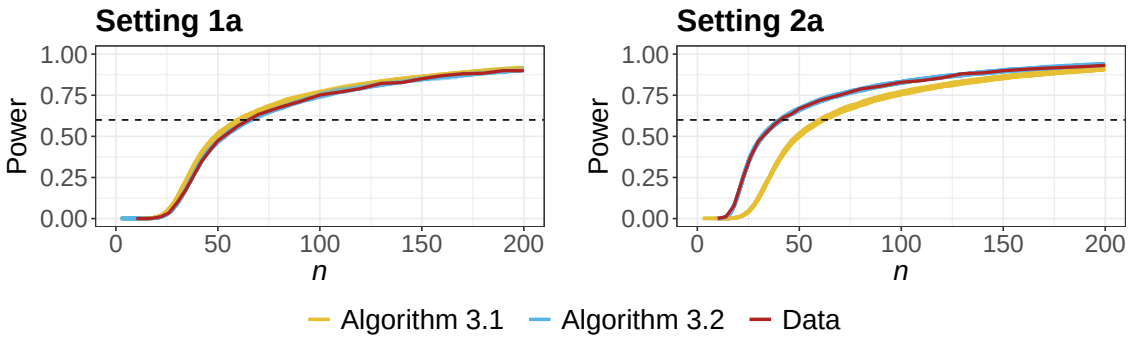


Figure B.2: 100 power curves obtained via Algorithms 3.1 (yellow) and 3.2 (blue), power curve estimated via simulated data (red), and target power $1-\beta$ (dotted line) for Settings 1a and 2a with hypothesis tests with imbalanced sample sizes.

and 3.3. That is, the variability in the marginal limiting distribution of $\hat{\boldsymbol{\eta}}_{2,qn}$ is scaled to reflect the larger ($q > 1$) or smaller ($0 < q < 1$) sample size in the second group. Similar modifications are also made when taking the normal approximation to the posterior of θ in (3.5), (3.6), and (3.9). No other modifications are required to account for imbalanced sample sizes.

Lastly, we evaluate the performance of our power curve approximation procedure with several scenarios that have imbalanced sample sizes. We reuse Settings 1a and 2a from Section 3.6.1. The only differences between this numerical study and the one conducted in Section 3.6.1 are those described in the previous paragraphs. We choose $q = 2$ (i.e., $n_2 = 2n_1$) for this numerical study. This reflects the male ($j = 2$) provider group having *roughly* twice as many observations as the female ($j = 1$) one in the motivating example from Section 3.3. For each setting, we obtained 100 approximations to the power curve using Algorithm 3.4 with $\zeta = \{\text{Alg. } 3.1, \text{Alg. } 3.2\}$. These results are depicted in Figure B.2. We again reach similar conclusions as those drawn for Settings 1a and 2a using Figure 3.4. We note that imbalanced sample size determination with Bayesian hypothesis tests is considered more formally in Chapter 5.

Appendix C

Additional Material for Chapter 4

C.1 Fisher Information for the Conditional Multinomial Model

Lemma C.1. *Let $\boldsymbol{\theta} = (Z_1, Z_2, ..., Z_{w-1})$ be the conditional probabilities defined in (4.2) for the standard multinomial model with w categories. Then, the inverse Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$ is diagonal for all possible $(Z_1, Z_2, ..., Z_{w-1}) \in (0,1)^{w-1}$.*

To prove Lemma C.1, we consider fixed parameters $\{\boldsymbol{p}_{w-1,0} = (p_{1,0}, ..., p_{w,0}) : 0 < p_{1,0}, ..., p_{w,0} < 1, \sum_{v=1}^{w} p_{v,0} = 1\}$. The inverse of the Fisher information $\mathcal{I}(\boldsymbol{p}_{w-1,0})$ is given by $\mathcal{I}(\boldsymbol{p}_{w-1,0})^{-1} = n^{-1}\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}_{s,t} = p_{s,0}(1-p_{s,0})$ if $1 \leq s = t \leq w-1$ and $-p_{s,0}p_{t,0}$ otherwise. We let $\hat{p}_v$ and $\hat{p}_{v'}$ for $1 \leq v < v' \leq w-1$ be MLEs for the multinomial parameters. We denote the vector of all such MLEs as $\hat{\boldsymbol{p}}_{w-1} = (\hat{p}_1, ..., \hat{p}_{w-1})$. We define analogues to the transformations in (4.2) for the fixed parameter values as $(Z_{1,0}, Z_{2,0}, ..., Z_{w-1,0})$.

We let $Z_{v,0} = g_v(\boldsymbol{p}_{w-1,0})$ for $v = 1, ..., w-1$, and define $\hat{Z}_v = g_v(\hat{\boldsymbol{p}}_{w-1})$. By the multivariate delta method, the asymptotic covariance $ACov(\cdot,\cdot)$ between $g_v(\hat{\boldsymbol{p}}_{w-1})$ and $g_{v'}(\hat{\boldsymbol{p}}_{w-1})$ for $1 \leq v < v' \leq w-1$ is

$$ACov(g_v(\hat{\boldsymbol{p}}_{w-1}), g_{v'}(\hat{\boldsymbol{p}}_{w-1})) = \frac{1}{n}\sum_{s=1}^{w-1}\sum_{t=1}^{w-1}\frac{\partial g_v}{\partial p_{s,0}}\frac{\partial g_{v'}}{\partial p_{t,0}}\boldsymbol{\Sigma}_{s,t}. \tag{C.1}$$

We use induction to show that $ACov(g_v(\hat{\boldsymbol{p}}_{w-1}), g_{v'}(\hat{\boldsymbol{p}}_{w-1})) = 0$ for all $1 \leq v < v' \leq w-1$. We first show that $ACov(g_1(\hat{\boldsymbol{p}}_2), g_2(\hat{\boldsymbol{p}}_2)) = 0$. This base case corresponds to the model

with $w = 3$. We have that

$$\frac{\partial g_1}{\partial \boldsymbol{p}_{2,0}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \frac{\partial g_2}{\partial \boldsymbol{p}_{2,0}} = \frac{p_{2,0}}{(1-p_{1,0})^2} \begin{bmatrix} 1 \\ (1-p_{1,0})/p_{2,0} \end{bmatrix}.$$

Because $\partial g_1 / \partial p_{2,0} = 0$, it follows that $ACov(\hat{Z}_1, \hat{Z}_2) = 0$ for the base case:

$$\begin{aligned} n \times ACov(g_1(\hat{\boldsymbol{p}}_2), g_2(\hat{\boldsymbol{p}}_2)) &= \frac{\partial g_1}{\partial p_{1,0}} \frac{\partial g_2}{\partial p_{1,0}} \boldsymbol{\Sigma}_{1,1} + \frac{\partial g_1}{\partial p_{1,0}} \frac{\partial g_2}{\partial p_{2,0}} \boldsymbol{\Sigma}_{1,2} \\ &= \frac{p_{1,0} p_{2,0}}{(1-p_{1,0})^2} \left[1 - p_{1,0} - p_{2,0}(1-p_{1,0})/p_{2,0} \right] \\ &= 0. \end{aligned} \tag{C.2}$$

For the inductive hypothesis, we assume that $ACov(g_v(\hat{\boldsymbol{p}}_{l-1}), g_{v'}(\hat{\boldsymbol{p}}_{l-1})) = 0$ for all $1 \leq v < v' \leq l-1$. This implies that the result for the base case holds for an arbitrary multinomial model with $w = l$ categories. For the inductive conclusion, we show that this result also holds for an arbitrary model with $w = l+1$ categories. With $l+1$ categories, we have that

$$\frac{\partial g_1}{\partial \boldsymbol{p}_{l,0}} = \begin{bmatrix} 1 \\ \mathbf{0}_{l-1} \end{bmatrix}, \quad \frac{\partial g_v}{\partial \boldsymbol{p}_{l,0}} = \frac{p_{v,0}}{\left(1 - \sum_{t=1}^{v-1} p_{t,0}\right)^2} \begin{bmatrix} \mathbf{1}_{v-1} \\ \left(1 - \sum_{t=1}^{v-1} p_{t,0}\right)/p_{v,0} \\ \mathbf{0}_{l-v} \end{bmatrix}, \text{ for } v = 2, ..., l-1,$$

$$\text{and} \quad \frac{\partial g_l}{\partial \boldsymbol{p}_{l,0}} = \frac{p_{l,0}}{\left(1 - \sum_{t=1}^{l-1} p_{t,0}\right)^2} \begin{bmatrix} \mathbf{1}_{l-1} \\ \left(1 - \sum_{t=1}^{l-1} p_{t,0}\right)/p_{l,0} \end{bmatrix}. \tag{C.3}$$

Because of the upper triangular structure of the partial derivatives in (C.3) and the inductive hypothesis, $ACov(g_v(\hat{\boldsymbol{p}}_l), g_{v'}(\hat{\boldsymbol{p}}_l)) = 0$ for all $1 \leq v < v' \leq l-1$. We therefore just need to consider $ACov(g_v(\hat{\boldsymbol{p}}_l), g_l(\hat{\boldsymbol{p}}_l))$ for all $1 \leq v \leq l-1$. We now derive general expressions for these asymptotic covariances.

We first derive the expression for $ACov(g_1(\hat{\boldsymbol{p}}_l), g_l(\hat{\boldsymbol{p}}_l))$. Similar to (C.2), we find

$$\begin{aligned} n \times ACov(g_1(\hat{\boldsymbol{p}}_l), g_l(\hat{\boldsymbol{p}}_l)) &= \sum_{s=1}^{l} \frac{\partial g_1}{\partial p_{1,0}} \frac{\partial g_l}{\partial p_{s,0}} \boldsymbol{\Sigma}_{1,s} \\ &= \frac{p_{1,0} p_{l,0}}{\left(1 - \sum_{t=1}^{l-1} p_{t,0}\right)^2} \left[\left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) - p_{l,0} \left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) / p_{l,0} \right] \\ &= 0 \end{aligned}$$

We now find an expression for $ACov(g_v(\hat{\boldsymbol{p}}_l), g_l(\hat{\boldsymbol{p}}_l))$, $v = 2, ..., l-1$. By matrix multiplication, (C.3), and the expression for $\boldsymbol{\Sigma}$, we obtain

$$\begin{aligned} n \times ACov(g_v(\hat{\boldsymbol{p}}_l), g_l(\hat{\boldsymbol{p}}_l)) &= \sum_{s=1}^{v} \sum_{t=1}^{l} \frac{\partial g_v}{\partial p_{s,0}} \frac{\partial g_l}{\partial p_{t,0}} \boldsymbol{\Sigma}_{s,t} \\ &= \frac{p_{v,0} p_{l,0}}{\left(1 - \sum_{t=1}^{v-1} p_{t,0}\right)^2 \left(1 - \sum_{t=1}^{l-1} p_{t,0}\right)^2} \\ &\quad \times \left\{ \sum_{s=1}^{v-1} p_{s,0} \left[\left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) - p_{l,0} \left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) / p_{l,0} \right] + \right. \\ &\quad \left. \left(1 - \sum_{s=1}^{v-1} p_{s,0}\right) \left[\left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) - p_{l,0} \left(1 - \sum_{t=1}^{l-1} p_{t,0}\right) / p_{l,0} \right] \right\} \\ &= 0 \end{aligned}$$

Therefore, $ACov(g_v(\hat{\boldsymbol{p}}_l), g_{v'}(\hat{\boldsymbol{p}}_l)) = 0$ for all $1 \leq v < v' \leq l$. This result holds for arbitrary $\boldsymbol{p}_{w-1,0}$. By mathematical induction, $ACov(\hat{Z}_v, \hat{Z}_{v'}) = 0$ for all $1 \leq v < v' \leq w-1$ for an arbitrary multinomial model with $\{w \in \mathbb{N} : w \geq 3\}$ categories. Therefore, $\mathcal{I}(\boldsymbol{\theta})^{-1}$ is diagonal for all possible $(Z_1, Z_2, ..., Z_{w-1}) \in (0,1)^{w-1}$.

C.2 More Simulations for the Calibration of Credible Sets

Here, we investigate how the prior dependence structure impacts the calibration of Bayesian credible sets for an example where the prior dependence structure is not a chronically

rejected one. We consider the standard gamma model parameterized by $\boldsymbol{\theta} = (\alpha, \lambda)$, where α and λ are respectively the shape and rate parameters. The correlation between α and λ dictated by the inverse Fisher information matrix is $1/\sqrt{\alpha\psi_1(\alpha)}$, where $\psi_1(\cdot)$ is the trigamma function. The correlation $1/\sqrt{\alpha\psi_1(\alpha)}$ is a positive and increasing function for all $\alpha > 0$. When the conditions for Theorem 4.1 hold, the joint posterior of $\boldsymbol{\theta}$ will be unable to retain negative dependence structures between α and λ. However, positive dependence structures between α and λ do not satisfy the conditions for chronic rejection outlined in Corollary 4.1 – even if the magnitude of the prior dependence is not retained a posteriori.

We define the prior predictive distribution of $\boldsymbol{Y}^{(n)}$ for these simulations by joining a GAMMA$(1000, 5000)$ prior for α and a GAMMA$(1000, 800)$ prior for λ with a Gaussian copula parameterized with Pearson's $\rho = 0.4$. For each of 10000 simulation repetitions, we approximated the posterior of $\boldsymbol{\theta}|\boldsymbol{y}^{(n)}$ using sampling-resampling methods (Rubin, 1988) with "nature's" analysis prior. The proposal distribution was the posterior of $\boldsymbol{\theta}|\boldsymbol{y}^{(n)}$ obtained by independently joining the marginal priors for α and λ, and we sampled from the proposal distribution using Markov chain Monte Carlo methods. For each posterior, we approximated its 95% HPD set for α and λ using two-dimensional kernel density estimation (Ripley, 2002). Empirical coverage was estimated as the proportion of simulation repetitions for which the parameter value $\boldsymbol{\theta}_0 = (\alpha_0, \lambda_0)$ used to generate the gamma data was contained in this HPD set. We implemented this process for $n = \{10^1, 10^2, 10^3, 10^4, 10^5\}$. We then repeated this process for analysis priors $p(\boldsymbol{\theta})$ that joined the marginal gamma priors from "nature's" prior with a Gaussian copula parameterized by Pearson's $\rho = \{0, 0.05, \ldots, 0.95\} \setminus \{0.4\}$. The results from this numerical study are visualized in Figure C.1.

Given the marginal GAMMA$(1000, 5000)$ prior for α that defines $p_D(\boldsymbol{\theta})$, the posterior dependence structure between α and λ should converge to that of a Gaussian copula with Pearson's ρ ranging between 0.86 and 0.9. We therefore expect the strength of the positive dependence between α and λ to increase along with the sample size n for nearly all ρ values for the analysis prior considered in Figure C.1. As in Figure 4.3, the empirical coverage in Figure C.1 exceeds the nominal level for small sample sizes n when the strength of the dependence between the components in $\boldsymbol{\theta}$ is slightly understated. As n increases, the impact of prior dependence is once again reduced and the empirical coverage approaches roughly 95% for all ρ values considered. In general, the results from this numerical study support the recommendations for prior dependence structure specification provided in Section 4.4.2.

