Methods for Improving Efficiency of Planned Missing Data Designs

by

Paul M. Imbriano

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Trivellore E. Raghunathan, Chair
Professor Michael R. Elliott
Professor Roderick J. Little
Assistant Professor Ritesh Mistry
Research Associate Professor James Wagner

Paul M. Imbriano

pimbri@umich.edu

ORCID iD: 0000-0001-6216-5431

Acknowledgments

I would like to thank my advisor, Trivellore Raghunathan, for all of his help and guidance over the years on my thesis. Raghu, you have been a wonderful advisor and it has been a pleasure working with you. I would also like to thank the rest of my committee members, Michael Elliott, Roderick Little, Ritesh Mistry, and James Wagner for their helpful suggestions.

To all my friends at the University of Michigan and in Ann Arbor, thank you for all the fun moments and great memories. I would not have been able to get through graduate school without you. I will miss you all. I will not forget the nights of board games, dinners, and movies, and the Saturday mornings of basketball. I especially want to thank Jingchunzi Shi and my roommate, Alan Kwong.

Finally, I would like to thank my parents, Michael and Vicki Imbriano, and my sister, Laura, for all of their support over the years.

Contents

List of Figures

List of Tables

ABSTRACT

Any survey specifically constructed so that at least some variables are unobserved on a subset of participants is a planned missing data design, where missing data represent an intentional feature of the study. Use of planned missing data designs can potentially reduce costs, improve data quality, and reduce unplanned missing data, and advancements in missing data methodology and multiple imputation software make planned missing data designs more attractive than before. Two commonly used planned missing data designs are two-phase sampling and split questionnaire design. Two-phase sampling is used to improve the efficiency of an estimate for a single outcome that is costly to measure, while the split questionnaire design is primarily used to reduce survey length.

First, we propose new methods for selecting our second phase sample in two-phase surveys to reduce the variance of our estimate. When our outcome variable is continuous, we can use the data collected in Phase I for selecting our Phase II sample in order to increase the precision of the estimates. For other instances, we propose an adaptive sampling method to select Phase II samples in order to improve estimation of the quantity of interest.

Next, we examine the performance of several design allocations for implementing a split questionnaire survey in a longitudinal study. While many papers examined the administration of split questionnaire designs in cross-sectional studies, research in applying these methods to longitudinal studies has been limited. For our project, we focus on the commonly used 3-form design and propose several methods to administer the 3-form split questionnaire in a longitudinal study. Using simulations and data from the Health and Retirement Study, we compare the performance of each proposed design under several correlation structures.

Finally, we propose a method for improving variable allocation in split questionnaire designs. We establish a criterion that allows us to determine which variable allocations minimize the loss of information due to missing data. We use the Kullback-Leibler divergence between the posterior distribution of the parameters with missing data and the posterior distribution without missing data to locate optimal designs. After establishing a criterion for comparing designs, we propose a search algorithm to find optimal variable allocations, as it would be difficult to enumerate all possible designs as the number of variables grows.

Chapter 1

Introduction

In most instances, missing data are treated as a nuisance. In the presence of missing data, we usually require additional methods for statistical analysis that are unnecessary when no values are missing, often making analysis more difficult and time consuming. In addition, missing values result in a loss of statistical power due to a reduction in the effective sample size, and can potentially bias results from analysis. For that reason, most studies do their best to try to avoid missing data. Interestingly, advancements in missing data methodology and software for handling missing data, especially the multiple imputation approach, have made missing data less problematic for analysis. Some have advocated that designing a survey to purposely include missing data could improve that study (Graham et al., 1996, 2006; Littvay, 2009). Any study purposely constructed to include missing values is a planned missing data design. Planned missing data designs are typically used to either reduce study costs or improve the quality of collected data by reducing the burden placed on participants. It is also important to note that researchers directly control missing data resulting from the study design, which means planned missing data will not create any systematic bias in results for most methods of statistical analysis. Due to this, missing values resulting directly from these designs will either be Missing Completely at Random (MCAR) or Missing at Random (MAR), which allows us to easily use multiple imputation, maximum likelihood, or Bayesian approaches to analyze our data. Graham et al. (2006) describe two popular types of planned missing data designs, which they refer to as the 2–method measurement design and the 3–form design.

The 2–method measurement design described by Graham et al. (2006) comes from the two-phase sampling method proposed by Neyman (1938). It is used to either decrease the variance of an estimate compared to a traditional study design or decrease the cost of a study. Two-phase survey sampling was originally developed for measuring the population mean of a costly outcome to measure. Neyman suggested breaking the sampling into two

parts or phases. In Phase I, we measure a cheaper variable correlated with the outcome on a large sample of the population. In Phase II, we use that surrogate variable to stratify the initial sample and then measure our outcome on a subsample. Due to subsampling, the more expensive outcome is missing on a large proportion of study participants by design. In general, using the information from the Phase I variable or variables to aid in estimation of the outcome variable produces a smaller variance than using the Phase II outcome by itself. For instances where both variables are highly correlated and measuring the surrogate variable is much cheaper than the desired outcome, this design provides a more efficient estimate for the mean of our outcome variable than a traditional survey design that costs the same (McNamee, 2003; Graham et al., 2006).

Most literature on the topic of two-phase sampling focuses on using this method to estimate disease prevalence (Alonzo, 2003; McNamee, 2003, 2004; Gao et al., 2000; Shrout & Newman, 1989). Typically, an inexpensive screening test is conducted in Phase I. In Phase II, a costly but more accurate gold standard disease diagnostic is administered after stratifying by the initial screening test. Two-phase sampling can also be used for case-control studies (Breslow & Cain, 1988; Cain & Breslow, 1988). Pierce & Burgess (2013) examined the use of two-phase sampling in the context of instrumental variables and Mendelian randomization, where either an exposure or outcome is only available for a subsample of study participants. In their paper, they discuss two-phase sampling as a potential method for estimating the effect of an exposure on an outcome by incorporating data from previous genetic association studies. For continuous outcomes, such as biomarkers, the population mean can be estimated either through stratification on Phase I data or with a regression estimator, which uses both linear regression estimates and the sample mean of our outcome to estimate the population mean (Sitter, 1977). Linear regression is used to adjust for differences between the Phase I and Phase II samples. This estimator will provide a more precise estimate of the population mean, compared to using the Phase II sample mean on its own, when the Phase I variable or variables are correlated with our outcome in Phase II.

Both the Great Smokey Mountain study (Costello et al., 1996) and the paper by Patton et al. (2000) use two-phase sampling for estimating the prevalence of psychiatric disorders in a population. These studies administered a screening test questionnaire to a random sample of the population of interest. Participants failing the screening test, along with a random sample of those passing, were recruited into Phase II and underwent further diagnostics for psychiatric disorders.