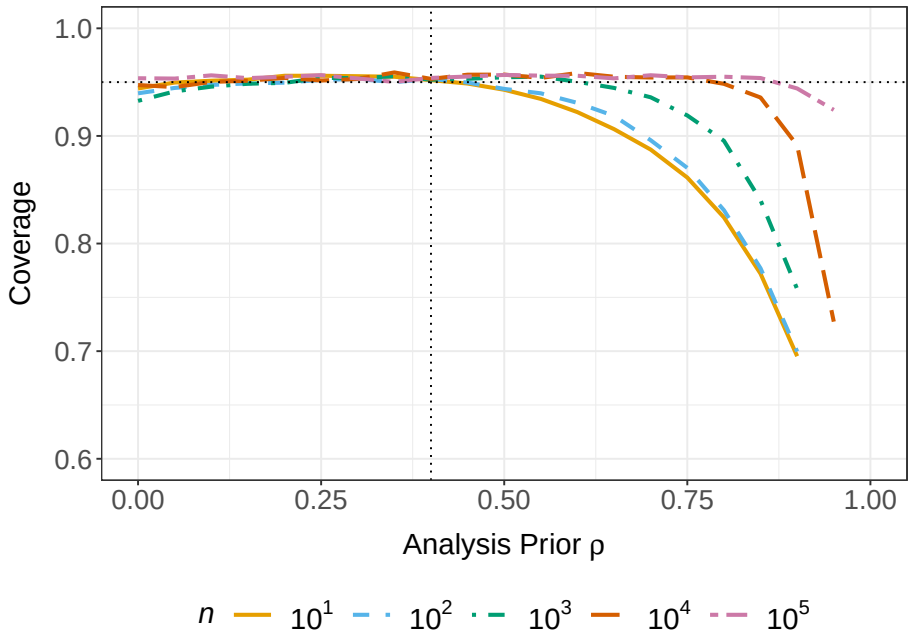


Figure C.1: Empirical coverage of 95% HPD for the gamma parameter $\boldsymbol{\theta} = (\alpha, \lambda)$ across 10000 posteriors. The horizontal dotted line denotes the nominal coverage, and the vertical one denotes "nature's" prior.

C.3 Proof of Theorem 4.2

Proof of Theorem 4.2*(a)*. The posterior mode $\boldsymbol{\theta}^{(k)}$ minimizes $-\log(p_k(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}))$ for $k = 1, 2$. We have that

$$\begin{aligned} &-\log(p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})) \\ &= -l(\boldsymbol{\theta}; \boldsymbol{y}^{(n)}) - \sum_{j=1}^{d} \log(f_j(\theta_j)) - \log(c_1(\boldsymbol{u})) + (\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))) + A, \end{aligned} \tag{C.4}$$

where $l(\boldsymbol{\theta}; \boldsymbol{y}^{(n)})$ is the log-likelihood function and the constant A reflects marginal likelihood term. The first three terms in (C.4) are equal to $-\log(p_1(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}))$. When $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$, it follows that $-\nabla_{\boldsymbol{\theta}}\log(p_1(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})) = 0$ and $-\nabla_{\boldsymbol{\theta}}\log(p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})) = -\nabla_{\boldsymbol{\theta}}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big]$. For large sample sizes n, the conditions for Theorem 4.2 ensure that the log-posterior $-\log(p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}))$ converges to its quadratic approximation about $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. We use $\boldsymbol{\theta}^{(1)}$ as a starting point for the Newton–Raphson method (Nocedal and Wright, 2006) with

$-\log(p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)}))$ to find $\boldsymbol{\theta}^{(2)}$. Because the quadratic approximation is appropriate for large samples, only one iteration of the Newton–Raphson method is required to approximate $\boldsymbol{\theta}^{(2)}$. We let $\mathcal{J}(\cdot)$ be the observed information. It follows that

$$\boldsymbol{\theta}^{(2)} \approx \boldsymbol{\theta}^{(1)} - \mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1}\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big]. \tag{C.5}$$

The conditions for Theorem 4.2 also guarantee that the posterior mode $\boldsymbol{\theta}^{(1)}$ is approximately equal to the approximately normal MLE $\hat{\boldsymbol{\theta}}^{(1)}$ for large samples $\boldsymbol{y}^{(n)}$. Because the MLE is consistent, $\mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1} \approx \mathcal{J}(\boldsymbol{\theta}_0)^{-1}$ for sufficiently large samples. By (4.8), we have that $\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big] \neq \mathbf{0}$, so $\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))$ can be approximated by a plane (with common gradient) in a neighbourhood of $\boldsymbol{\theta}_0$. It follows that $\mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1}\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}}\big[\log(c_1(\boldsymbol{u}))-\log(c_2(\boldsymbol{u}))\big]$ will be roughly constant for all large samples $\boldsymbol{y}^{(n)}$ generated from $m(\cdot|\boldsymbol{\theta}_0)$. For an arbitrarily large sample $\boldsymbol{y}^{(n)}$, we have that $\boldsymbol{\theta}^{(2)} \approx \boldsymbol{\theta}^{(1)} + \boldsymbol{b}$ for some constant $\boldsymbol{b}$ that does not depend on $\boldsymbol{y}^{(n)}$ by (C.5). Since the posterior concentrates around $\boldsymbol{\theta}_0$ as the sample size n increases, this common perturbation $\boldsymbol{b}$ will decrease in magnitude. $\boldsymbol{\theta}^{(1)}$ is approximately normally distributed about $\boldsymbol{\theta}_0$ for large samples, so these small perturbations will shift $\boldsymbol{\theta}^{(2)} \approx \boldsymbol{\theta}^{(1)} + \boldsymbol{b}$ closer to $\boldsymbol{\theta}_0$ with probability of roughly 0.5 due to the symmetry of the normal distribution. □

Proof of Theorem 4.2*(b)*. The results from (C.4) and (C.5) hold true as in part *(a)*. Because $\boldsymbol{u}_0$ is a local minimum of $\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))$, the local linear approximation of this function is not serviceable at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. However, this implies that $-\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big]$ should be directed toward $\boldsymbol{\theta}_0$ for large samples $\boldsymbol{y}^{(n)}$ when the quadratic approximation to the log-posterior is appropriate and $\boldsymbol{\theta}^{(1)} \approx \boldsymbol{\theta}_0$. We let the orthonormal basis $\mathcal{B}$ be composed of the eigenvectors of $\mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1}$. Since $\mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1}$ is a positive definite matrix, the angle between $-\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big]$ and $-\mathcal{J}(\boldsymbol{\theta}^{(1)})^{-1}\nabla_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}}\big[\log(c_1(\boldsymbol{u})) - \log(c_2(\boldsymbol{u}))\big]$ will be acute. With respect to $\mathcal{B}$, the perturbation from $\boldsymbol{\theta}^{(1)}$ to $\boldsymbol{\theta}^{(2)}$ induced by the Newton-Raphson method is then directed in the same orthant of $\boldsymbol{\Theta} \subseteq \mathbb{R}^d$ that contains $\boldsymbol{\theta}_0$. It follows that for large samples, $\boldsymbol{\theta}^{(2)}$ cannot be further from $\boldsymbol{\theta}_0$ than $\boldsymbol{\theta}^{(1)}$ due to a perturbation in the wrong direction. However, $\boldsymbol{\theta}^{(2)}$ could still be further from $\boldsymbol{\theta}_0$ than $\boldsymbol{\theta}^{(1)}$ if the magnitude of the perturbation is too large. We argue that this cannot occur for an arbitrarily large sample $\boldsymbol{y}^{(n)}$ because both $p_1(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ and $p_2(\boldsymbol{\theta}|\boldsymbol{y}^{(n)})$ will concentrate around $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ maximizes $\log(c_2(\boldsymbol{u})) - \log(c_1(\boldsymbol{u}))$ in a neighbourhood of this fixed point. If this perturbation is too large for a given sample $\boldsymbol{y}^{(n)}$, this behaviour cannot persist as $n \to \infty$. For large samples, these small perturbations by the Newton-Raphson method will therefore shift $\boldsymbol{\theta}^{(2)}$ closer to $\boldsymbol{\theta}_0$ with probability approaching 1. □

Appendix D

Additional Material for Chapter 5

D.1 Additional Content for Theorem 5.1

D.1.1 Proof of Theorem 5.1

We prove Theorem 5.1 in two stages. We first prove a simpler version of Theorem 5.1 where the design prior $p_D(\boldsymbol{\eta})$ is degenerate (i.e., $p_D(\boldsymbol{\eta}^*) = 1$ for some $\boldsymbol{\eta}^* = (\boldsymbol{\eta}_1^*, \boldsymbol{\eta}_2^*)$ and 0 otherwise). This simpler version of Theorem 5.1 is incorporated into Theorem 3.1 from Chapter 3. Here, we provide a similar proof to that detailed Appendix B.1.3 because the notation differs slightly when we accommodate imbalanced sample sizes.

Under the conditions for Theorem 5.1 in the simpler setting, the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n_j}$ converges in probability to $\boldsymbol{\eta}_j^*$ for $j = 1, 2$. The following result also holds for $\mathcal{J}_j(\tilde{\boldsymbol{\eta}}_{j,n})/n_j$ in (5.6):

$$\frac{1}{n_j}\mathcal{J}_j(\tilde{\boldsymbol{\eta}}_{j,n_j}) = \left[-\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\partial^2}{\partial\boldsymbol{\eta}_j^2}\log(f(y_{ij};\boldsymbol{\eta}_j)) - \frac{1}{n_j}\frac{\partial^2}{\partial\boldsymbol{\eta}_j^2}\log(p_j(\boldsymbol{\eta}_j))\right]_{\boldsymbol{\eta}_j=\tilde{\boldsymbol{\eta}}_{j,n_j}} \xrightarrow{P} \mathcal{I}(\boldsymbol{\eta}_j^*). \quad \text{(D.1)}$$

Because $\tilde{\boldsymbol{\eta}}_{j,n_j} - \hat{\boldsymbol{\eta}}_{j,n_j} \xrightarrow{P} 0$, the mean and variance of the normal distribution in (5.6) respectively approximate

$$\hat{\theta}_n = g(\hat{\boldsymbol{\eta}}_{1,n_1}) - g(\hat{\boldsymbol{\eta}}_{2,n_2}) \quad \text{and} \quad \frac{1}{n}\mathcal{I}(\hat{\theta}_n)^{-1} = \sum_{j=1}^{2}\frac{1}{n_j}\left[\frac{\partial g}{\partial\boldsymbol{\eta}}^T\mathcal{I}(\boldsymbol{\eta})^{-1}\frac{\partial g}{\partial\boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}_{j,n_j}},$$

for large $n = n_1 = n_2/q$ by the continuous mapping theorem. For simplicity, we do not incorporate q into the notation for $\hat{\theta}_n$ in this appendix.

Similar to in Appendix B.1.3, we extend the notation from (5.7) to estimate the posterior probabilities that comprise $\mathcal{P}^{\delta}_{n,q,\boldsymbol{Y}^{(n,q)},(5.6)}$ when data $\boldsymbol{Y}^{(n,q)}$ are generated. For $\mathcal{P}^{\delta}_{n,q,\boldsymbol{Y}^{(n,q)},(5.6)}$, the fraction inside the standard normal CDF of (5.7) converges to the following normal distribution:

$$\sqrt{n}\left(\frac{\delta-\theta^*}{\sqrt{\mathcal{I}(\hat{\theta}_n)^{-1}}}-\frac{\hat{\theta}_n-\theta^*}{\sqrt{\mathcal{I}(\hat{\theta}_n)^{-1}}}\right)\xrightarrow{d}\mathcal{N}\left(\frac{\delta-\theta^*}{\sqrt{\mathcal{I}(\theta^*)^{-1}}},1\right). \tag{D.2}$$

This result follows by the asymptotic normality of the MLEs $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$, the continuous mapping theorem because $g(\cdot)$ is differentiable at $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$, and Slutsky's theorem since $\mathcal{I}(\hat{\theta}_n)^{-1}\xrightarrow{P}\mathcal{I}(\theta^*)^{-1}$. When psuedorandom sequences $\boldsymbol{U}\overset{\text{i.i.d.}}{\sim}\mathcal{U}([0,1]^{2d})$ are input into Algorithm 5.1, the maximum likelihood estimates for $\mathcal{P}^{\delta}_{n,q,\boldsymbol{U},Alg.5.1}$ are generated from a distribution that coincides exactly with the right side of (D.2), and the impact of the prior is negligible for large n. For the simplified case with degenerate design priors, another application of the continuous mapping theorem with the function $\Phi(\cdot)$ and Scheffé's lemma (Williams, 1991) prompt the result $\|\mathcal{P}^{\delta}_{n,q,\boldsymbol{Y}^{(n,q)},(5.6)}-\mathcal{P}^{\delta}_{n,q,\boldsymbol{U},Alg.5.1}\|_{TV}\xrightarrow{P}0$.

We use the previous result to prove Theorem 5.1 for the case with nondegenerate design priors $p_D(\boldsymbol{\eta})$. The conditions for Theorem 5.1 ensure that the previous result holds for all $\boldsymbol{\eta}^*$ such that $p_D(\boldsymbol{\eta}^*)>0$. Theorem 5.1 assumes that $\boldsymbol{\eta}^*\sim p_D(\boldsymbol{\eta})$. For nondegenerate $p_D(\boldsymbol{\eta})$, we must integrate with respect to $\boldsymbol{\eta}$: Theorem 5.1 holds true by yet another application of the continuous mapping theorem and Scheffé's lemma. □

D.2 Proof of Lemma 5.1

D.2.1 Proof of Parts (a) to (c)

Parts (a), (b), and (c) of Lemma 5.1 respectively correspond to parts (a), (b), and (c) of Lemma 3.1. The differences between these results involve notation. Whereas we referred to the fixed parameter value as θ_0 in Chapter 3, we use θ^* to denote the anticipated value for θ drawn from the relevant design prior. In Section 5.5.2, we described additional differences in notation that involve $\hat{\boldsymbol{\eta}}_n$, $\hat{\boldsymbol{\eta}}_{1,n_1}$, and $\hat{\boldsymbol{\eta}}_{2,n_2}$ along with the functions $g(\cdot)$ and $g_*(\cdot)$. We therefore proved parts (a), (b), and (c) of Lemma 5.1 in Section B.2.

D.2.2 Proof of Part (d)

To prove part (d) of Lemma 5.1, we introduce simplified notation, where $a(\delta_U, \theta^*) = a$ and $a(\delta_L, \theta^*) = c$. We note that $b(\boldsymbol{u}_r) = b$ is the same for both endpoints of the interval (δ_L, δ_U). These simplifications yield the following result:

$$\begin{aligned} &\log(p_{n,q,\boldsymbol{u}_r}^{\delta_U-\delta_L}) - \log\left(1 - p_{n,q,\boldsymbol{u}_r}^{\delta_U-\delta_L}\right) \\ &\approx \log\left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right) - \log\left(1 - \left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right)\right). \end{aligned} \tag{D.3}$$

The first derivative of (D.3) with respect to n is

$$\begin{aligned} &\frac{d}{dn}\left[\log\left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right) - \log\left(1 - \left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right)\right)\right] \\ &= \frac{a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)}{2\sqrt{n}(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right))} + \frac{a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)}{2\sqrt{n}\left(1 - (\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right))\right)}. \end{aligned} \tag{D.4}$$

We consider the limit of this derivative as $n \to \infty$ in three cases. In the first case, we consider $\theta^* \in (\delta_L, \delta_U)$. In this setting, $\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right) \to 1$ as $n \to \infty$. Therefore, the limit of the first fraction in (D.4) as $n \to \infty$ is 0. The second fraction can be written in an indeterminate form, so we consider its limiting behaviour using L'Hopital's rule. We have that

$$\begin{aligned} &\lim_{n\to\infty} \frac{\frac{a}{\sqrt{n}}\phi\left(a\sqrt{n}+b\right) - \frac{c}{\sqrt{n}}\phi\left(c\sqrt{n}+b\right)}{2(1 - (\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)))} \\ &= \lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(a\sqrt{n}+b\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(c\sqrt{n}+b\right)}{2\left(a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)\right)}. \end{aligned} \tag{D.5}$$

We must consider the limiting behaviour of (D.5) in cases. For the points in the green

region where $\theta^* \in (\delta_L, \delta_U)$, $a > 0$ and $c < 0$. When $|a| < |c|$, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(a\sqrt{n}+b\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(c\sqrt{n}+b\right)}{2\left(a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)\right)} \\
&= \lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\frac{c\phi\left(c\sqrt{n}+b\right)}{\phi\left(a\sqrt{n}+b\right)}}{2\left(a - \frac{c\phi\left(c\sqrt{n}+b\right)}{\phi\left(a\sqrt{n}+b\right)}\right)} \\
&= \lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\exp\left(\frac{-1}{2}\left[(c\sqrt{n}+b)^2 - (a\sqrt{n}+b)^2\right]\right)}{2\left(a - c\exp\left(\frac{-1}{2}\left[(c\sqrt{n}+b)^2 - (a\sqrt{n}+b)^2\right]\right)\right)} \\
&= \frac{a^2}{2}.
\end{aligned} \tag{D.6}$$

The last step of (D.6) follows because the limit of the exponential term in the numerator and denominator is 0 when $|a| < |c|$. When $|a| > |c|$, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(a\sqrt{n}+b\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(c\sqrt{n}+b\right)}{2\left(a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)\right)} \\
&= \lim_{n\to\infty} \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right)\exp\left(\frac{-1}{2}\left[(a\sqrt{n}+b)^2 - (c\sqrt{n}+b)^2\right]\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)}{2\left(a\exp\left(\frac{-1}{2}\left[(a\sqrt{n}+b)^2 - (c\sqrt{n}+b)^2\right]\right) - c\right)} \\
&= \frac{c^2}{2}.
\end{aligned} \tag{D.7}$$

The last step of (D.7) follows because the limit of the exponential term in the numerator and denominator is 0 when $|a| > |c|$. When $a = -c$, the limit in (D.5) is $0.5 \times (a^3 - c^3)/(a - c) = a^2/2 = c^2/2$. Therefore, the limit of the first derivative in (D.4) is $\min\{a^2, c^2\}/2$ when $\theta^* \in (\delta_L, \delta_U)$.

In the second case for (D.4), we consider points in the red region, where a and c have

the same sign. When $\theta^* > \delta_U$, $c < a < 0$, and $0 < c < a$ when $\theta^* < \delta_L$. In either case, $\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right) \to 0$ as $n \to \infty$. Therefore, the limit of the second fraction in (D.4) as $n \to \infty$ is 0. The first fraction can be written in an indeterminate form, so we consider its limiting behaviour using L'Hopital's rule. We have that

$$\begin{aligned} &\lim_{n\to\infty} \frac{\frac{a}{\sqrt{n}}\phi\left(a\sqrt{n}+b\right) - \frac{c}{\sqrt{n}}\phi\left(c\sqrt{n}+b\right)}{2(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right))} \\ &= \lim_{n\to\infty} -1 \times \frac{a\left(a^2 + \frac{ab}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(a\sqrt{n}+b\right) - c\left(c^2 + \frac{cb}{\sqrt{n}} + \frac{1}{n}\right)\phi\left(c\sqrt{n}+b\right)}{2\left(a\phi\left(a\sqrt{n}+b\right) - c\phi\left(c\sqrt{n}+b\right)\right)}. \end{aligned} \tag{D.8}$$

The limit in (D.8) is just -1 times the limit in (D.5). Therefore, the limit of the first derivative in (D.4) is $-\min\{a^2, c^2\}/2$ when $\theta^* \notin [\delta_L, \delta_U]$.

The third and final case for (D.4) is when $\theta^* \in \{\delta_L, \delta_U\}$. In this scenario, we conclude that the limit of both fractions in (D.4) is 0 without appealing to L'Hopital's rule because $\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right) \to 0.5$ as $n \to \infty$. Thus, the limit of (D.4) as $n \to \infty$ is 0. We emphasize that $a = 0$ if $\theta^* = \delta_U$ and $c = 0$ if $\theta^* = \delta_L$. Thus, the limit of the first derivative in (D.4) is $\min\{a^2, c^2\}/2 = 0$ when $\theta^* \in \{\delta_L, \delta_U\}$.

Putting the three cases together, we obtain part (d) of Lemma 5.1:

$$\begin{aligned} &\lim_{n\to\infty} \frac{d}{dn}\left[\log\left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right) - \log\left(1 - \left(\Phi\left(a\sqrt{n}+b\right) - \Phi\left(c\sqrt{n}+b\right)\right)\right)\right] \\ &= \begin{cases} \dfrac{\min\{a^2, c^2\}}{2}, & \text{if } \theta^* \in [\delta_L, \delta_U] \\ \dfrac{-\min\{a^2, c^2\}}{2}, & \text{if } \theta^* \notin [\delta_L, \delta_U]. \end{cases} \quad \square \end{aligned} \tag{D.9}$$

D.3 Additional Content for the Multinomial Model

D.3.1 Relaxing the Approximate Normality Assumption for the MLE

The mappings between posteriors and $[0,1]^{2d}$ prompted by Algorithm 5.1 are suitable when the sampling distributions of the MLEs $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ are approximately normal. The

conditions for Theorem 5.1 guarantee that this approximate normality holds in the limiting case as $n \to \infty$. However, the quality of these normal approximations may be poor for moderate n with multinomial models where some of the individual probabilities in $\boldsymbol{p}$ are close to 0 or 1. The illustrative example described in Section 5.3 is one such model: it is unlikely that children are very dissatisfied with either porridge sample as $\hat{p}_{11} = 0.0278$ and $\hat{p}_{21} = 0.0146$. Moreover, the design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ assign substantial prior weight for p_{11} and p_{21} to probabilities that are even closer to 0.

The sufficient statistics for the multinomial model in group $j = 1, 2$ are $T_{j\dagger}(\boldsymbol{y}^{(n,q)}) = \{T_{jv}(\boldsymbol{y}^{(n,q)})\}_{v=1}^{w}$, where $T_{jv}(\boldsymbol{y}^{(n,q)}) = \sum_{i=1}^{n_j} \mathbb{I}(y_{ij} = v)$. Instead of simulating $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ from their approximately normal limiting distributions as in Algorithm 5.1, we generate approximate sufficient statistics using a continuous approximation to the binomial CDF. When $X \sim \text{BIN}(n, p_*)$, we approximate the discrete binomial variable by a continuous variable X^* such that

$$X^* \sim \begin{cases} \mathcal{U}(0, 0.5), & \text{with } Pr(X = 0) \\ \mathcal{U}(v - 0.5, v + 0.5), & \text{with } Pr(X = v) \text{ for } v = 1, ..., n-1 \\ \mathcal{U}(n - 0.5, n), & \text{with } Pr(X = n). \end{cases} \tag{D.10}$$

It can be shown that

$$\mathbb{E}(X^*) = np_* + \frac{(1-p_*)^n - p_*^n}{4} \text{ and} \tag{D.11}$$

$$\text{Var}(X^*) = np_*(1-p_*) + \frac{1}{12} - \frac{(1-p_*)^n(8np_* + 1) - p_*^n(8np_* - 1)}{16}.$$

From (D.11), it follows that for any $p_* \in (0, 1)$ with large n, $\mathbb{E}(X^*) \approx \mathbb{E}(X) = np_*$ and $\text{Var}(X^*) \approx \text{Var}(X) = np_*(1 - p_*)$. Sufficient statistics for the multinomial model in each group *could* be obtained by iteratively sampling from (discrete) binomial distributions. However, we illustrate that such a solution would prevent us from using linear approximations to $\text{logit}(p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L})$ to explore segments of sampling distributions of posterior probabilities in Appendix D.3.3. Instead, we use the approach to map posteriors to $[0, 1]^{2d}$ for $d = w - 1$ presented in Algorithm D.1 along with Algorithm 5.2 to conduct the numerical studies in Sections 5.5 and 5.6. In Algorithm D.1, we refer to X^* from (D.10) as $X^*(n, p_*)$ to emphasize the parameters of the binomial distribution being approximated.

The components of $X^*(n, p_*)$ are defined using the draw from the design prior $\boldsymbol{\eta}_j^*$. In Lines 3 to 5 of Algorithm D.1, we account for having a noninteger number of observations to allocate to the remaining $w - v + 1$ multinomial categories. If the remaining number of observations is noninteger, we take the ceiling of this number to be n when considering

Algorithm D.1 Alternative Mapping of Posteriors to $[0,1]^{2d}$ for the Multinomial Model

1: **procedure** MapMultinomial($f(y;\boldsymbol{\eta}_1^*)$, $f(y;\boldsymbol{\eta}_2^*)$, $g(\cdot)$, n, q, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$)
2: **for** j in 1:2 **do**
3: **for** v in 1:$(w-1)$ **do**
4: Let $T_{jv}(\boldsymbol{y}^{(n,q)})$ be the $u_{(j-1)(w-1)+v}$-quantile of $X^*(n_j - \lfloor\sum_{k=1}^{v-1} T_{jv}(\boldsymbol{y}^{(n,q)})\rfloor, Z_{jv*})$.
5: Multiply $T_{jv}(\boldsymbol{y}^{(n,q)})$ by $(n_j - \sum_{k=1}^{v-1} T_{jv}(\boldsymbol{y}^{(n,q)}))/(n_j - \lfloor\sum_{k=1}^{v-1} T_{jv}(\boldsymbol{y}^{(n,q)})\rfloor)$.
6: Let $T_{jw}(\boldsymbol{y}^{(n,q)})$ be $n_j - \sum_{k=1}^{v-1} T_{jw}(\boldsymbol{y}^{(n,q)})$
7: Use $T_{j\dagger}(\boldsymbol{y}^{(n,q)})$ to obtain the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n_j}$ via optimization.
8: Use $\tilde{\boldsymbol{\eta}}_{1,n_1}(\boldsymbol{u})$, $\tilde{\boldsymbol{\eta}}_{2,n_2}(\boldsymbol{u})$, $T_{1\dagger}(\boldsymbol{y}^{(n,q)})$, $T_{2\dagger}(\boldsymbol{y}^{(n,q)})$, and $g(\cdot)$ to obtain (5.6).

the continuous approximation to the binomial model X^*. The $\{Z_{jv*}\}_{v=1}^{w-1}$ terms in Line 4 of Algorithm 5.1 are the anticipated values for the conditional multinomial probabilities in (5.4). In Line 5, we apply a proportional decrease to the generated sufficient statistic to account for the noninteger number of remaining observations. While Algorithm D.1 is tailored to the multinomial model, a similar process could be applied for other discrete distributions (e.g., the binomial or Poisson models) if the sampling distributions of $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$ are not approximately normal for moderate sample sizes.

We show that Algorithm D.1 leads to better performance than Algorithm 5.1 for the illustrative example with moderate sample sizes in Appendix D.3.3. Moreover, Lemma 5.1 still holds true when using Algorithm D.1 instead of Algorithm 5.1. The variable X^* approximately follows a binomial distribution for large n. For sufficiently large sample sizes, the binomial distribution approximates the normal distribution. The result in part (a) of Lemma 5.1 that involves the conduits for the data $\hat{\boldsymbol{\eta}}_n$ is therefore true for large sample sizes when using Algorithm D.1. In contrast, that result holds for any $n_1, n_2 > 0$ when using Algorithm 5.1. The remainder of the proof of Lemma 5.1 in Appendix D.2 can be applied without modifications when Algorithm D.1 is used.

D.3.2 Benefits of Quasi-Monte Carlo Methods

We now assess the impact of using Sobol' sequences with our design procedure based on sampling distribution segments. In Section 5.5.3, we implemented 1000 sample size calculations using Algorithm 5.2 for the illustrative example with Sobol' sequences of length $m = 8192$. Here, we repeated that process using Algorithm 5.2 with pseudorandom sequences of length $m = 8192$. We repeated that process again using pseudorandom se-

quences of length $m = 24000$. Given these 1000 sample size calculations, Figure D.1 depicts the density curves for the recommended sample size n (left) and critical value γ (right) corresponding to each of the three settings considered.

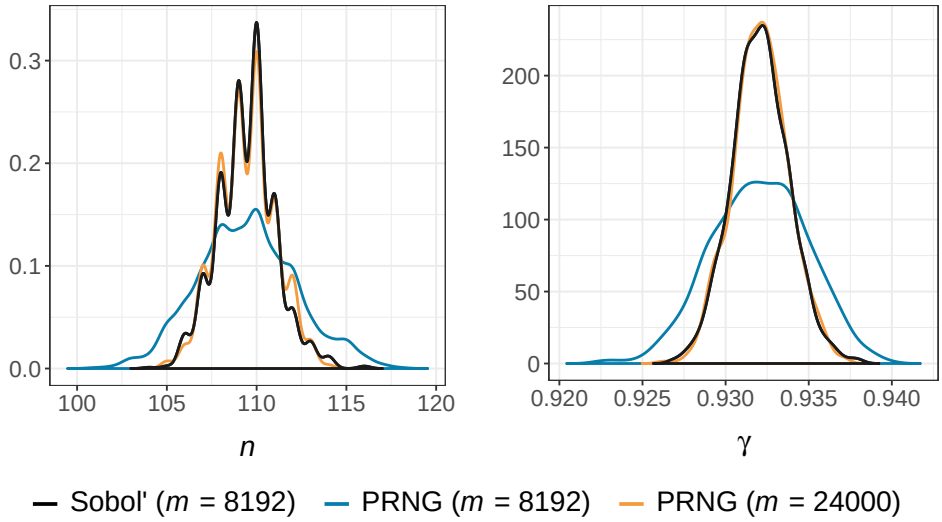


Figure D.1: Density plots of recommendations for the sample size n (left) and critical value γ (right) over 1000 simulation repetitions with Sobol' and pseudorandom (PRNG) sequences of various lengths m.

Using Sobol' sequences gives rise to optimal (n, γ) recommendations that are more precise than those acquired with pseudorandom sequences. The alignment between the black and orange density curves illustrates that the (n, γ) recommendations obtained using Sobol' sequences with length $m = 8192$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 24000$. For this illustrative example, Sobol' sequences therefore allow us to implement simulation-based design with the same level of precision using approximately one third of the simulation repetitions.