In addition to the studies explicitly using the design proposed by Neyman, several large-

scale studies utilize sub-sampling similar to two-phase sampling in order to save costs. One such study, the Sacramento Area Latino Study on Aging (Mayeda et al., 2013), explored numerous health related outcomes in older Mexican Americans, including disease and cognitive impairment. Annual questionnaire and lab test variables were collected on all participants, and all subjects underwent a multistage screening process to diagnose dementia and cognitive impairment (Haan et al., 2003). Participants failing initial screening tests were subject to further examination, while only a subset of individuals passing these tests received further evaluation, resulting in a subset of individuals having an observed dementia diagnosis. This process was very similar to what is done in a typical two-phase study; however, traditional two-phase studies would focus on measuring a single outcome and not include dozens of other important outcomes of interest. Another study, the Health and Retirement study, also incorporated sub-sampling, where only a subset of participants were selected for the "enhanced interview,"(Crimmins et al., 2008a,b). Only for this subset were measurements taken of biomarkers and physical characteristics, similar to a two-phase study. Additionally, some studies may take a subsample from a larger study to measure a specific outcome. Thus, some of the methods used in two-phase study designs can be applied to a wide range of studies using sub-sampling that are not explicitly conducted as a two-phase study.

The other planned missing design described by Graham et al. (2006), the 3–form design, is an example of a split questionnaire design, which reduces the number of questions that any individual participant responds to without reducing the total number of questions in the study (Raghunathan & Grizzle, 1995). This is done by dividing the survey questionnaire into multiple components and having each participant receive a fraction of the components. For the 3–form design, the survey questionnaire is divided into four components (X,A,B,C). Each participant receives component X and two of the three other components. This results in a total of three survey forms, (X,A,B), (X,A,C), and (X,B,C), which are usually given in equal proportions to participants (Graham et al., 2006; Rhemtulla & Little, 2012). The 3–form design achieves an approximate 25% reduction in survey length. We could use a different number of components and splits if we desire further reduction in survey length. The 3–form design has been used in multiple studies (Graham et al., 2006).

A split questionnaire design is used for the purpose of lowering the burden and fatigue placed on study participants by reducing the length of the survey. This is useful because longer surveys have been shown to have higher nonresponse rates (Adams & Darwin, 1982; Dillman et al., 1993; Roszkowski & Bean, 1990; Beckett et al., 2016), and item nonresponse tends to occur more towards the end of a questionnaire (Raghunathan & Grizzle, 1995).

Participants are more likely to lose interest in a longer study, which affects the quality and accuracy of responses (Herzog & Bachman, 1981; Gonzalez & Eltinge, 2007; Peytchev & Peytcheva, 2017). Since a split questionnaire design reduces the burden of a survey, it makes item nonresponse less likely to occur, making us less likely to encounter missing not at random (MNAR) data (Rhemtulla & Little, 2012; Jorgensen et al., 2014; Kaplan & Su, 2016). As a result, missing data are less likely to introduce bias in our analysis.

The split questionnaire design provides a simple way to reduce the survey length without forcing researchers to omit any variables of interest. This can be beneficial for large national surveys that examine many outcomes related to the population of interest, such as the National Health and Nutrition Examination Survey (NHANES). Graham et al. (2006) demonstrated how using a split questionnaire design allows for the testing of many more hypotheses than simply using a standard questionnaire form with the same length. In practice, the split questionnaire design was found to produce estimates similar to those obtained using a standard questionnaire containing all variables of interest, though it causes a decrease in power due to the lower number of observations per variable (Raghunathan & Grizzle, 1995; Littvay, 2009). The loss of power can be mitigated by using an increased sample size, which may be easy to obtain when the split questionnaire design decreases the cost per participant (Littvay, 2009).

The split questionnaire design originates from multiple matrix sampling, where participants were assigned a random sample of available items within a study (Shoemaker, 1973; Raghunathan & Grizzle, 1995). Multiple matrix sampling places no constraint on the items assigned, making it possible that some items never appear together during the study. Split questionnaire design explicitly assigns items beforehand so that all two-way associations can be estimated. Multiple matrix sampling has frequently been used in education assessment (Shoemaker & Shoemaker, 1981). Students are evaluated on several subjects, but a full evaluation on an individual student would be extremely disruptive. Since investigators are interested in evaluating the student population instead of individuals, they can implement multiple matrix sampling to lower the burden on students. This also makes the study more likely to be approved by administrators. For multiple matrix sampling, each student receives both a set of common items administered to all participants and a randomly selected subset of items. This type of design has been used in the Kentucky Instructional Results Information System, the Massachusetts Comprehensive Assessment System, the National Assessment of Educational Progress, and the Dutch National Assessment Program (Childs & Jaciw, 2003).

In this paper, we further explore both two-phase sampling and split questionnaire design and propose ways of implementing these designs to further improve analysis. Chapter 2 focuses entirely on two-phase sampling, where we explore several methods for selecting our subsample with the goal of further lowering the standard errors of our estimates. We present several different methods for selecting the Phase II sample, depending on the distribution of our variables, and demonstrate the gain in efficiency from using these methods. We propose a novel method for choosing our Phase II subsample when our outcome variable is continuous and we have multiple Phase I variables. In these instances, we can select our subsample using the distribution of our Phase I variables to lower the variance of our estimate for the mean of our Phase II variable or the parameter estimates from the regression of our Phase II variable on the Phase I variables. For other cases, we derive estimates for the mean of our Phase II outcome and the relationship between our outcome and the Phase I variables. We then determine the sample size allocation that minimizes the variance of these estimates. Since the variance often depends on parameters from the conditional distribution of our Phase II variable given the Phase I variables, which are likely unknown beforehand, we propose an adaptive approach for selecting our subsample in order to improve efficiency.

Chapters 3 and 4 focus on implementation of split questionnaire designs. In chapter 3, we investigate several designs for administering a split questionnaire design in a longitudinal or panel study. There has not been very much work done to explore longitudinal split questionnaire designs (Jorgensen et al., 2014). Due to the repeated measures component of these studies, implementation of a split questionnaire differs for longitudinal studies compared to cross–sectional studies. In the chapter, we keep the split questionnaire forms the same at each visit and examine how rotation of those forms affect estimation. We propose six different methods for rotating the forms. Our goal is to determine under what circumstances we would prefer one type of form rotation over another. We then use simulations and data from the Health and Retirement Study to examine the performance of each of our proposed longitudinal designs under a number of different data structures.

In chapter 4, we propose a method for creating split questionnaire designs based on minimizing the loss of information due to missing data. We want to determine an optimal method to place variables into different components of a split questionnaire so that the inference obtained for the data we observe is as close as possible to the inference we would have obtained from using a tradition questionnaire. Since prior information on the joint distribution of our variables is necessary for determining an optimal design, we propose using a Bayesian framework. This allows us to incorporate our prior knowledge. Within

this framework, we use the Kullback-Leibler divergence between our observed data posterior distribution and the posterior distribution with no missing data to determine which design minimizes the loss of information. We also propose a search algorithm for locating good split questionnaire designs, as it is almost impossible to perform an exhaustive search of all designs. Finally, we discuss the limitations of our methods and future directions to explore for two-phase sampling and split questionnaire design in chapter 5.

Chapter 2

Sample Selection to Improve Estimation Efficiency in Two-phase Studies

2.1 Introduction

Two-phase survey sampling or double sampling was first proposed by Neyman (1938) for measuring the population mean of a variable, Y, which is costly to measure. Neyman suggested that if there were a variable, X, correlated with the outcome of interest, then first the variable X should be measured on a large sample, and, stratifying by X, then measure Y on a subsample. For instances where X is highly correlated with Y and the cost of measuring X is small compared to Y, this design provides a more precise estimate for the population mean of Y than a conventional survey design with the same cost (McNamee, 2003). Since Neyman's proposal, much of the two-phase sampling literature has focused on estimation of disease prevalence, where variables collected in Phase I are used as an inexpensive screening test and the more costly disease diagnostic is measured in Phase II (Alonzo, 2003; McNamee, 2003, 2004; Gao et al., 2000; Shrout & Newman, 1989); however, two-phase studies are not limited to these designs. Breslow and Cain describe design and analysis of two-phase case-control studies using logistic regression (Breslow & Cain, 1988; Cain & Breslow, 1988). Pierce & Burgess (2013) consider using two-phase sampling to lower the costs of observational studies using Mendelian randomization by only measuring either the exposure or outcome on subsample of individuals. Two-phase studies can also be used for continuous outcomes, such as biomarkers. When Y is continuous, the population mean can be estimated either through stratification on Phase I data or with the regression estimator, which uses the conditional distribution of Y given X from linear regression to more precisely estimate the mean of Y(Sitter, 1977). Examples of studies implementing the two-phase design for disease diagnosis include the Great Smokey Mountain study (Costello et al., 1996) and the paper by

Patton et al. (2000), both of which employed two-phase sampling to estimate the prevalence of psychiatric disorders.

In addition to studies using two-phase sampling as proposed by Neyman, some large-scale studies may employ sub-sampling and other similar designs due to cost limitations. These large-scale studies with sub-sampling differ from traditional two-phase sampling as there are typically multiple variables of interest to investigators. An example of this can be seen in the Sacramento Area Latino Study on Aging, which examined many diseases in an older Mexican American population (Mayeda et al., 2013). The study conducted annual questionnaires and lab tests on all participants, but, due to costs, used a screening process to diagnose dementia and cognitive impairment, resulting in an observed diagnosis on only a subset of participants (Haan et al., 2003). The Health and Retirement study also only selected a subset of participants to measure biomarkers and physical characteristics (Crimmins et al., 2008a,b). Thus, two-phase sampling can be an effective way to reduce survey costs even for studies that are not explicitly interested in a single outcome measurement. Any large survey that performs lab tests, genotyping, or anthropometric measurements on a subset of participants could be considered a two-phase design. The methods presented in this chapter can be applied to a wide range of studies using sub-sampling.

Throughout this chapter, we will let X denote the variable or variables obtained in Phase I and let Y denote the variable collected in Phase II. Most literature on two-phase designs focuses on estimation of the population mean for Y; however, for some studies, investigators may be more interested in the relationship between Y and X, i.e. regression analysis. In this chapter, we discuss new methods for selecting our Phase II sample based on the estimand of interest. Our selection methods assume Phase I variables are available for everyone prior to Phase II. In section 2.2, we detail how to choose our sample when we have multiple X variables and Y is continuous. In section 2.3, we describe estimation and sample selection for a single binary X and Y when estimating the mean of Y, risk difference, and natural logarithm of the odds ratio. In section 2.4, we extend the methods used in section 2.3 to when we have an additional categorical variable measured in Phase I. In section 2.5, we extend our method to the case of a binary X variable and normally distributed Y variable. Finally, we apply our methods to data from NHANES in section 2.6 and discuss our conclusions in section 2.7.

2.2 Sample Selection for Continuous Y

In this section, we discuss how to improve estimation for the mean of Y and how to improve parameter estimates from the regression of Y and X when Y is continuous.

2.2.1 Estimation of the Mean of Y

Let X = $(X_1, ...X_p)$ denote the p variables collected in Phase I of the study and let Y denote the variable collected on the Phase II subsample. Let n and r denote the sample size for Phase I and Phase II, respectively. When Y is continuous, we can estimate the mean of Y using the regression estimator (Sitter, 1977)

$$\widehat{\mu}_Y = \overline{Y} + (\overline{X}_n - \overline{X}_r)^T \widehat{\beta^*}, \tag{2.1}$$

where $\widehat{\beta^*}$ is a column vector containing the OLS estimates for $\beta_1, ..., \beta_p$ in the regression of Y on X, $\overline{Y}$ is the sample mean for Y, $\overline{X}_n$ is a column vector of the sample means for X among all study participants, and $\overline{X}_r$ is a column vector of the sample means for X among participants in Phase II. Note that X does not contain a column of 1s and β^* does not contain an intercept term. If X is correlated with Y, $\widehat{\mu}_Y$ is a more efficient estimate than $\overline{Y}$. If the Phase II sample is a simple random subsample of Phase I, $\widehat{\mu}_Y$ is an unbiased estimate for μ_Y even if the regression model is misspecified.

Suppose $Y_i = \beta_0 + X_i^T \beta^* + \epsilon_i$, $E(\epsilon_i|X_i) = 0$, $V(\epsilon_i|X_i) = \sigma^2$, and the ϵ_i are independent, where β_0 is the intercept term from linear regression. We can calculate the variance of $\widehat{\mu}_Y$ under simple random sampling using the law of total expectation, $V(\widehat{\mu}_Y) = E\{V(\widehat{\mu}_Y|X)\} + V\{E(\widehat{\mu}_Y|X)\}$.

First note that the regression estimator can be written as

$$\widehat{\mu}_Y = \widehat{\beta_0} + \overline{X}_n^T \widehat{\beta^*},$$

with $\widehat{\beta_0}$ being the OLS estimate for the intercept in the regression of Y on X. Then, we have

$$\begin{aligned} E\{V(\widehat{\mu}_Y|X)\} &= E\{V(\overline{Y}) + (\overline{X}_n - \overline{X}_r)^T V(\widehat{\beta^*})(\overline{X}_n - \overline{X}_r)\} \\ &= \frac{\sigma^2}{r} + \frac{\sigma^2}{r-1} E\{(\overline{X}_n - \overline{X}_r)^T (S_r^2)^{-1} (\overline{X}_n - \overline{X}_r)\}, \end{aligned}$$

where S_r^2 is the sample variance of X for individuals in Phase II of the study.