D.3.3 Illustrative Analysis Based on the Approximate Normality of the MLE

We demonstrate why using Algorithm 5.1 instead of Algorithm D.1 with the illustrative example yields unsuitable performance. The numerical study from Section 5.5.3 implemented 1000 repetitions of the sample size calculation for the illustrative example using

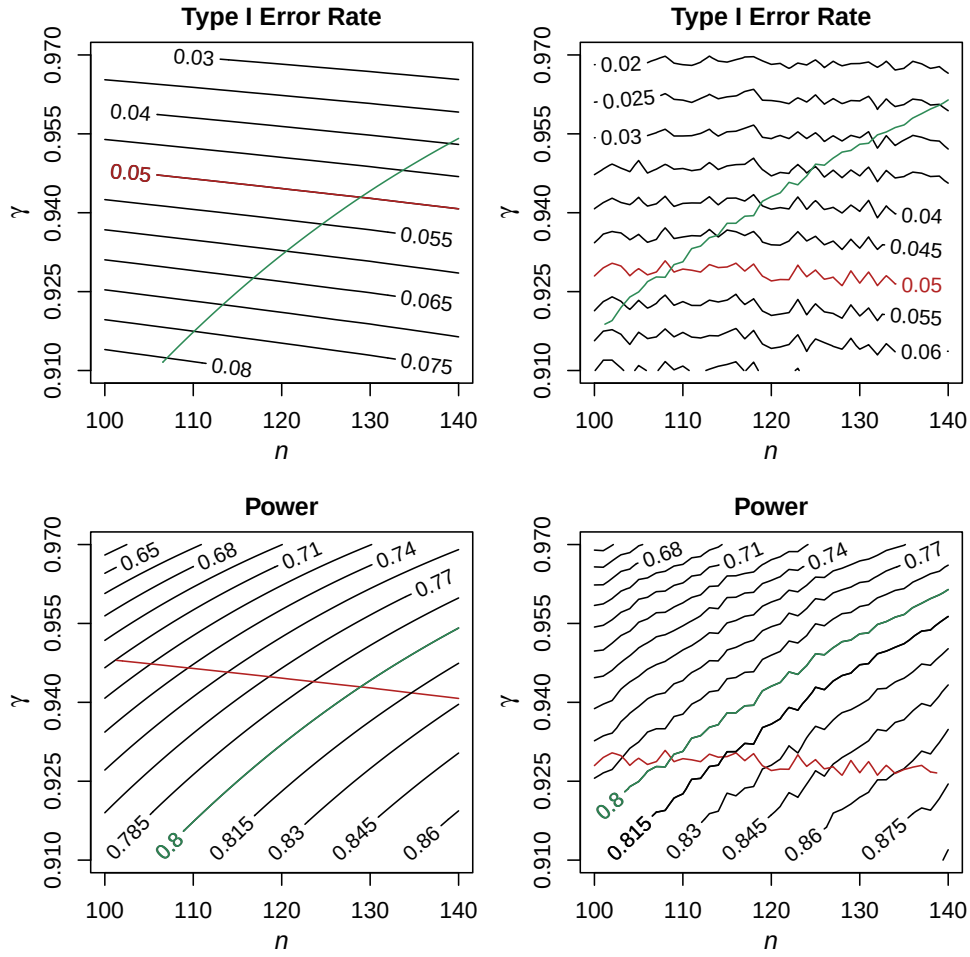


Figure D.2: Left: Averaged contour plots for the type I error rate and power from 1000 sample size calculations with Algorithm 5.1. Right: Contour plots estimated by simulating data.

Algorithm D.1. Here, we repeated this process for the same sample size calculation using Algorithm 5.1 instead of Algorithm D.1. Following a process similar to that in Section 5.6, we averaged contour plots corresponding to the 1000 repetitions of this sample size calculation. The contour plots for the type I error rate and power are given in the left column of Figure D.2. The contour plots are formatted as those in Section 5.6. Based on these plots, the smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours is 129. There is a substantial discrepancy between 129 and the recommendation from the

averaged contour plots in Section 5.6 of $n = 109$. Moreover, the median recommended critical value was $\gamma = 0.9440$ when using Algorithm 5.1 compared to $\gamma = 0.9321$ prompted by Algorithm D.1.

The contour plots in the right column of Figure D.2 were created by simulating $m = 81920$ samples from the prior predictive distributions for $n = \{100, 101, ..., 140\}$ following the process detailed in Section 5.2. Again, the contours in the right plots are jagged since $q = 1.25 \notin \mathbb{Z}$. Unlike in Section 5.6 with Algorithm D.1, the plots in the two columns are not similar when using Algorithm 5.1. This dissimilarity occurs because the sampling distributions for the relevant MLEs are not approximately normal for the multinomial categories with small probabilities. When implementing Algorithm 5.1 with the multinomial example for group $j = 1$ and 2, we generate maximum likelihood estimates for the logits of $\{Z_{jv}\}_{v=1}^{w-1}$ defined in (5.4). This process ensures we do not generate maximum likelihood estimates for any components of $\boldsymbol{p}_j$ that are not between 0 and 1. Figure D.3 illustrates that the sampling distribution of the MLE is not approximately normal for a sample size of $n = 109$ with a multinomial model where $Z_{11*} = p_{11*} = 3/108$. This anticipated value for Z_{11*} coincides with $\hat{p}_{11}$: the observed proportion of children that were assigned a Likert score of 1 in the comparison group for the illustrative example in Section 5.3.

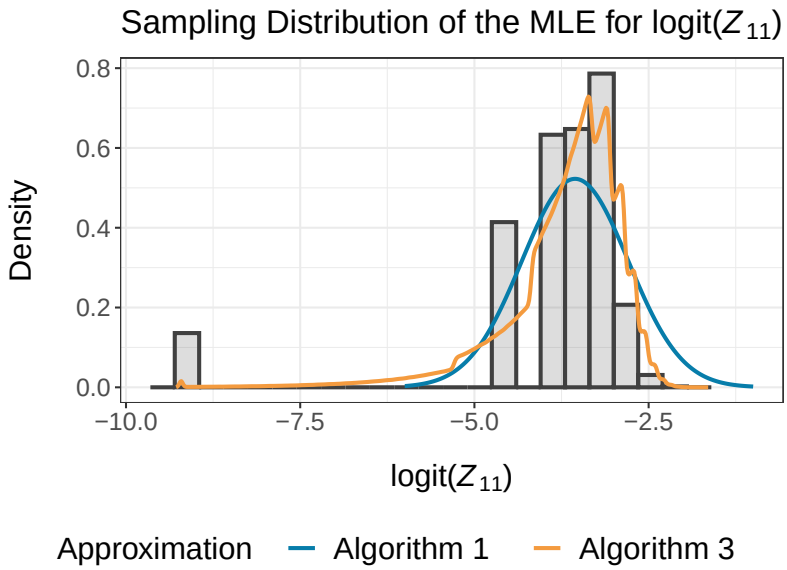


Figure D.3: Histogram of maximum likelihood estimates for the logit of Z_{11} according to the selected binomial distribution. Density curves for the approximations to this distribution prompted by Algorithms 5.1 (blue) and D.1 (orange) are also provided.

The histogram in Figure D.3 was created by simulating 10^6 observations from the BIN$(109, 3/108)$ distribution. For each binomial sample, we took the mean of the Bernoulli observations as $\hat{Z}_{11}$. To ensure that all the maximum likelihood estimates are finite, the histogram shows the sampling distribution of $\text{logit}(\max(\hat{Z}_{11}, 0.0001))$. The blue and orange density curves visualize the approximations to the sampling distribution of the MLE used by Algorithms 5.1 and D.1, respectively. Algorithm 5.1 cannot accommodate the skewness of the true sampling distribution for the optimal sample size of $n = 109$, which is why Algorithm D.1 yields better performance for the illustrative example. For the BIN(n, p_*) model, it is standard practice to require that $np_* > 5$ to invoke the normal approximation to the binomial distribution that Algorithm 5.1 relies on. For this example, nZ_{11*} was 3.028. As such, we recommend using Algorithm D.1 instead of Algorithm 5.1 in scenarios where $p_D(\boldsymbol{\eta}^*)$ assigns substantial prior weight to multinomial models where any of the categorical probabilities violate the $np_* > 5$ condition. It may even be advisable to consider Algorithm D.1 in situations where $np_* = 5 + \epsilon$ for some small $\epsilon > 0$.

To conclude this subsection, we demonstrate why Algorithm D.1 is more suitable than approaches that directly leverage discrete binomial distributions. We selected three Sobol' sequence points from the green region for the illustrative example. For each of these points $\boldsymbol{u}_r \in [0, 1]^{2d+1}$, we used Algorithm D.1 to estimate the logit of $p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L}$ at sample sizes $n = \{90, 91, ..., 150\}$ for the illustrative example. We then modified this process to estimate the logits of $p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L}$ by generating sufficient statistics using CDF inversion on their exact binomial distributions with the same three points from the Sobol' sequence. Figure D.4 visualizes the logit of $p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L}$ as a function of n for both types of approximations, where the functions approximated using the discrete binomial distributions are given by the solid curves. The functions approximated via Algorithm D.1 are depicted using the dotted curves, and the results for each Sobol' sequence point are grouped by colour.

We can generate sufficient statistics for the multinomial model using their exact binomial distributions; however, that process prevents the logit of $p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L}$ from being a smooth function for moderate n as a result of the binomial distribution's discreteness. It is therefore problematic to use the linear approximations to those functions to select segments from sampling distributions of posterior probabilities. Algorithm D.1 and Algorithm 5.1 have linear approximations to the logit of $p_{n,q,\boldsymbol{u}_r}^{\delta_U - \delta_L}$ that are of better quality. Since Algorithm D.1 yields suitable performance as illustrated in Section 5.6, we recommend that method over ones that involve discrete distributions.

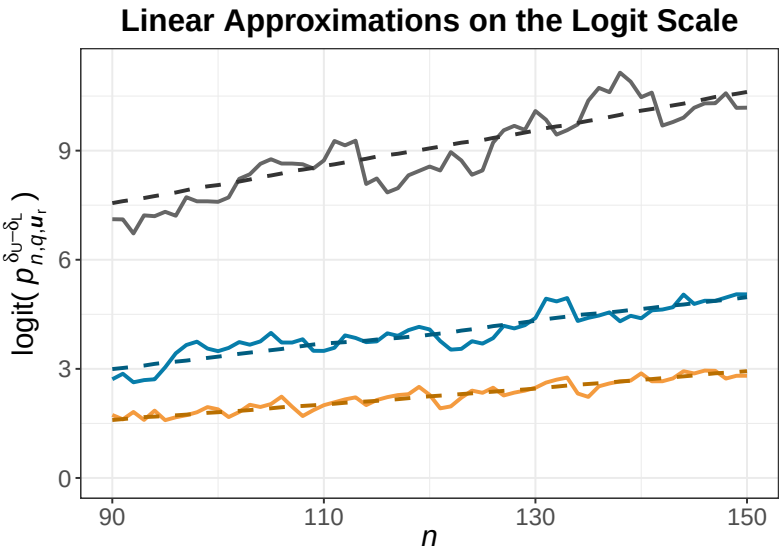


Figure D.4: The logits of $p^{\delta_U-\delta_L}_{n,q,\boldsymbol{u}_r}$ as a function of n for three Sobol' sequence points from the green region for the illustrative example. The curves were created using the discrete binomial model (solid) and Algorithm D.1 (dotted).

D.4 Additional Content for Non-Exponential Family Models

D.4.1 Alternative Method to Map Posteriors to Hypercubes

Algorithm 5.1 must be adapted when the selected model is not a member of the exponential family. Here, we extend the hybrid approach for mapping posteriors to $[0,1]^{2d}$ from Algorithm 3.3 that accounts for the priors when low-dimensional sufficient statistics cannot be recovered from the maximum likelihood estimates $\hat{\boldsymbol{\eta}}_{1,n_1}$ and $\hat{\boldsymbol{\eta}}_{2,n_2}$. That hybrid approach leverages the following result, which holds true when $\boldsymbol{\eta}_j \approx \hat{\boldsymbol{\eta}}_{j,n_j}$ for sufficiently large n_j:

$$\log(p_j(\boldsymbol{\eta}_j|\boldsymbol{y}^{(n)})) \approx l(\hat{\boldsymbol{\eta}}_{j,n_j};\boldsymbol{y}^{(n,q)}) - \frac{n_j}{2}(\boldsymbol{\eta}_j - \hat{\boldsymbol{\eta}}_{j,n_j})^T\mathcal{I}(\hat{\boldsymbol{\eta}}_{j,n_j})(\boldsymbol{\eta}_j - \hat{\boldsymbol{\eta}}_{j,n_j}) + \log(p_j(\boldsymbol{\eta}_j)). \tag{D.12}$$

This result follows from the second-order Taylor approximation to the log-posterior of $\boldsymbol{\eta}_j$ around $\hat{\boldsymbol{\eta}}_{j,n_j}$, where the observed information is replaced with the (expected) Fisher information. An approximation to the posterior mode is the value that maximizes the

right side of (D.12): $\ddot{\boldsymbol{\eta}}_{j,n_j}$. This approximation to the posterior mode was denoted by $\boldsymbol{\eta}^*_{j,n_j}$ in Chapter 3.

We consider the following normal approximation to the posterior of θ:

$$\mathcal{N}\left(g(\ddot{\boldsymbol{\eta}}_{1,n_1}) - g(\ddot{\boldsymbol{\eta}}_{2,n_2}), \sum_{j=1}^{2}\left[\frac{\partial g}{\partial \boldsymbol{\eta}}^T \ddot{\mathcal{J}}_j(\boldsymbol{\eta})^{-1}\frac{\partial g}{\partial \boldsymbol{\eta}}\right]_{\boldsymbol{\eta}=\ddot{\boldsymbol{\eta}}_{j,n_j}}\right), \tag{D.13}$$

$$\text{where} \quad \ddot{\mathcal{J}}_j(\boldsymbol{\eta}) = n_j\mathcal{I}(\boldsymbol{\eta}) - \frac{\partial^2}{\partial \boldsymbol{\eta}^2}\log(p_j(\boldsymbol{\eta})).$$

The observed information is again replaced with the Fisher information in $\ddot{\mathcal{J}}_j(\boldsymbol{\eta})$ of (D.13) since we do not generate samples $\boldsymbol{y}^{(n,q)}$. Algorithm D.2 details how we map a single point $\boldsymbol{u} \in [0,1]^{2d}$ to the posterior approximation in (D.13). Algorithm D.2 is effectively the same mapping process as Algorithm 3.3, where we now accommodate imbalanced sample size determination such that n_1 and n_2 may not be equal.

Algorithm D.2 Mapping Posteriors to $[0,1]^{2d}$ with Hybrid Method when $n_1 \neq n_2$

1: **procedure** MapHybridUnequal($f(y;\boldsymbol{\eta}^*_1)$, $f(y;\boldsymbol{\eta}^*_2)$, $g(\cdot)$, n, q, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$)
2: Generate $\hat{\boldsymbol{\eta}}_{1,n_1}(\boldsymbol{u})$ and $\hat{\boldsymbol{\eta}}_{2,n_2}(\boldsymbol{u})$ using Lines 2 to 4 of Algorithm 5.1.
3: **for** j in 1:2 **do**
4: Obtain $\ddot{\boldsymbol{\eta}}_{j,n_j}$ as $\arg\max_{\boldsymbol{\eta}_j}$ of the right side of (D.12) anchored at $\boldsymbol{\eta}_{j,n_j} = \hat{\boldsymbol{\eta}}_{j,n_j}(\boldsymbol{u})$.
5: Use $\ddot{\boldsymbol{\eta}}_{1,n_1}$, $\ddot{\boldsymbol{\eta}}_{2,n_2}$, and the partial derivatives of $g(\cdot)$ to obtain (D.13).