Looking at the last term,

$$\frac{\sigma^2}{r-1}(\overline{X}_n - \overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_n - \overline{X}_r)$$
$$= \frac{\sigma^2(n-r)}{nr(r-1)}\frac{(n-r)r}{n}(\overline{X}_{n-r} - \overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_{n-r} - \overline{X}_r),$$

where $\overline{X}_{n-r}$ is the sample average of X for individuals in Phase I, but not Phase II. When X follows a multivariate normal distribution, we have

$$\frac{(n-r)r}{n}(\overline{X}_{n-r} - \overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_{n-r} - \overline{X}_r) \sim T^2_{p,r-1},$$

where $T^2_{p,r-1}$ denotes a Hotelling's T-squared distribution with p and $r-1$ degrees of freedom. The above distribution still holds approximately for large samples even when X does not follow a multivariate normal. Taking the expectation of the Hotelling's T-squared distribution, we get

$$E(V(\widehat{\mu}_Y|X)) = \frac{\sigma^2}{r} + \frac{\sigma^2(n-r)p}{nr(r-p-2)}.$$

For small sample sizes and when X is not multivariate normal, we can approximate this expectation using an empirical distribution. We can calculate this distribution by repeatedly choosing random samples of size r and computing the quantity $(\overline{X}_n - \overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_n - \overline{X}_r)$. We would then take the average value to approximate the expectation.

$$\begin{aligned} V\{E(\widehat{\mu}_Y|X)\} &= V\{E(\overline{X}_n^T\widehat{\beta}^* + \widehat{\beta}_0)\} \\ &= V(\overline{X}_r^T\beta^* + \beta_0) \\ &= \frac{1}{n}\beta^{*T}\Sigma\beta^* \\ &= \frac{\sigma^2 R^2}{(1-R^2)n}, \end{aligned}$$

where Σ is the variance-covariance matrix for X and R^2 is the coefficient of determination from the OLS regression of Y on $X_1, ..., X_p$.
Therefore,

$$V(\widehat{\mu}_Y) = \frac{\sigma^2}{r} + \frac{\sigma^2 R^2}{(1-R^2)n} + \frac{\sigma^2(n-r)p}{nr(r-p-2)} \tag{2.2}$$

for large samples or when X is multivariate normal.

Note that the second term and the last term in the variance tend towards zero as the Phase I and Phase II sample size increase, respectively. With all else constant, increasing the number of X covariates that Y is regressed on will increase the variance of $\widehat{\mu}_Y$, unless those additional covariates reduce σ^2 enough to offset the increase in p. The contribution of the last term to the variance can be considerable if the Phase II sample size is not substantially larger than the number of covariates.

Note that $\sigma^2 = \sigma_Y^2(1 - R^2)$, where σ_Y^2 is the unconditional variance of Y. Plugging this quantity into (2.2), the variance of the regression estimator can be written as

$$\frac{\sigma_Y^2}{r}\left\{1 - R^2 + \frac{R^2 r}{n} + \frac{(1-R^2)(n-r)p}{n(r-p-2)}\right\}.$$

If R^2 is small, the variance of the regression estimator for μ_Y is greater than σ_Y^2/r, the variance of $\overline{Y}$. The regression estimator should be more efficient for a moderately large R^2, provided we have not used up too many degrees of freedom by conditioning on a large number of covariates. Table 2.1 shows the relative efficiency of the two-phase regression estimator compared to $\overline{Y}$ (the ratio of the variances of $\widehat{\mu}_Y$ and $\overline{Y}$) at different values of R^2 and p. A value greater than 1 indicates that $\overline{Y}$ is more efficient. From the table, we can see that the regression estimator is always less efficient when R^2 equals zero, but the regression estimator is far more efficient than using $\overline{Y}$ on its own for large values of R^2. Additionally, the regression estimator becomes less efficient if we increase the number of covariates (p) without changing R^2.

Table 2.1: Variance of the two-phase regression estimator ($\widehat{\mu}_Y$) divided by the variance of $\overline{Y}$ at different values of R^2 and p, when $n = 500$ and $r = 50$.

p	$R^2 = 0.00$	$R^2 = 0.25$	$R^2 = 0.50$	$R^2 = 0.75$
1	1.02	0.79	0.56	0.33
5	1.10	0.85	0.60	0.35
10	1.24	0.95	0.67	0.38
20	1.64	1.25	0.87	0.49

The last term in (2.2) comes from the expectation

$$\frac{\sigma^2}{r-1} E\left\{(\overline{X}_n - \overline{X}_r)^T (S_r^2)^{-1} (\overline{X}_n - \overline{X}_r)\right\}$$

under simple random sampling. If we could select our Phase II sample so that $\overline{X}_n = \overline{X}_r$, the

last term in the variance equation disappears, reducing the variance. Getting $\overline{X}_n = \overline{X}_r$ is difficult when there is a large number of covariates for which we would need to match Phase I and Phase II means. Instead, we use the distribution of the last term to select our sample. The last term is proportional to

$$\frac{r-p}{p(r-1)}\frac{(n-r)r}{n}(\overline{X}_{n-r}-\overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_{n-r}-\overline{X}_r),$$

which is distributed as $F_{p,r-p}$ exactly when X is multivariate normal and approximately for large samples. In these instances, we can use the quantiles of the F-distribution to choose an upper bound for our last term and only select Phase II samples satisfying this bound. We propose first choosing a subsample for Phase II using simple random sampling, and then computing the F-value using the available data from Phase I. If our F-value is greater than the selected upper bound, we reject that sample and randomly select another sample. We repeat this process until we obtain a F-value below our targeted upper bound. This method is similar to the approach for balancing covariates in re-randomization described by Morgan & Rubin (2012). The use of quantiles assures us that we can select a feasible upper bound, and enables us to calculate, on average, how many samples we would need to reject before finding a suitable sample.

In order to find the variance of $\widehat{\mu}_Y$ under this sampling scheme, we need to calculate the expectation of a truncated F-distribution, which can be approximated through simulation. A faster approach, however, is to use a χ^2 approximation to the F-distribution. If $C \sim F_{v,w}$, then vC converges to χ^2_v as w increases (Li & Martin, 2002). Taking $q = vC$, $q \sim \frac{w}{w-2}\chi^2_v$ provides a better approximation to the quantiles and expectation of an $F_{v,w}$ distribution. From this, we can approximate the expected value of C, when $C < b$, by taking $\frac{w}{w-2}$ times the expectation of a χ^2_v distribution truncated at $b^* = bv$. We have

$$E(q|q<b^*) = \frac{w}{w-2}\frac{vP(b^*,v+2)}{P(b^*,v)},$$

where $P(b^*, v)$ is the cdf for the χ^2_v distribution evaluated at b^*.

$$\text{Let} \quad \Delta(b,v) = \frac{P(bv,v+2)}{P(bv,v)}, \text{ then} \quad E(C|C<b) \approx \Delta(b,v)E(C).$$

Thus for any selected upper bound, b, we have

$$V(\widehat{\mu}_{Y_b}) \approx \frac{\sigma^2}{r} + \frac{\sigma^2 R^2}{(1-R^2)n} + \Delta(b,p)\frac{\sigma^2(n-r)p}{nr(r-p-2)}, \tag{2.3}$$

with $\widehat{\mu}_{Y_b}$ denoting the estimate for Y under bounded sampling. Note that $0 < \Delta(b,p) < 1$, and we can think of $\{1 - \Delta(b,p)\} \times 100$ as the percent reduction in the last term of the variance. Under our restricted sampling method, $\widehat{\mu}_{Y_b}$ is still an unbiased estimate for the mean of Y since $E(\overline{X}_n) = E(\overline{X}_r)$.

Figure 2.1 displays how well our proposed scaled χ^2 distribution approximates the F-distributed quantity from the regression estimator using various values of n, r, and p. From the figure, the χ^2 approximation appears to work reasonably well across many different sample sizes and number of covariates. Data were generated from a multivariate normal distribution.

A major advantage to using a χ^2_v approximation for estimating the variance is that it allows us to work backwards to compute the upper bound needed to achieve a specified reduction in variance, if possible, over using simple random sampling. There is a limit to the amount this method can decrease the variance, however, with the maximum decrease occurring as $\Delta(b,p)$ approaches zero. For any percent decrease less than this maximum, we can algebraically solve for the $\Delta(b,p)$ needed to achieve this decrease and then find our approximate bound from the ratio of the cdfs of the χ^2 distribution. Similarly, when using simple random sampling, we can calculate the Phase II sample size required to match the variance of our estimate when selection is based on an upper bound. We can work backwards to find the bound needed to achieve a specific increase in sample size over simple random sampling, provided the specified sample size increase is less than the maximum possible sample size increase. We will go into more details on how these computations work in the following subsections.

Figure 2.2 displays the approximate percent reduction in variance from bounded sampling with a fixed $\Delta(b,p)$ of 0.15. We are using equation (2.3) to approximate variance. From the figure, we can see how this sampling method reduces the variance of the regression estimator for a different number of covariates and varying Phase II sample sizes. This method has its greatest effect when the Phase II sample size is small relative to the number of X covariates. We can only achieve a modest reduction in variance for a single covariate unless the Phase II sample size is very small, while for twenty covariates, we can achieve notable variance reductions for even moderate sample sizes. This is because the use of our bound reduces the size of the last term in our variance, which is related to the number of covariates that we are conditioning on. When using the bounded sampling method, we are no longer as heavily penalized for adding additional covariates to our regression estimator. The difference in efficiency at the same value of R^2 between using 1 versus 20 covariates is nowhere near as

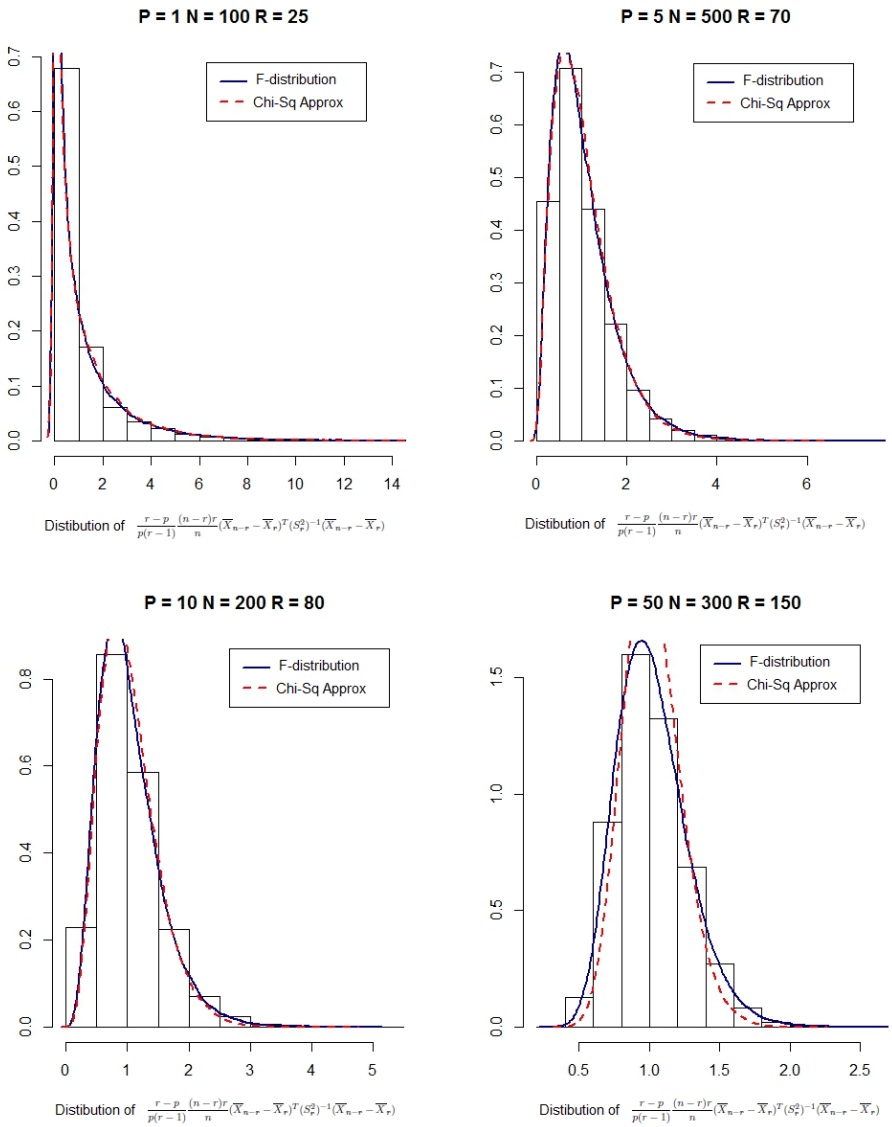

Figure 2.1: Distribution of $\frac{r-p}{p(r-1)}\frac{(n-r)r}{n}(\overline{X}_{n-r}-\overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_{n-r}-\overline{X}_r)$ from simulations at various values of n, r, and p. Kernel density plots from the F-distribution and our χ^2 approximation are overlayed.

pronounced as what we saw in table 2.1 when we were using simple random sampling to select the Phase II sample.