The results from Theorem 5.1 and Lemma 5.1 hold true when using Algorithm D.2 instead of Algorithm 5.1. The result in Theorem 5.1 is straightforward because $\ddot{\boldsymbol{\eta}}_{j,n_j} - \hat{\boldsymbol{\eta}}_{j,n_j}$ converges in probability to 0, and $\ddot{\mathcal{J}}_j(\ddot{\boldsymbol{\eta}}_{j,n_j})/n_j$ converges in probability to $\mathcal{I}(\boldsymbol{\eta}^*_j)$ following similar logic to (D.1). Parts (a) and (b) of Lemma 5.1 follow from Appendix D.2.1. To prove the result in part (c) of Lemma 5.1, we argue that $\underline{\tau}_r^{(n)} \approx \mathcal{I}(\theta^*)^{-1}/n$ for sufficiently large n once again by similar logic to (D.1) when Algorithm D.2 is used. No modifications to Appendix D.2.2 are required to prove part (d) of Lemma 5.1.

D.4.2 Illustrative Analysis with the Weibull Model

We now reconsider the data set from the ENIGH 2020 survey (INEGI, 2021) that was introduced in Section 3.3. That data set split lower-middle income households from the

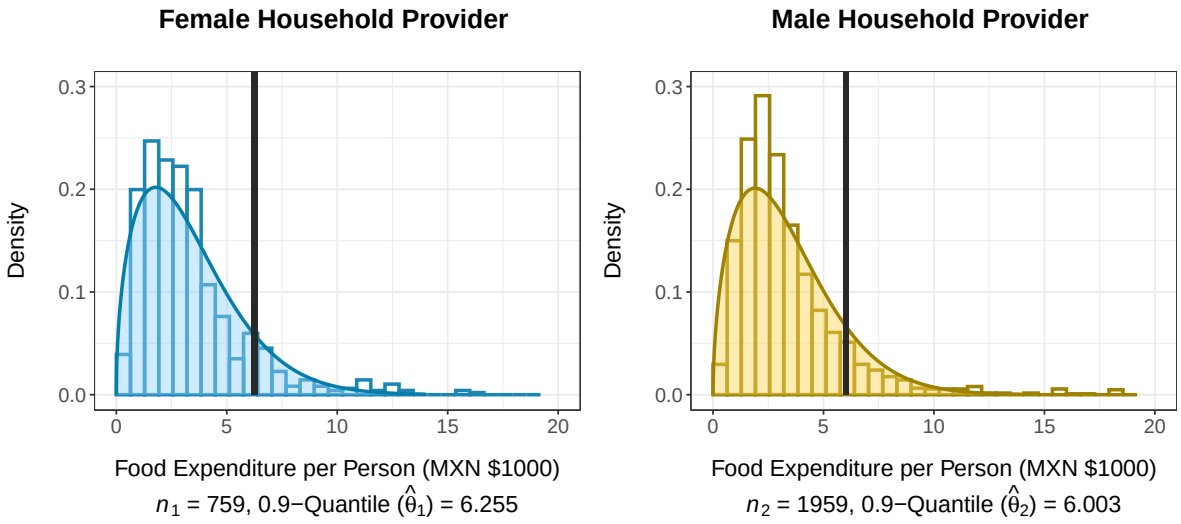


Figure D.5: Distribution of quarterly food expenditure per person in each group.

Mexican state of Aguascalientes into two groups based on the sex of the household's main provider. The datum y_{ij} collected for each household $i = 1, ..., n_j,\ j = 1, 2$ is its quarterly expenditure on food per person measured in thousands of Mexican pesos (MXN \$1000). We again exclude the 0.41% of households that report zero quarterly expenditure on food to accommodate the Weibull model's positive support. This respectively yields $n_1 = 759$ and $n_2 = 1959$ observations in the female ($j = 1$) and male ($j = 2$) provider groups that are visualized in Figure D.5.

Unlike in Section 3.3, we compare the 0.9-quantile for each distribution. That is, $\theta_j = F_j^{-1}(0.9)$, where $F_j(\cdot)$ is the cumulative distribution function for distribution $j = 1, 2$. We use the ratio θ_1/θ_2 to consider whether the 0.9-quantiles of food expenditure in the female and male provider groups are practically equivalent. The observed 0.9-quantiles of quarterly food expenditure per person (in MXN \$1000) are $\hat{\theta}_1 = 6.255$ and $\hat{\theta}_2 = 6.003$. We assign uninformative GAMMA$(2, 1)$ priors to both the shape ν_j and scale ι_j parameters of the Weibull model for group $j = 1, 2$. We let $\boldsymbol{\eta}_j = (\nu_j, \iota_j)$ for $j = 1, 2$. We obtain 10^5 posterior draws for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ using MCMC methods. The Weibull distributions characterized by the posterior means for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are superimposed on the histograms in Figure D.5. Since the Weibull distribution is a reasonable model for these data, we use this example to illustrate our extensions for non-exponential family models in this section.

We next choose design and analysis priors for this example. Because the ENIGH survey is conducted biennially, we choose design priors for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ using data from the ENIGH

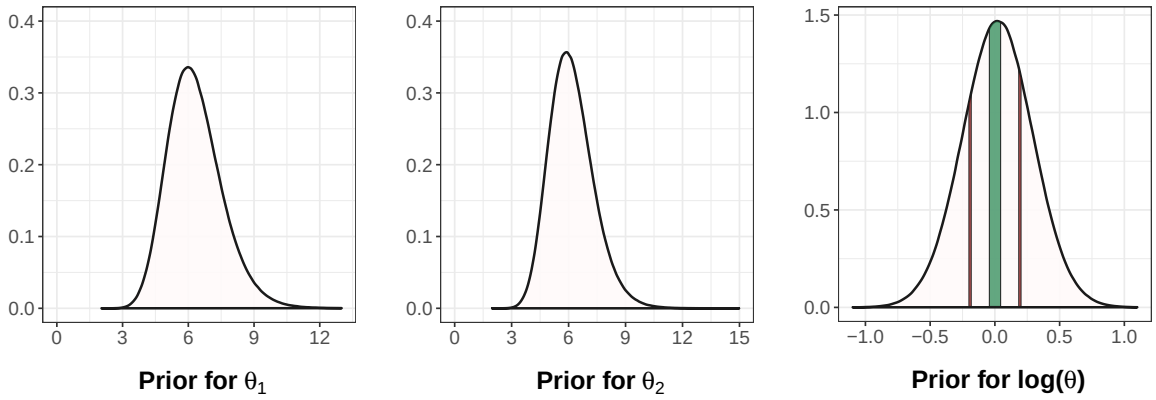


Figure D.6: Induced design priors for θ_1 (left), θ_2 (center), and $\log(\theta)$ (right). The green and red regions of the θ-space are visualized on the logarithmic scale on the right plot.

2018 survey (INEGI, 2019). We used the inflation-adjusted data set of food expenditures from 2018 created in Section 3.6.1. For group j, we obtained posteriors for ν_j and ι_j given these data and GAMMA$(2, 1)$ priors for each parameter. To define priors, we consider gamma distributions that have the same modes with variances that are larger by factors of 30 and 100 for groups 1 and 2, respectively. These distributions considered for illustration are GAMMA$(38.07, 26.23)$ for ν_1, GAMMA$(34.92, 10.02)$ for ι_1, GAMMA$(38.35, 25.09)$ for ν_2, and GAMMA$(37.51, 10.70)$ for ι_2.

We use those gamma distributions to obtain the design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$ for this example. Figure D.6 visualizes the priors for θ_1, θ_2, and $\log(\theta) = \log(\theta_1) - \log(\theta_2)$ that are induced by those gamma distributions. For illustration, we choose the interval (δ_L, δ_U) to be $(1.2^{-1}, 1.2)$. This choice indicates that a 20% relative increase or decrease in the 0.9-quantile is not of practical importance. We define the region $\mathcal{G} = (1.05^{-1}, 1.05)$ to be centered around 1 on the relative scale. For this example, the region $\mathcal{R} = (1.225^{-1}, 1.2^{-1}) \cup (1.2, 1.225)$ is contiguous with both endpoints of the interval (δ_L, δ_U). These red and green regions are depicted on the logarithmic scale in the right plot of Figure D.6. We now define design priors of $p_{D_1}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta} | \log(\theta) \sim \mathcal{U}(\log(\mathcal{G})))$ and $p_{D_0}(\boldsymbol{\eta}) \propto p_D(\boldsymbol{\eta} | \log(\theta) \sim \mathcal{U}(\log(\mathcal{R})))$, where $p_D(\boldsymbol{\eta})$ is created by independently joining the gamma priors from the previous paragraph. Here, the conditioning used to define these priors ensures that θ is uniformly distributed over the red and green regions on the logarithmic scale. We independently join marginal GAMMA$(2, 1)$ priors for ν_j and ι_j to obtain an analysis prior $p_j(\boldsymbol{\eta}_j)$ for group $j = 1, 2$. Our final inputs for Algorithm 5.2 are $m = 8192$, $m_0 = 512$, $\alpha = 0.1$, $\beta = 0.3$, $q = 1$.

When using Algorithm D.2 to map posteriors to $[0,1]^{2d}$, Algorithm 5.2 returned an optimal design characterized by $(n, \gamma) = (163, 0.8795)$. For reference, we considered the precision of the (n, γ) recommendations with Sobol' and pseudorandom sequences for this Weibull example using the process to create Figure D.1. The results from that numerical study suggested the (n, γ) recommendations obtained using Sobol' sequences with length $m = 8192$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 4 \times 10^4$.

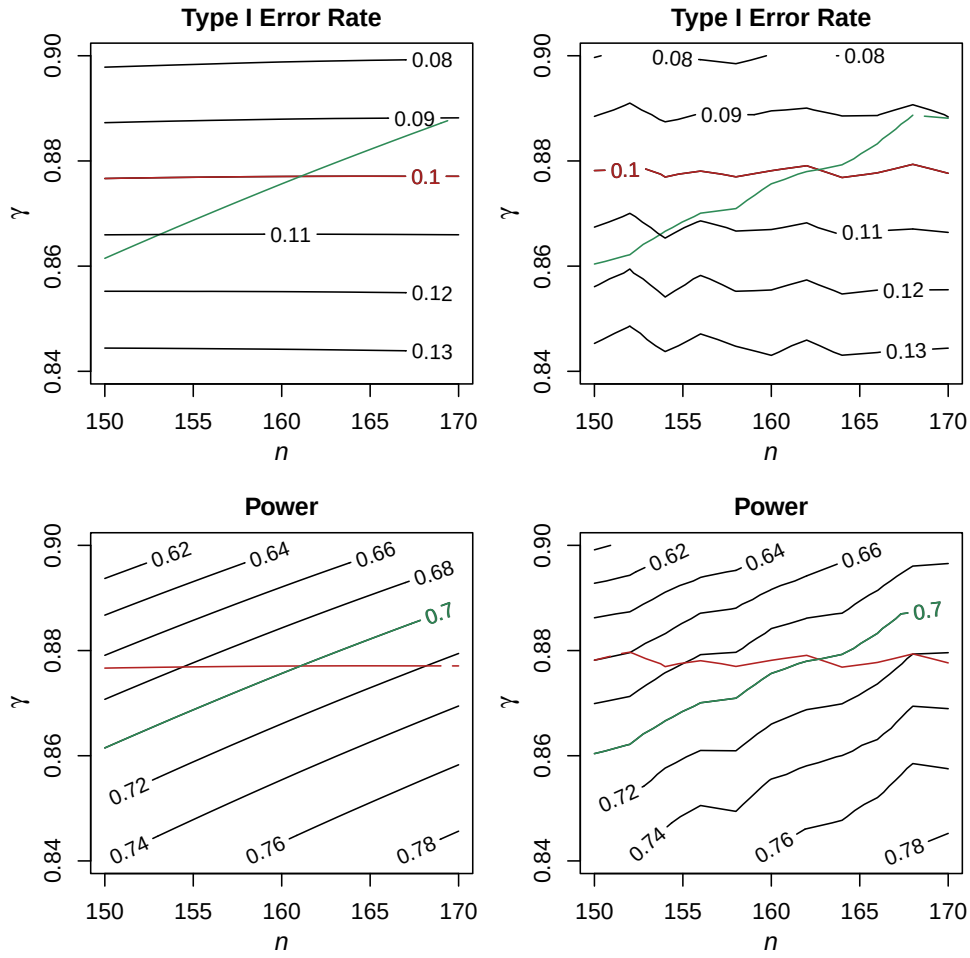


Figure D.7: Left: Averaged contour plots for the type I error rate and power from 1000 sample size calculations with Algorithm D.2 and the Weibull example. Right: Contour plots estimated by simulating data.

As in Section 5.5.3, we repeated the sample size calculation from the previous paragraph 1000 times with different Sobol' sequences $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$. For each repetition, the optimal design coincided with the (n, γ) recommendation obtained by exploring entire sampling distributions of posterior probabilities with nontargeted approaches using the same Sobol' sequences. For this Weibull example, Algorithm 5.2 took roughly 25 seconds on a standard laptop without parallelization to return an optimal design for the illustrative example. The modified version of Algorithm 5.2 that explored entire sampling distributions of posterior probabilities took approximately 90 seconds. Using the process described in Section 5.6, we averaged contour plots for the type I error rate and power corresponding to the 1000 repetitions of the sample size calculation for the Weibull example. These plots are given in the left column of Figure D.7. Based on these plots, the smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours is 162.

The contour plots in the right column of Figure D.7 were created by simulating $m = 40960$ samples from the prior predictive distributions for $n = \{150, 152, ..., 170\}$ following the process detailed in Section 5.2. The contours in the right plots are less jagged than those for the multinomial example since $q = 1$. However, the contours in the right column are still more jagged than those in the left column because the left plots consider the sampling distributions of posterior probabilities with the same points $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$ for *each* sample size. The sampling distributions of posterior probabilities in the right plots are estimated independently for each value of n considered. The smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours in the right plots is $n = 163$. The plots in the left and right columns are similar, which illustrates that using Algorithm D.2 to map posteriors to $[0, 1]^{2d}$ for this non-exponential family example prompts suitable performance.