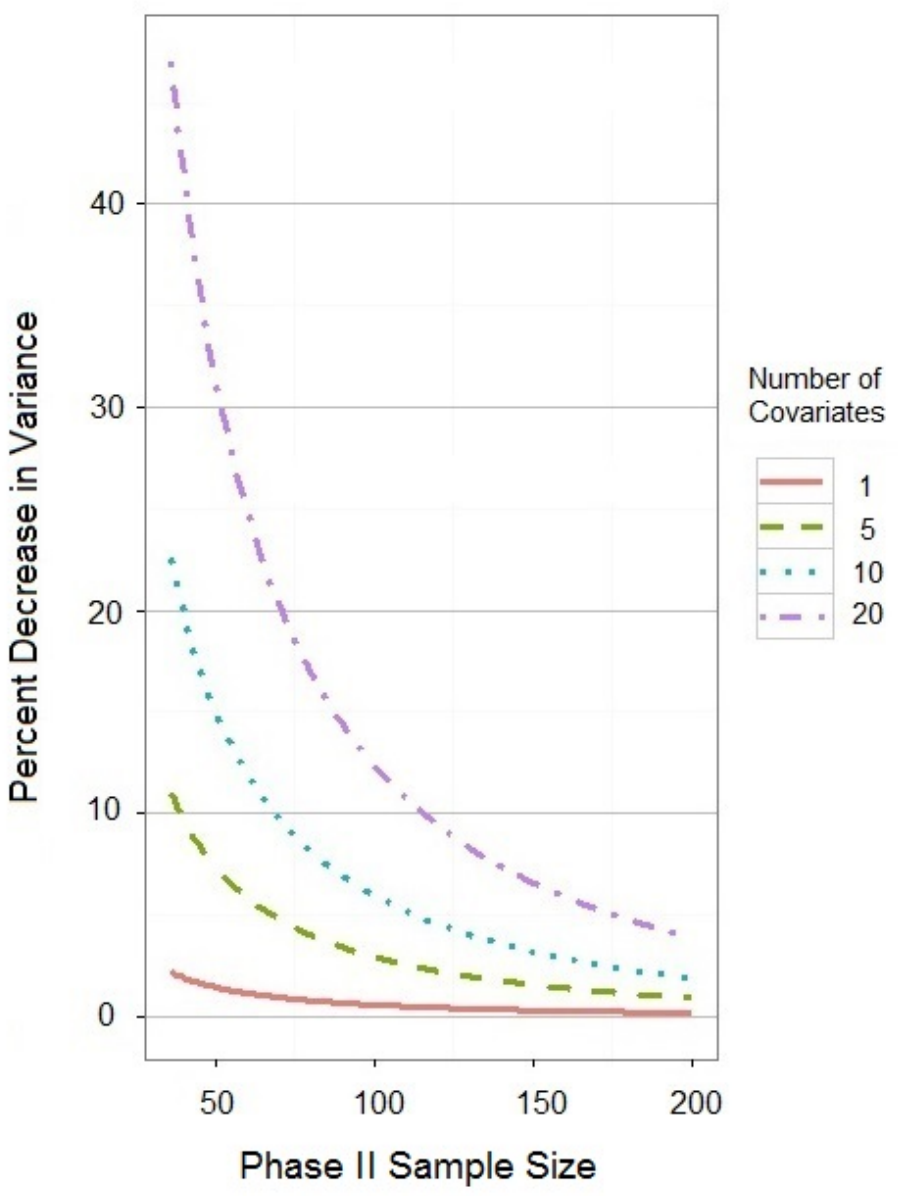

Figure 2.2: Decrease in variance from using bounded sampling over simple random sampling as a function of the Phase II sample size and the number of X covariates, with a fixed Phase I sample size of 500, $\Delta(b, p)$ of 0.15, and R^2 of 0.5.

For small sample sizes where X is not multivariate normal, we would instead use the empirical distribution of $(\overline{X}_n - \overline{X}_r)^T (S_r^2)^{-1} (\overline{X}_n - \overline{X}_r)$. We can approximate the quantiles by repeatedly selecting samples of size r and computing that quantity. We could use those quantiles to select an upper bound. We would then approximate the expectation under bounded

sampling by taking the average value of our computed quantity when we are restricted to samples greater than the selected bound.

Computing Variance Reduction from Bounded Sampling

Here we examine the approximate variance reduction from using bounded sampling for large sample sizes. Taking the ratio of $V(\widehat{\mu}_{Y_b})$ and $V(\widehat{\mu}_Y)$ we have

$$\frac{V(\widehat{\mu}_{Y_b})}{V(\widehat{\mu}_Y)} \approx \frac{1 + \frac{R^2 r}{(1-R^2)n} + \Delta(b,p)\frac{(n-r)p}{n(r-p-2)}}{1 + \frac{R^2 r}{(1-R^2)n} + \frac{(n-r)p}{n(r-p-2)}},$$

and

$$1 - \frac{V(\widehat{\mu}_{Y_b})}{V(\widehat{\mu}_Y)} = \frac{\{1 - \Delta(b,p)\}\frac{(n-r)p}{n(r-p-2)}}{1 + \frac{R^2 r}{(1-R^2)n} + \frac{(n-r)p}{n(r-p-2)}},$$

which only depends on R^2, n, r, p, and b and not on σ^2.

As the Phase I sample gets very large compared to the Phase II sample size, as $n \to \infty$,

$$1 - \frac{V(\widehat{\mu}_{Y_b})}{V(\widehat{\mu}_Y)} \approx \frac{\{1 - \Delta(b,p)\}\frac{p}{(r-p-2)}}{1 + \frac{p}{(r-p-2)}},$$

and the variance reduction no longer depends on R^2.

The maximum variance reduction achievable with this method occurs when $\Delta(b,p) = 0$. For very large n, this upper bound is approximately

$$\frac{\frac{p}{(r-p-2)}}{1 + \frac{p}{(r-p-2)}},$$

and we can see that a greater variance reduction occurs when the number of parameters is large compared to the Phase II sample size. If r = 5p + 2, the largest percent reduction is 20%, if r = 10p + 2, the largest reduction would be 10%, and if r = 20p + 2, this reduction would be 5%.

Not ignoring the second term of the variance, the maximum possible variance reduction, denoted as α_0, is

$$\alpha_0 = \frac{\frac{(n-r)p}{n(r-p-2)}}{1 + \frac{R^2 r}{(1-R^2)n} + \frac{(n-r)p}{n(r-p-2)}}.$$

The actual reduction for a given bound, denoted as α, is

$$\alpha = \frac{\{1 - \Delta(b,p)\}\frac{(n-r)p}{n(r-p-2)}}{1 + \frac{R^2 r}{(1-R^2)n} + \frac{(n-r)p}{n(r-p-2)}}.$$

If we desire a certain percent reduction in variance, α, then we can calculate the upper bound needed to achieve this reduction, when $\alpha < \alpha_0$, using

$$\Delta(b,p) = 1 - \alpha - \frac{\alpha\{1 - R^2(n-r)\}n(r-p-2)}{(1-R^2)n(n-r)p}.$$

Once we have solved for $\Delta(b,p)$, we can approximate b using statistical software.

Sample Size Gained from Bounded Sampling

Here we look at the equivalent Phase II sample size necessary under simple random sampling in order to achieve the same variance as using bounded sampling for large sample sizes. Letting m denote the additional Phase II sample size, we have

$$\frac{1}{r}\left\{1 + \Delta(b,p)\frac{(n-r)p}{n(r-p-2)}\right\} = \frac{1}{r+m}\left\{1 + \frac{(n-r-m)p}{n(r+m-p-2)}\right\}.$$

The maximum sample size increase achievable, m_0, occurs when $\Delta(b,p) = 0$.

$$m_0 = \frac{\sqrt{(nr + rp - np - 2n)^2 - 4n(r^2p - nrp)} - (nr + rp - np - 2n)}{2n}.$$

If a certain sample size gain is desired, we can specify a value for m, where $m < m_0$, and determine the bound necessary to achieve that sample size gain. Our bound can be found taking

$$\Delta(b,p) = \frac{r(n-r-m)(r-p-2)}{(r+m)(n-r)(r+m-p-2)} - \frac{mn(r-p-2)}{p(r+m)(n-r)}$$

and solving for b from $\Delta(b,p)$ using statistical software.

2.2.2 Estimation of Regression Parameters

For many studies, the relationship between X and Y is more important than the mean of Y. In those cases, we propose an alternative criterion for selecting the Phase II sample that

increases the precision of the regression parameter estimates.

Let $Z = (1_{n\times 1}, X_1, ..., X_p)$ denote the collection of X covariates plus a column of 1s corresponding to the intercept term in a regression model. In two-phase studies, only individuals having Y measured contribute to the regression of Y on Z. Using ordinary least squares regression, we have

$$\widehat{\beta} = (Z_r^T Z_r)^{-1} Z_r^T Y,$$

where Z_r are the Z variables for individuals in Phase II. The variance of $\widehat{\beta}$ from this regression is given as $\sigma^2(Z_r^T Z_r)^{-1}$. We can see that samples with larger values of $Z_r^T Z_r$ will produce a smaller variance for $\widehat{\beta}$. The Z variable that corresponds to $\widehat{\beta}_0$ will always be a column of 1s and will not be affected by choosing large values of $Z_r^T Z_r$. As result, we can ignore the intercept term and focus on the other regression parameters, $\widehat{\beta}^*$, whose variances are given by

$$\sigma^2\{(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)\}^{-1},$$

where σ^2 is the conditional variance of Y given X, X_r is the $r \times p$ matrix of X values for individuals in Phase II, $\overline{X}_r$ is a $p\times 1$ vector containing the Phase II sample means of $X_1, ...X_p$, and $1_{r\times 1}$ is a $r\times 1$ vector of 1s. X_r does not contain a column of 1s. $X_r - 1_{r\times 1}\overline{X}_r^T$ is simply the X values for individuals from Phase II centered at their Phase II mean. If $X_i \sim N_p(\mu_X, \Sigma)$, $(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T) \sim W_p(\Sigma, r-1)$, with $W_p(\Sigma, r-1)$ denoting a p dimensional Wishart distribution with $r-1$ degrees of freedom and scale matrix Σ. For other instances, we can instead find the empirical distribution by re-sampling.

We propose using bounded sampling to lower the variance of our regression parameters, similar to our method for lowering the variance of our estimate of the mean. In this instance, we only want to select Phase II subsamples where $(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)$ is large; however, since $(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)$ is a $p\times p$ positive definite matrix, determining an upper bound is more complicated. We recommend using $log|(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)|$ and only selecting Phase II samples where this quantity is greater than some specified lower bound.

Taking the determinant reduces the bound to a scalar, which is easier to work with. We can easily approximate the distribution of $log|(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)|$ when X is multivariate normal by simulating from a $W_p(\Sigma, r-1)$ and taking the appropriate transformation, enabling us to again use approximate quantiles in selecting an achievable

bound. Applying the natural logarithm to the determinant makes the distribution asymptotically normal and the limiting distribution has a closed form (Cai et al., 2015), which allows us to determine approximate quantiles without using simulations when X is multivariate normal and our sample size is large. We can instead use the empirical distribution of $log|(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)|$ to find quantiles when X is not multivariate normal. We can estimate the expected value of $[(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)]^{-1}$ when using this bound, through simulations or from the empirical distribution. We can then approximate the reduction in the variance of $\widehat{\beta}^*$ for a selected bound compared to simple random sampling. The percent reduction in variance is the same across all regression parameters when X is multivariate normal.

When interest lies in estimating both the mean and regression parameters, we can select a Phase II sample using both the upper bound for the F-distribution and the lower bound for the regression parameters. Selecting a sample using both bounds will simultaneously improve estimation of the mean of Y and the parameters from the regression of Y on X. When X is multivariate normal, we need only examine each individual bound on its own. For other instances, we have to consider the joint distribution of both bounds, which we can calculate empirically.

2.2.3 Simulations

We performed simulations to evaluate how the use of two bounds would affect the estimates and variances for mean and regression parameters. Using a fixed Phase I sample size of 500, X was generated from a multivariate Gaussian distribution. We then selected observations for the Phase II sample and generated values for Y conditional on X from a normal distribution. We computed $\widehat{\mu}_Y$ and $\widehat{\beta}$ and their variance from the data. Table 2.2 displays the results for parameter estimates and their variances based on selecting our Phase II sample using either simple random sampling or using the bounds for improving the mean and regression parameters. The table displays the average of the parameter estimates and variances under each sampling scheme. We also calculated the empirical variance of each estimate, which was similar to the average variance. The use of either bound for selecting the second phase sample did not affect the bias of parameter estimates. From the table, we can see that the mean bound lowers the variance of $\widehat{\mu}_Y$ and $\widehat{\beta}_0$ but does not have much of an effect on the other parameters, while the regression bound does not affect the variance of those two parameters but improves the precision of the other parameter estimates. Using both bounds simultaneously lowers the variance of all parameters. The variance of each parameter when

using both bounds is similar to its variance when using each bound on its own. The variance of $\widehat{\mu}_Y$ and $\widehat{\beta}_0$ are nearly lowered to that obtained when using a sample size of 55 for simple random sampling in Phase II. For the other regression parameters, the variance is the same as the variance obtained from a Phase II sample size of 60 when using simple random sampling, equivalent to a 20% increase in Phase II sample size.

Table 2.2: Simulation results for parameter estimates and variances for simple random sampling versus using either the F-distribution bound for mean estimation, the bound for regression parameter estimation, or both bounds simultaneously.

		SRS			Bounded	r = 50	
Type	Parameter	r = 50	r = 55	r = 60	Mean	Regression	Double
Estimate	μ_Y	10.00	10.03	10.00	10.00	10.01	10.01
	β_0	10.00	10.02	9.99	10.01	10.03	10.01
	β_1	1.00	0.99	0.99	1.00	0.99	1.00
	β_2	-1.00	-0.99	-1.01	-1.00	-1.01	-1.00
	β_3	2.00	2.01	2.01	1.99	2.01	2.00
	β_4	-3.00	-3.00	-3.00	-2.99	-2.99	-2.99
	β_5	0.98	1.00	1.00	0.99	1.01	1.00
Variance	μ_Y	1.268	1.156	1.070	1.174	1.267	1.172
	β_0	1.09	0.982	0.893	1.00	1.069	0.997
	β_1	0.291	0.260	0.236	0.282	0.236	0.237
	β_2	0.213	0.191	0.174	0.209	0.174	0.174
	β_3	0.242	0.216	0.197	0.238	0.197	0.196
	β_4	0.438	0.391	0.356	0.431	0.357	0.356
	β_5	0.639	0.565	0.521	0.625	0.521	0.519

2.3 Sample Selection for Single Binary X and Y

Thus far, we have discussed selecting our Phase II sample to improve parameter estimation for a continuous Y. We now examine sample selection to improve the efficiency of our estimators for the case of a single covariate, X, where X and Y are both binary. We first discuss the maximum likelihood and Bayesian posterior estimates and variances for the mean, risk difference, and log odds ratio for two-phase designs and the sample size allocation that will minimize that variance. We then discuss our adaptive two-phase sampling approach to lower the variance of these estimators and present simulation results. Similar methods to those presented in this section can be applied to cases where we have a single binary X and a normal Y.