D.4.3 Mapping Posteriors with Misspecified Priors

The approximation to the log-posterior of $\boldsymbol{\eta}_j$ in (D.12) is only valid when $\boldsymbol{\eta}_j \approx \hat{\boldsymbol{\eta}}_{j,n_j}$ for sufficiently large n_j. However, the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n_j} = \arg\max_{\boldsymbol{\eta}_j} p_j(\boldsymbol{\eta}_j|\boldsymbol{y}^{(n,q)})$ may not be near the maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{j,n_j}$ if the prior is misspecified for group $j = 1, 2$. In such scenarios, the approximation to the posterior mode $\ddot{\boldsymbol{\eta}}_{j,n_j}$ prompted by maximizing (D.12) may not be accurate. These concerns were not as relevant when using Algorithm 3.3 from Chapter 3 with degenerate design priors. In that chapter, we used Algorithm 3.3 with informative analysis priors for $\boldsymbol{\eta}_j$ whose modes aligned with the design values $\boldsymbol{\eta}_{j,0}$. This practice was sensible in that context because it would be philosophically inconsistent to use *informative* analysis priors that do not align with our understanding of

the data generation process. Moreover, Algorithm 3.3 yielded suitable performance with uninformative analysis priors.

When using informative analysis priors with nondegenerate design priors $p_{D_1}(\boldsymbol{\eta})$ and $p_{D_0}(\boldsymbol{\eta})$, prior misspecification is a near certainty. A particular set of analysis priors $p(\boldsymbol{\eta}_1)$ and $p(\boldsymbol{\eta}_2)$ cannot support both H_0 and H_1. If the induced analysis prior on θ is concentrated in the green (red) region of the θ-space, the set of analysis priors is misspecified when generating data under H_0 (H_1). Furthermore, it is likely that an informative analysis prior on θ that is concentrated in $\mathcal{G}$ ($\mathcal{R}$) is at least slightly misspecified when generating data under the corresponding hypothesis H_1 (H_0). For instance, even if the mode of the induced prior on θ is near the center of $\mathcal{G}$, the design prior $p_{D_1}(\boldsymbol{\eta})$ will generate $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$ values from throughout the green region. Given $\boldsymbol{\eta}_1^*$ and $\boldsymbol{\eta}_2^*$ values from the extremities of $\mathcal{G}$, an informative analysis prior might still be slightly misspecified.

It is therefore pertinent to have a method to map posteriors to $[0,1]^{2d}$ that accommodates misspecified priors. For models in the exponential family, the posterior approximation in (5.6) makes such accommodations because it is centered around the true posterior mode. The posterior approximation for non-exponential family models in (D.12) is instead centered around an approximation to the posterior mode that relies on the suitability of the quadratic approximation to the log-likelihood anchored at the maximum likelihood estimate $\hat{\boldsymbol{\eta}}_{j,n_j}$. Here, we propose a foundation for posterior approximation that accommodates prior misspecification without requiring sufficient statistics for non-exponential family models. This foundation still assumes that the quadratic approximation to the log-likelihood function $l(\boldsymbol{\eta}_{j,n_j}; \boldsymbol{y}^{(n,q)})$ is suitable; however, we no longer require that this approximation is anchored around $\hat{\boldsymbol{\eta}}_{j,n_j}$.

Since the posterior of $\boldsymbol{\eta}_j$ will be concentrated around the posterior mode $\tilde{\boldsymbol{\eta}}_{j,n_j}$ for sufficiently large n_j, this framework allows us to anchor the quadratic approximation to the log-likelihood around a value for $\boldsymbol{\eta}_j$ that is closer to $\tilde{\boldsymbol{\eta}}_{j,n_j}$ than $\hat{\boldsymbol{\eta}}_{j,n_j}$. When this quadratic approximation is not anchored around $\hat{\boldsymbol{\eta}}_{j,n_j}$, we cannot assume that the coefficient for its linear term is 0. We therefore use the approximation to the log-posterior of $\boldsymbol{\eta}_j$ provided in (D.14) that is anchored around $\boldsymbol{\eta}_j = \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}$.

$$\begin{aligned}\log(p_j(\boldsymbol{\eta}_j|\boldsymbol{y}^{(n)})) \approx l(\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}; \boldsymbol{y}^{(n,q)}) + n_j \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\eta}_j}\log(f(y;\boldsymbol{\eta}_j))\right]_{\boldsymbol{\eta}_j = \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)} \vee \hat{\boldsymbol{\eta}}_{j,n_j}} (\boldsymbol{\eta}_j - \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}) \\ - \frac{n_j}{2}(\boldsymbol{\eta}_j - \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)})^T \mathcal{I}(\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)})(\boldsymbol{\eta}_j - \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}) + \log(p_j(\boldsymbol{\eta}_j)).\end{aligned} \tag{D.14}$$

We will use (D.14) in an iterative manner. That is, the value that maximizes the right side of (D.14) for a given $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}$ will be denoted by $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e+1)}$. The final value of $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e+1)}$ considered will serve as $\ddot{\boldsymbol{\eta}}_{j,n_j}$: the approximation to the posterior mode that is used in (D.13).

Unlike the approximation in (D.12), the approximation in (D.14) does not constrain the coefficient for the linear term of the quadratic approximation of the log-likelihood to be 0. Because we cannot obtain sufficient statistics when generating MLEs for non-exponential family models, we suggest obtaining the coefficient for the linear term as follows. First, we obtain the expectation of $\log(f(y;\boldsymbol{\eta}_j))$ as a function of $\boldsymbol{\eta}_j$. This function is 0 if the same value for $\boldsymbol{\eta}_j$ is used for all occurrences in this equation (which is typical). However, we substitute $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}$ into this equation for all occurrences of $\boldsymbol{\eta}_j$ arising from parameters in the log-likelihood function; $\hat{\boldsymbol{\eta}}_{j,n_j}$ is instead substituted into this equation for all occurrences of $\boldsymbol{\eta}_j$ arising from integrating over the data. The notation $\boldsymbol{\eta}_j = \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)} \vee \hat{\boldsymbol{\eta}}_{j,n_j}$ in (D.14) represents this procedure.

By implementing this procedure, we obtain a coefficient for the linear term that is not 0 when $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)} \neq \hat{\boldsymbol{\eta}}_{j,n_j}$. This procedure may not obtain a perfect approximation to the log-posterior of $\boldsymbol{\eta}_j$, but it is an approximation we can obtain without sufficient statistics. Algorithm D.3 formalizes how we iteratively use the approximation in (D.14) to map posteriors to $[0,1]^{2d}$ in the presence of prior misspecification for non-exponential family models. We initialize $\ddot{\boldsymbol{\eta}}_{1,n_1}^{(0)}$ and $\ddot{\boldsymbol{\eta}}_{2,n_2}^{(0)}$ as their generated maximum likelihood estimates. To implement Algorithm D.3, we must also choose a tolerance $\epsilon > 0$ to define convergence for the approximations to the posterior modes $\ddot{\boldsymbol{\eta}}_{1,n_1}$ and $\ddot{\boldsymbol{\eta}}_{2,n_2}$

Algorithm D.3 Iterative Mapping of Posteriors to $[0,1]^{2d}$ with Misspecified Priors

1: **procedure** MapIterative($f(y;\boldsymbol{\eta}_1^*)$, $f(y;\boldsymbol{\eta}_2^*)$, $g(\cdot)$, n, q, $\boldsymbol{u}$, $p_1(\boldsymbol{\eta}_1)$, $p_2(\boldsymbol{\eta}_2)$, ϵ)
2: Generate $\hat{\boldsymbol{\eta}}_{1,n_1}(\boldsymbol{u})$ and $\hat{\boldsymbol{\eta}}_{2,n_2}(\boldsymbol{u})$ using Lines 2 to 4 of Algorithm 5.1.
3: Let $\ddot{\boldsymbol{\eta}}_{1,n_1}^{(0)} = \hat{\boldsymbol{\eta}}_{1,n_1}(\boldsymbol{u})$ and $\ddot{\boldsymbol{\eta}}_{2,n_2}^{(0)} = \hat{\boldsymbol{\eta}}_{2,n_2}(\boldsymbol{u})$.
4: **for** j in 1:2 **do**
5: `converged` $\leftarrow$ FALSE; $e \leftarrow 0$
6: **while** `converged` = FALSE **do**
7: Obtain $\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e+1)}$ as $\arg\max_{\boldsymbol{\eta}_j}$ of (D.14) anchored at $\boldsymbol{\eta}_{j,n_j} = \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}$.
8: **if** $\|\ddot{\boldsymbol{\eta}}_{j,n_j}^{(e+1)} - \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e)}\|_2 \leq \epsilon$ **then**
9: `converged` $\leftarrow$ TRUE
10: $\ddot{\boldsymbol{\eta}}_{j,n_j} \leftarrow \ddot{\boldsymbol{\eta}}_{j,n_j}^{(e+1)}$
11: $e \leftarrow e + 1$
12: Use $\ddot{\boldsymbol{\eta}}_{1,n_1}$, $\ddot{\boldsymbol{\eta}}_{2,n_2}$, and the partial derivatives of $g(\cdot)$ to obtain (D.13).

We now illustrate how the approximation to the log-posterior in (D.14) and Algorithm D.3 are used with the Bernoulli model. The Bernoulli model is a member of the exponential family, but the posterior of a single Bernoulli parameter θ has a known beta distribution when conjugate priors are used. Knowing the exact posterior of θ allows us to compare the performance of the posterior approximations in (D.12) and (D.14). We focus on a single group of data here for this simple, informal illustration.

As in Section 3.5.4, we parameterize the Bernoulli model in terms of its canonical parameter $\eta = \log(\theta) - \log(1-\theta) \in \mathbb{R}$ to improve the quality of the normal approximation to the posterior. The corresponding inverse transformation is $\theta = \exp(\eta) \div (1 + \exp(\eta))$. Given this parameterization, we have that

$$\frac{\partial}{\partial \eta} \log(f(y;\eta)) = y - \frac{\exp(\eta)}{1+\exp(\eta)}. \tag{D.15}$$

We omit the subscript j for the group number when considering η and the sample size n because we have a single group of data in this setting. Since $\mathbb{E}(y) = \theta = \exp(\eta) \div (1 + \exp(\eta))$, the result in (D.15) prompts

$$\mathbb{E}\left[\frac{\partial}{\partial \eta} \log(f(y;\eta))\right]_{\eta=\ddot{\eta}_n^{(e)} \vee \hat{\eta}_n} = \frac{\exp(\hat{\eta}_n)}{1+\exp(\hat{\eta}_n)} - \frac{\exp(\ddot{\eta}_n^{(e)})}{1+\exp(\ddot{\eta}_n^{(e)})}.$$

We suppose that θ was assigned a BETA(α, β) prior that induces a prior on η. For completeness, maximizing (D.14) for this illustrative example is equivalent to maximizing

$$\begin{aligned} & n\left[\frac{\exp(\hat{\eta}_n)}{1+\exp(\hat{\eta}_n)} - \frac{\exp(\ddot{\eta}_n^{(e)})}{1+\exp(\ddot{\eta}_n^{(e)})}\right](\eta - \ddot{\eta}_n^{(e)}) \\ & - \frac{n\exp(\ddot{\eta}_n^{(e)})}{2(1+\exp(\ddot{\eta}_n^{(e)}))^2}(\eta - \ddot{\eta}_n^{(e)})^2 + \alpha\eta + (\alpha+\beta)\log(1+\exp(\eta)). \end{aligned} \tag{D.16}$$

The final two terms in (D.16) are equal to the log-prior of η less a normalizing constant. These final two terms incorporate the Jacobian that arises from converting between η and θ. It is also not necessary to account for the constant $l(\ddot{\eta}_n^{(e)}; \boldsymbol{y}^{(n)})$ from (D.14) when maximizing (D.16).

We next apply Algorithms D.2 and D.3 to consider posterior approximation for η when θ is assigned a BETA$(20, 80)$ prior. For illustration, we suppose that $u \in [0, 1]$ corresponding to $\hat{\theta}_{100} = 0.35$ was input into both algorithms. In this case, the prior mode for θ of 0.1939 differs greatly from $\hat{\theta}_{100}$. The exact posterior of θ given $\hat{\theta}_{100}$ is a BETA$(55, 145)$ distribution. This beta prior and posterior for θ induce a prior and posterior on η. The induced prior and

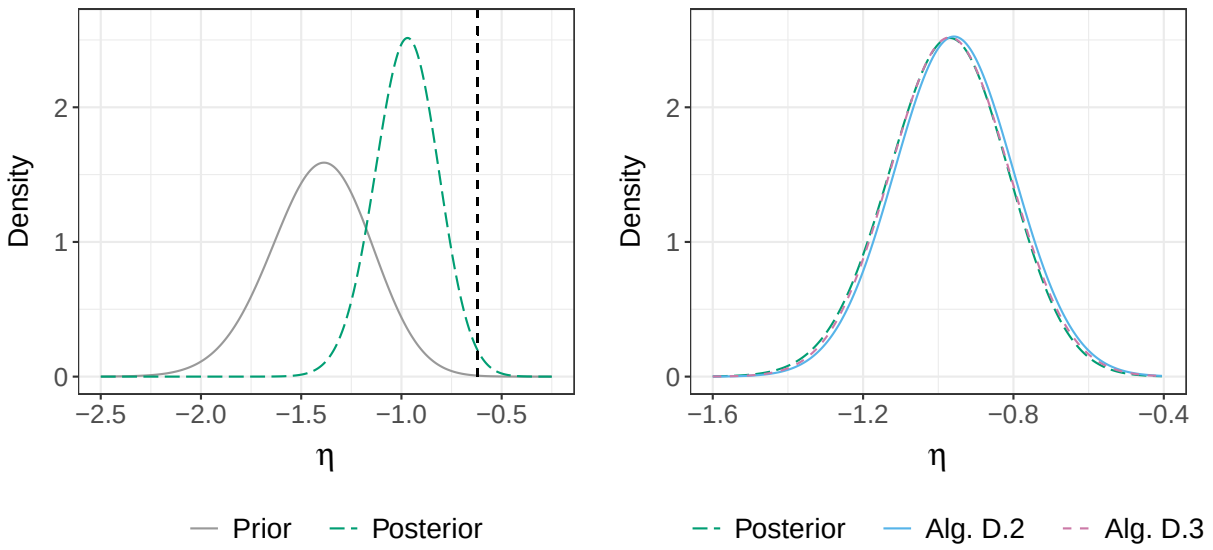


Figure D.8: Left: Induced prior and posterior on η with $\hat{\eta}_{100}$ denoted by the dotted line. Right: Exact posterior of η along with the approximations provided by Algorithms D.2 and D.3.

posterior are visualized in the left plot of Figure D.8. The maximum likelihood estimate on the η-scale is depicted by the dotted vertical line. There is a clear discrepancy between the posterior mode $\tilde{\eta}_{100}$ and maximum likelihood estimate $\hat{\eta}_{100}$.

The right plot of Figure D.8 compares the exact posterior of η to the posterior approximations that result from the mappings in Algorithms D.2 and D.3. The posterior approximation from Algorithm D.2 exhibits noticeable bias and is shifted to the right of the true posterior for η. The posterior approximation from Algorithm D.3 was obtained with $\epsilon = 10^{-4}$. The final value of e when implementing Algorithm D.3 for this example was 2. Therefore, only one iteration of Algorithm D.3 was required to better approximate the posterior mode $\tilde{\eta}_{100}$, and the second iteration was used to confirm convergence. The results visualized in Figure D.8 informally suggest that iterative procedures can be used to improve the quality of posterior approximation in the presence of prior misspecification. It follows that using such posterior approximations may improve the accuracy of our mappings between posterior probabilities and $[0, 1]^{2d}$ when the priors are misspecified. The work in this subsection is preliminary and more formal investigation with models that are not members of the exponential family is required. Nevertheless, this foundation for future work attests to the broad applicability of design with sampling distribution segments.

D.5 Two-Group Comparisons with Additional Covariates

D.5.1 Posterior Mapping with Linear Regression

When Bayesian linear regression is used to account for additional covariates, we propose a framework to map posteriors of these regression coefficients to the unit hypercube $[0,1]^d$. For reasons explained later, we denote the simulation dimension for posterior mapping as d instead of $2d$ used in Algorithm 5.1. The simulation dimension d for regression settings depends on the number of covariates in the linear model, which we denote by k in this section. We consider linear models of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \tag{D.17}$$

where y is the response variate, $x_1 = \mathbb{I}(\text{Group} = 1)$ is the binary treatment indicator, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent error terms for observations $i = 1, \ldots, n_1 + n_2$. For illustration, we suppose that the remaining $k-1$ covariates are such that $x_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$ independently for $l = 2, 3, \ldots, k$. The normality assumption for the additional covariates could be relaxed in future work, and we discuss the theoretical implications of that relaxation in Appendix D.5.2.

For two-group comparisons facilitated via (D.17), the posterior distribution of interest is that of β_1. We use a conjugate normal-inverse-gamma prior (Koch, 2007) for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ and the error term variance σ_ε^2 for simplicity. The normal-inverse-gamma distribution has a location parameter $\boldsymbol{\mu}_0$ along with a matrix parameter $\boldsymbol{\lambda}_0$, and two scalar parameters a_0 and b_0. Under this conjugate prior, the posterior of β_1 follows a 3-parameter t-distribution. We discuss how to consider alternative priors with linear regression models in Appendix D.5.2. While the sampling distribution for the MLEs of $\boldsymbol{\beta}$ and σ_ε^2 are approximately normal for large sample sizes, the estimated $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_\varepsilon^2$ do not comprise sufficient statistics for the data $\{y_i, x_{1i}, x_{2i}, \ldots, x_{ki}\}_{i=1}^{n_1+n_2}$. We therefore require an alternative method to Algorithm 5.1 to map posteriors of β_1 to $[0,1]^d$. In this case, the dimension of the hypercube with the model in (D.17) is $d = 0.5k(k+5)$, which is not a multiple of 2 for all $k \geq 2$.

We let $x_{0i} = 1$ for $i = 1, \ldots, n_1 + n_2$. The sufficient statistics for the model in (D.17) are $\left\{\sum_{i=1}^{n_1+n_2} x_{li} x_{vi}\right\}_{0 \leq l \leq v \leq k}$ and $\left\{\sum_{i=1}^{n_1+n_2} x_{li} y_i\right\}_{0 \leq l \leq k}$. Several of these sufficient statistics are determined once we select the sample sizes n_1 and n_2. The remaining sufficient statistics must be simulated to define the posterior of β_1 and compute relevant posterior probabilities. As mentioned in Appendix D.5.2, there is a one-to-one mapping between the sufficient

statistics listed earlier in this paragraph and the sample means in groups 1 and 2 for $x_2, \ldots, x_k, \varepsilon$ along with the sample covariance matrix for $x_2, x_3, \ldots, x_k, \varepsilon$.

This sample covariance matrix can be drawn according to a Wishart distribution (Wishart, 1928). We use the Bartlett decomposition (Bartlett, 1934) of a matrix $\boldsymbol{E}$ from a k-variate Wishart distribution with scale matrix $\mathbf{V}$ and n degrees of freedom: $\boldsymbol{E} = \boldsymbol{L}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{L}^T$, where $\boldsymbol{L}$ is the Cholesky factor of $\mathbf{V}$ and

$$\boldsymbol{A} = \begin{pmatrix} c_1 & 0 & 0 & \cdots & 0 \\ n_{21} & c_2 & 0 & \cdots & 0 \\ n_{31} & n_{32} & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{k1} & n_{k2} & n_{k3} & \cdots & c_{k1} \end{pmatrix} \tag{D.18}$$

such that $c_l^2 \sim \chi^2_{n-l+1}$ and $n_{lv} \sim \mathcal{N}(0,1)$ independently. The matrix $\boldsymbol{L}$ is determined by the selected normal distributions for $x_2, x_3, \ldots, x_k, \varepsilon$. If the explanatory covariates and error terms are mutually independent as required by standard linear regression assumptions, then $\boldsymbol{L}$ is diagonal.

Algorithm D.4 details our procedure to map posteriors of β_1 to $[0,1]^d$. We now elaborate on several components of Algorithm D.4. First, Algorithm D.4 requires that we characterize the data generation process by choosing values for the regression coefficients $\boldsymbol{\beta}$ and parameters for the normal distributions of $x_2, \ldots, x_k, \varepsilon$. These coefficients and normal parameters could either take the same values for each simulation repetition as in Chapter 3 or be drawn according to a design prior. We must also specify the parameters for the normal-inverse-gamma prior: $\boldsymbol{\mu}_0$, $\boldsymbol{\lambda}_0$, a_0, and b_0.

In Line 2 of Algorithm D.4, we assign qn observations to group 1 and n observations to group 2. This choice is made to align with the numerical studies in Appendix D.5.3; however, we emphasize that n observations were assigned to group 1 and qn observations were assigned to group 2 in the main portion of Chapter 5. This change predominantly involves notation and does not impact Chapter 5's theoretical results. Without loss of generality, the $n_1 + n_2$ observations are ordered so that the first $n_1 = qn$ observations are assigned to group 1. In Lines 4 to 8, we generate the group sample means for the $k-1$ additional covariates and the error term ε using CDF inversion with quantiles from the normal distributions selected as inputs for Algorithm D.4. We obtain the overall sample means for the combined data from both groups using algebra in Line 9.

In Lines 10 to 12, we use the Bartlett decomposition in (D.18) to generate the sample covariance matrix of $x_2, \ldots, x_k, \varepsilon$ for the combined data from both groups using CDF

Algorithm D.4 Mapping Posteriors to $[0,1]^d$ with Linear Regression

1: **procedure** MapLinear($\boldsymbol{\beta}$, $\{\mu_l\}_{l=2}^k$, $\{\sigma_l^2\}_{l=2}^k$, σ_ε^2, n, q, $\boldsymbol{u}$, $\boldsymbol{\mu}_0$, $\boldsymbol{\lambda}_0$, a_0, b_0)
2: Let $n_1 = qn$ and $n_2 = n$
3: Let $\sum_{i=1}^{n_1+n_2} x_{1i} = n_1$
4: **for** l in $2{:}k$ **do**
5: Let $n_1^{-1} \sum_{i=1}^{n_1} x_{li} = \mu_l + \Phi^{-1}(u_{2(l-2)+1})\sigma_l/\sqrt{n_1}$
6: Let $n_2^{-1} \sum_{i=n_1+1}^{n_2} x_{li} = \mu_l + \Phi^{-1}(u_{2(l-2)+2})\sigma_l/\sqrt{n_2}$
7: Let $n_1^{-1} \sum_{i=1}^{n_1} \varepsilon_{ji} = \Phi^{-1}(u_{2k-1})\sigma_\varepsilon/\sqrt{n_1}$
8: Let $n_2^{-1} \sum_{i=n_1}^{n_2} \varepsilon_{ji} = \Phi^{-1}(u_{2k})\sigma_\varepsilon/\sqrt{n_2}$
9: Calculate $\{\bar{x}_l\}_{l=2}^k$ using Lines 4 to 6 and $\bar{\varepsilon}$ using Lines 7 and 8
10: **for** l in $(2k+1){:}(k^2+5k)/2$ **do**
11: Going through each row of (D.18) from left to right, generate the $(l-2k)^{\text{th}}$ nonzero element of $\boldsymbol{A}$ using the u_l-quantile of the relevant normal or chi-squared distribution.
12: Let the sample covariance matrix for $x_2, x_3, \ldots, x_{k+1}, \varepsilon$ be $\boldsymbol{E} = \boldsymbol{L}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{L}^T$.
13: Obtain the sufficient statistics for the linear model using $\boldsymbol{E}$, Line 9, and algebra.
14: Use the sufficient statistics and prior hyperparameters $\boldsymbol{\mu}_0$, $\boldsymbol{\lambda}_0$, a_0, and b_0 to obtain the 3-parameter t-distribution posterior for β_1.

inversion with normal and chi-squared distributions. Given the simulated first moments of the covariates and error terms from the previous paragraph and the second central moments generated in Lines 10 to 12, we can calculate the raw second sample moments using algebra. The raw second sample moments $\left\{\sum_{i=1}^{n_1+n_2} x_{li}x_{vi}\right\}_{0\le l\le v\le k}$ and $\left\{\sum_{i=1}^{n_1+n_2} x_{li}y_i\right\}_{0\le l\le k}$ prompt the posterior of β_1, which we obtain in Line 14 of Algorithm D.4 using the specified prior hyperparameters. After explaining how this algorithm aligns with our theoretical results from Chapter 5 in the next subsection, we describe how Algorithm D.4 can be used with a slightly modified version of Algorithm 5.2 in Appendix D.5.3.

D.5.2 Connections to Theoretical Results from Chapter 5

We first discuss the connection between Algorithm D.4 and Theorem 5.1. In Appendix D.5.1, we assume that the additional $k-1$ covariates $x_2, \ldots, x_k$ are normally distributed. As such, the statistics simulated using CDF inversion in Algorithm D.4 follow the relevant normal and chi-squared distributions exactly. Algorithm D.4 also uses the exact posterior of β_1, which is a t-distribution, instead of a large-sample normal approximation thereof.

When using a pseudorandom sequence $\boldsymbol{U} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1]^d)$ with Algorithm D.4, the sampling distribution of posterior probabilities $\mathcal{P}^{\delta}_{n,q,\boldsymbol{U},Alg.D.4}$ is therefore the same as $\mathcal{P}^{\delta}_{n,q,(D.17)}$, the sampling distribution prompted by generating data $\boldsymbol{Y}^{(n,q)}$ and $\boldsymbol{X}^{(n,q)}$ according to the linear model in (D.17).

However, we would need to consider analogues to Theorem 5.1 for more general linear regression settings. For instance, if we used a non-conjugate prior for $\boldsymbol{\beta}$ and σ^2_ε, we would have approximated the posterior of β_1 using the Laplace approximation (Gelman et al., 2020). This approximation to the exact posterior of β_1 is suitable for large sample sizes n under the conditions for the BvM theorem in Appendix B.1.1. Moreover, we may consider linear regression models where not all $k-1$ additional covariates are normally distributed. The sufficient statistics for the linear model are based on sums of functions of independent observations regardless of the underlying distributions for $x_2, \ldots, x_k$.

For large enough n, the joint sampling distribution of these sufficient statistics is approximately normal. We could therefore generate approximate sufficient statistics using univariate conditional normal CDF inversion under a variety of distributional assumptions for the covariates. A hybrid of Algorithms 5.1, D.1, and D.4 could be developed for settings where the normal approximation to the sampling distribution of discrete covariates is poor for moderate n. Under the conditions for the BvM theorem, analogues to Algorithm 5.1 regarding the convergence of the sampling distribution of posterior probabilities should exist for more general linear regression models.

We now discuss the connection between Algorithm D.4 and Lemma 5.1. Part (a) of Lemma 5.1 involves a normal conduit for the data $\hat{\boldsymbol{\eta}}_n$. Unlike for the two-group comparisons facilitated via Algorithms 5.1, D.1, or D.2, several conduits for the data in Algorithm D.4 pertain to both groups. These data conduits $\hat{\boldsymbol{\eta}}_n$ include $\{\bar{x}_{l(1)}\}_{l=2}^{k}$ and $\{\bar{x}_{l(2)}\}_{l=2}^{k}$ generated in Lines 4 to 6, $\{\bar{\varepsilon}_{(j)}\}_{j=1}^{2}$ generated in Lines 7 and 8, and $\{\hat{s}_{x_l,x_v}\}_{2\leq l\leq v\leq d}$ and $\{\hat{s}_{x_l,\varepsilon}\}_{2\leq l\leq d}$ generated in Lines 10 and 11. For the sample means listed above, the subscript in parentheses denotes the group number $j = 1$ or 2.

Given the process in Algorithm D.4, we show that all components of $\hat{\boldsymbol{\eta}}_n$ can be generated using the process outlined in part (a) of Lemma 5.1. The only change in notation that we must make for this part of the lemma involves the simulation dimension, which is now $d = 0.5k(k+5)$ instead of $2d+1$ for some $d \in \mathbb{N}$. The data conduits $\{\bar{x}_{l(1)}\}_{l=2}^{k}$, $\{\bar{x}_{l(2)}\}_{l=2}^{k}$, and $\{\bar{\varepsilon}_{(j)}\}_{j=1}^{2}$ are normal sample means. These conduits therefore satisfy the conditions for part (a) of Lemma 5.1 based on the proof in Appendix B.2.

To explain why the remaining data conduits $\{\hat{s}_{x_l,x_v}\}_{2\leq l\leq v\leq d}$ and $\{\hat{s}_{x_l,\varepsilon}\}_{2\leq l\leq d}$ satisfy the conditions for part (a) of Lemma 5.1, we reconsider the Bartlett decomposition for

the Wishart distribution from (D.18): $\boldsymbol{E} = \boldsymbol{L}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{L}^T$. The matrix $\boldsymbol{A}\boldsymbol{A}^T$ is symmetric, so only the lower-triangular component of this matrix contains unique entries. Going through the lower-triangular component of $\boldsymbol{A}\boldsymbol{A}^T$ from top to bottom in each column, each successive component only depends on *one* new normal or chi-squared variable. Since $\boldsymbol{L}$ is also lower triangular, the top rightmost element of $\boldsymbol{E}$ only depends on c_1^2 from (D.18); $\boldsymbol{E}_{1,1}$ is related to the sample variance for x_2 in both groups. Given c_1^2, $\boldsymbol{E}_{1,2}$ (related to the sample covariance for x_2 and x_3 in both groups) only depends on n_{21} from (D.18). Similar results hold true for all unique components of $\boldsymbol{E}$, which is the unnormalized covariance matrix. The chi-squared variables used in (D.18) approximate normal variables as $n \to \infty$. In such settings, each component of $\{\hat{s}_{x_l,x_v}\}_{2\leq l\leq v\leq d}$ and $\{\hat{s}_{x_l,\varepsilon}\}_{2\leq l\leq d}$ generated using the iterative process in Algorithm D.4 depends only on the quantile of a single univariate normal distribution. Part (a) of Lemma 5.1 is therefore satisfied for all components of $\hat{\boldsymbol{\eta}}_n$ from Algorithm D.4 for sufficiently large n.

Part (b) of Lemma 5.1 involves an estimate for the characteristic of interest $\hat{\theta}_n$. Here, we let $\hat{\theta}_n = \hat{\beta}_{1,n}$ be the MLE for β_1 in (D.17) that corresponds to a sample size of n. Standard results for linear regression prompt the formula for the MLE of the regression coefficients $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, where $\boldsymbol{X}$ is the covariate matrix for (D.17) and $\boldsymbol{Y}$ is the vector of response variates. By considering the elements of the matrices $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{Y}$, we have that

$$\hat{\boldsymbol{\beta}}_1 = g\left(\left\{\sum_{i=1}^{n_1+n_2} x_{li}x_{vi}\right\}_{0\leq l\leq v\leq k}, \left\{\sum_{i=1}^{n_1+n_2} x_{li}y_i\right\}_{0\leq l\leq k}\right). \tag{D.19}$$

There is a one-to-one mapping between $\left\{\sum_{i=1}^{n_1+n_2} x_{li}x_{vi}\right\}_{0\leq l\leq v\leq k}$ and $\left\{\sum_{i=1}^{n_1+n_2} x_{li}y_i\right\}_{0\leq l\leq k}$ and the components of $\hat{\boldsymbol{\eta}}_n$: $\left\{\bar{x}_{l(1)}\right\}_{l=2}^{k}$, $\left\{\bar{x}_{l(2)}\right)_{l=2}^{k}$, $\left\{\bar{\varepsilon}_{(j)}\right\}_{j=1}^{2}$, $\{\hat{s}_{x_l,x_v}\}_{2\leq l\leq v\leq d}$, and $\{\hat{s}_{x_l,\varepsilon}\}_{2\leq l\leq d}$. From (D.19), it follows that

$$\hat{\beta}_1 = g_*\left(\left\{\bar{x}_{l(1)}\right\}_{l=2}^{k}, \left\{\bar{x}_{l(2)}\right)_{l=2}^{k}, \left\{\bar{\varepsilon}_{(j)}\right\}_{j=1}^{2}, \{\hat{s}_{x_l,x_v}\}_{2\leq l\leq v\leq d}, \{\hat{s}_{x_l,\varepsilon}\}_{2\leq l\leq d}\right)$$

The methods for linear regression presented in Algorithm D.4 satisfy the conditions for part (b) of Lemma 5.1 since $\hat{\beta}_1 = g_*(\hat{\boldsymbol{\eta}}_n)$ for some function $g_*(\cdot)$ that does not depend on n. The remainder of the proof for part (b) follows from Appendix B.2.

Part (c) of Lemma 5.1 also holds true for two reasons. First, the conditions for the BvM theorem are satisfied. Second, Algorithm D.4 uses the exact posterior of β_1, which is a t-distribution where the degrees of freedom increase as $n \to \infty$. Further discussion of this logic is provided in Appendix B.2. Lastly, the proof in Appendix D.2.2 can be used

to go from part (c) to part (d) of Lemma 5.1 without any modifications. Thus, the linear approximation to $\text{logit}(p_{n,q,\boldsymbol{u}_r}^{\delta_U-\delta_L})$ prompted by Algorithm D.4 as a function of n is a good global approximation for sufficiently large sample sizes.

D.5.3 Illustrative Example with Linear Regression

Here, we consider a two-group comparison with additional covariates that is based on a recent clinical trial funded by Novo Nordisk (Wilding et al., 2021). This clinical trial assessed the effectiveness of weekly semaglutide injections for the purpose of weight loss. In this clinical trial, patients in groups 1 and 2 were respectively given a weekly semaglutide injection or placebo for 68 weeks. One datum y_i that was collected for each patient $i = 1, \ldots, n_1 + n_2$ was their weight loss in kilograms (kg) over the course of the study. In total, $n_1 = 1306$ and $n_2 = 655$ patients were enrolled in this study. As detailed in Wilding et al. (2021), the two groups of patients were well balanced with respect to various additional covariates, and the patients who were given the semaglutide lost an average of 12.4 kg more than the patients who were given the placebo. This treatment was deemed statistically significant using p-values.

We now suppose that we want to design a two-group comparison for a phase I clinical trial of a similar semaglutide medication using information from Wilding et al. (2021). For illustration, we plan to use a Bayesian linear regression model with one additional covariate (i.e., $k = 2$). The regression model that we consider takes the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \tag{D.20}$$

where y is weight loss in kg, $x_1 = \mathbb{I}(\text{Group} = 1)$ is the binary treatment indicator, and x_2 is the patient's baseline waist circumference in centimetres (cm). Baseline waist circumference was one of several covariates measured by Wilding et al. (2021). We choose to include this covariate in our linear model because it is reasonable to expect that correlation between baseline waist circumference and weight loss is nonnegligible. As a continuous covariate, the baseline waist circumference for all patients could feasibly be normally distributed.

For this comparison, the characteristic of interest is $\theta = \beta_1$, the increased amount of weight loss (in kg) associated with taking the semiglutide injections. We choose the interval $(\delta_L, \delta_U) = (5, \infty)$ for illustration. This choice implies that we want to support the hypothesis $H_1 : \beta_1 > 5$. Because various side effects are typically experienced by patients who receive semaglutide treatments (Wilding et al., 2021), we want to observe a substantial weight loss of at least 5 kg associated with the semaglutide treatment to offset such side effects. Wilding et al. (2021) assigned patients to the treatment and control groups at a

2 to 1 ratio, and we follow this guidance for our hypothetical study. Using the notation from Algorithm D.4, we have that $q = 2$, $n_1 = 2n$, and $n_2 = n$.

We adopt a simpler process to define design priors for this example than the framework proposed in Section 5.4. The approach proposed here illustrates that our framework from Chapter 5 can be simplified and combined with elements of our approach in Chapter 3. For this regression example, we specify $p_{D_0}(\boldsymbol{\eta})$ as a degenerate prior. We must therefore choose design values for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ along with design values for the normal distributions of x_2 and ε. Here, we choose design values for the regression parameters of $\boldsymbol{\beta}^* = (-25.75, 5, 0.25)$. The design value $\beta_1^* = 5 = \delta_L$ is on the boundary of the hypotheses H_0 and H_1. The choice for $\beta_0^* = -25.75$ indicates that we expect patients in group 2 who are given the placebo to lose 3 kg on average. This assumption is reasonable since patients in both groups are given non-pharmaceutical interventions, such as counselling and diet plans. The design value $\beta_2^* = 0.25$ reflects a Pearson's correlation coefficient of roughly 0.3 between baseline waist circumference in cm and weight loss in kg.

Moreover, we suppose the baseline waist circumference x_2 for all patients follows a $\mathcal{N}(115, 14.5^2)$ distribution and the error terms ε follow a $\mathcal{N}(0, 10.07^2)$ distribution. The two groups in Wilding et al. (2021) were balanced with respect to various covariates, so we assume the distributions of x_2 and ε are the same in both groups. The choices defined in the previous two paragraphs define $p_{D_0}(\boldsymbol{\eta})$, which allows us to generate all conduits for the data $\hat{\boldsymbol{\eta}}_n$ under H_0. According to the notation from Section 5.4, the red region for this example is $\mathcal{R} = 5 = \delta_L$.

The design prior $p_{D_1}(\boldsymbol{\eta})$ for this example is a relatively simple nondegenerate prior. We use the (degenerate) design values from $p_{D_0}(\boldsymbol{\eta})$ for all data generation parameters except for β_1. In this case, β_1^* takes a value of 9 or 12 with probability 0.5. The choice illustrates an alternative way to incorporate uncertainty in the data generation process. Simpler methods that account for this uncertainty are useful when it is too difficult to specify a nondegenerate design prior as in Section 5.4, which may occur in settings where there are many additional covariates. The data generation process in $p_{D_1}(\boldsymbol{\eta})$ reflects a mixture of two scenarios: $\beta_1^* = 12$ reflects the previously demonstrated efficacy of semaglutide injections (Wilding et al., 2021), and $\beta_1^* = 9$ reflects a less optimistic scenario. As discussed shortly, how we define $p_{D_0}(\boldsymbol{\eta})$ and $p_{D_1}(\boldsymbol{\eta})$ for this example requires us to make slight modifications to Algorithm 5.2. The green region for this example is $\mathcal{G} = \{9, 12\}$ according to the notation from Section 5.4.

To illustrate the use of our methods with this regression example, an uninformative conjugate normal-inverse-gamma prior is used for $\boldsymbol{\beta}$ and σ_ε^2 with the following parameters: $\boldsymbol{\mu}_0 = (0, 0, 0)$, $\boldsymbol{\lambda}_0 = 0.01 \times \mathbb{I}_3$, $a_0 = 1$, and $b_0 = 1$ such that $\mathbb{I}_3$ is the 3×3 identity

matrix. For this example, we select the standard criteria for the operating characteristics of $\alpha = 0.05$ and $\beta = 0.2$. Unlike in the previous numerical studies for Chapter 5 and Appendix D, we use $m = 4096$ and $m_0 = 128$ for Algorithm 5.2. We use a smaller value of m than the previously used value of 8192 since we average over less variability in the $\boldsymbol{\eta}^*$ values prompted by these design priors $p_{D_0}(\boldsymbol{\eta})$ and $p_{D_1}(\boldsymbol{\eta})$. The value for m_0 is reduced to reflect this reduction in m.

The dimension of the hypercube is $d = 0.5k(k+5) = 7$ for this example. Unlike in our previous applications of Algorithm 5.2, we do not need to add an additional dimension to the hypercube to order our draws from the design priors $p_{D_0}(\boldsymbol{\eta})$ and $p_{D_1}(\boldsymbol{\eta})$. Since $p_{D_0}(\boldsymbol{\eta})$ is degenerate, all draws from this design prior will be identical and do not require reordering. The design prior $p_{D_1}(\boldsymbol{\eta})$, however, is nondegenerate. In this case, we can simply take $m/2$ draws where $\beta_1^* = 9$ and $m/2$ draws where $\beta_1^* = 12$. We obtain suitable results using Algorithm 5.2 if we order these draws to alternate between $\beta_1^* = 9$ and $\beta_1^* = 12$ since subsequences of the Sobol' sequence are also low discrepancy (Sobol', 1967). This fact can also be leveraged to order draws from more complicated design priors involving discrete mixtures.

When using Algorithm D.4 to map posteriors to $[0,1]^7$, Algorithm 5.2 returned an optimal design characterized by $(n, \gamma) = (40, 0.9554)$. For reference, we considered the precision of the (n, γ) recommendations with Sobol' and pseudorandom sequences for this regression example using the process to create Figure D.1. The results from that numerical study suggested the (n, γ) recommendations obtained using Sobol' sequences with length $m = 4096$ are roughly as precise as those obtained with pseudorandom sequences of length $m = 3.5 \times 10^4$.

As in Section 5.5.3, we repeated the sample size calculation from the previous paragraph 1000 times with different Sobol' sequences $\{\boldsymbol{u}_r^{(1)}\}_{r=1}^m$ and $\{\boldsymbol{u}_r^{(0)}\}_{r=1}^m$. For each repetition, the optimal design coincided with the (n, γ) recommendation obtained by always exploring entire sampling distributions of posterior probabilities using the same Sobol' sequences. For this regression example, Algorithm 5.2 took roughly 4 seconds on a standard laptop without parallelization to return an optimal design for the illustrative example. The modified version of Algorithm 5.2 that explored entire sampling distributions of posterior probabilities took approximately 14 seconds using Sobol' sequences with length $m = 4096$ and 118 seconds using pseudorandom sequnces with length $m = 3.5 \times 10^4$. Using the process described in Section 5.6, we averaged contour plots for the type I error rate and power corresponding to the 1000 repetitions of the sample size calculation for the regression example. These plots are given in the left column of Figure D.9. Based on these plots, the smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours is 40.

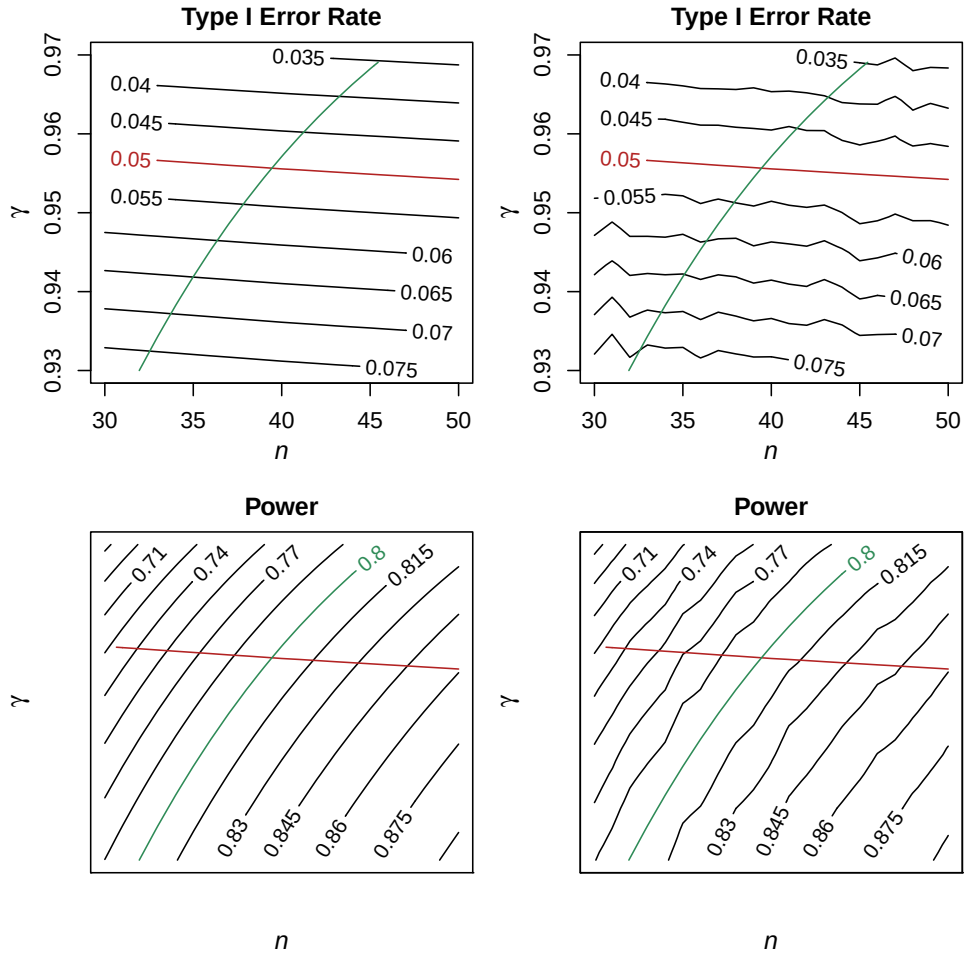


Figure D.9: Left: Averaged contour plots for the type I error rate and power from 1000 sample size calculations with Algorithm D.4 and the regression example. Right: Contour plots estimated by simulating data.

The contour plots in the right column of Figure D.9 were created by simulating $m = 10^5$ samples from the prior predictive distributions for $n = \{30, 31, ..., 50\}$ following the process detailed in Section 5.2. Again, the contours in the right column are more jagged since the sampling distributions of posterior probabilities in the right plots are estimated independently for each value of n considered. The smallest $n \in \mathbb{Z}^+$ to the right of the intersection of the green and red contours in the right plots is also $n = 40$. The plots in the left and right columns are similar, so using Algorithm D.4 to map posteriors to $[0, 1]^7$

for this regression example prompts suitable performance.

Appendix D.5 introduced a scalable design framework with two-group comparisons that account for additional covariates. These methods were developed for linear regression models. For certain generalized Bayesian linear models, sufficient statistics may be difficult or impossible to generate. Future work could combine the approaches in Appendices D.4 and D.5 to develop design methods with sampling distribution segments that accommodate more flexible and complex regression models.