2.3.1 Estimation of the Mean

Let $X_i \sim Bernoulli(p)$, $Y_i|X_i = 0 \sim Bernoulli(\pi_0)$, and $Y_i|X_i = 1 \sim Bernoulli(\pi_1)$. Then we can estimate the mean of Y using

$$\widehat{\mu}_Y = \widehat{p}\,\widehat{\pi}_1 + (1-\widehat{p})\widehat{\pi}_0,$$

where $\widehat{p}$, $\widehat{\pi}_0$, and $\widehat{\pi}_1$ are the maximum likelihood estimates for their respective parameter. The variance of $\widehat{\mu}_Y$ is approximately

$$\frac{p^2\pi_1(1-\pi_1)}{r_1} + \frac{(1-p)^2\pi_0(1-\pi_0)}{r_0} + \frac{p(1-p)(\pi_1-\pi_0)^2}{n},$$

where n is again the Phase I sample size, and r_0 and r_1 are the Phase II sample sizes when X = 0 and X = 1, respectively. Assuming the first phase has already taken place, then for a fixed Phase II sample size, $r = r_0 + r_1$, selecting

$$r_0 = \frac{r}{1 + \sqrt{\frac{p^2\pi_1(1-\pi_1)}{(1-p)^2\pi_0(1-\pi_0)}}} \tag{2.4}$$

would yield the smallest variance for the estimate μ_Y.

Unfortunately, implementation of the optimal design for improving the variance of our mean estimate requires some knowledge of the conditional probabilities of Y given X. If we possessed a priori information on π_0 and π_1, we might be interested in using a Bayesian estimator. Erkanli et al. (1997) discussed Bayesian estimation and optimal designs for two-phase studies. In their paper, both the first and second phase sample size could vary, but we only consider the case where the Phase I and Phase II sample sizes are fixed and Phase I has been completed. We can write the joint likelihood for X and Y as

$$\begin{aligned} &\prod_{i=1}^{n} f(X_i|p) \prod_{i=1}^{r} f(Y_i|X_i = a, \pi_0, \pi_1) \\ &= \prod_{i=1}^{n} f(X_i|p) \prod_{i=1}^{r_0} f(Y_i|\pi_0) \prod_{i=1}^{r_1} f(Y_i|\pi_1). \end{aligned}$$

If we use independent beta priors for p, π_0, and π_1, $\pi(p) \sim Beta(\alpha_p, \beta_p)$, $\pi(\pi_0) \sim Beta(\alpha_0, \beta_0)$, and $\pi(\pi_1) \sim Beta(\alpha_1, \beta_1)$, the posterior distributions are independent beta distributions.

The posterior distribution of the population mean of Y, $\pi(\mu_Y|Y, X, r_0, r_1, n)$,

$$= \pi(p|X, n)\pi(\pi_1|Y, r_1) + \{1 - \pi(p|X, n)\}\pi(\pi_0|Y, r_0),$$

with $\pi(p|X, n)$, $\pi(\pi_0|Y, r_0)$, and $\pi(\pi_1|Y, r_1)$ indicating the posterior distribution for p, π_0, and π_1. Then,

$$E\{\pi(\mu_Y|Y, X, r_0, r_1, n)\} = \widetilde{p}\,\widetilde{\pi}_1 + (1 - \widetilde{p})\widetilde{\pi}_0,$$

where $\widetilde{p}$, $\widetilde{\pi}_0$, and $\widetilde{\pi}_1$ are posterior expectations for p, π_0 and π_1. The posterior variance,

$$V\{\pi(\mu_Y|Y, X, r_0, r_1, n)\} \approx \frac{\widetilde{p}^2\widetilde{\pi}_1(1 - \widetilde{\pi}_1)}{r_1 + \alpha_1 + \beta_1 + 1} + \frac{(1 - \widetilde{p})^2\widetilde{\pi}_0(1 - \widetilde{\pi}_0)}{r_0 + \alpha_0 + \beta_0 + 1} + \frac{\widetilde{p}(1 - \widetilde{p})(\widetilde{\pi}_1 - \widetilde{\pi}_0)^2}{n + \alpha_p + \beta_p + 1},$$

is very similar to the MLE variance. For a fixed Phase I sample and a Phase II sample size of r, this variance is minimized by choosing

$$r_0 = \frac{r + \alpha_1 + \beta_1 + 1 - \frac{\sqrt{\widetilde{p}^2\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}{\sqrt{(1-\widetilde{p})^2\widetilde{\pi}_0(1-\widetilde{\pi}_0)}}(\alpha_0 + \beta_0 + 1)}{\frac{\sqrt{\widetilde{p}^2\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}{\sqrt{(1-\widetilde{p})^2\widetilde{\pi}_0(1-\widetilde{\pi}_0)}} + 1}. \tag{2.5}$$

2.3.2 Estimation of the Risk Difference

Another potential quantity of interest when Y and X are binary is the risk difference, $\pi_1 - \pi_0$, whose mle variance is given by

$$\frac{\pi_1(1 - \pi_1)}{r_1} + \frac{\pi_0(1 - \pi_0)}{r_0}.$$

For a fixed Phase II sample size, this is minimized by choosing

$$r_0 = \frac{r}{1 + \sqrt{\frac{\pi_1(1-\pi_1)}{\pi_0(1-\pi_0)}}}. \tag{2.6}$$

For Bayesian estimation, the posterior distribution for the risk difference (RD) is given by

$$\pi(RD|Y, r_0, r_1) = \pi(\pi_1|Y, r_1) - \pi(\pi_0|Y, r_0), \text{ with}$$

$$E\{\pi(RD|Y, r_0, r_1)\} = \widetilde{\pi}_1 - \widetilde{\pi}_0, \text{ and}$$

$$V\{\pi(RD|Y, r_0, r_1)\} = \frac{\widetilde{\pi}_1(1-\widetilde{\pi}_1)}{r_1 + \alpha_1 + \beta_1 + 1} + \frac{\widetilde{\pi}_0(1-\widetilde{\pi}_0)}{r_0 + \alpha_0 + \beta_0 + 1}.$$

This is minimized for a fixed Phase I and Phase II sample size by selecting

$$r_0 = \frac{r + \alpha_1 + \beta_1 + 1 - \frac{\sqrt{\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}{\sqrt{\widetilde{\pi}_0(1-\widetilde{\pi}_0)}}(\alpha_0 + \beta_0 + 1)}{\frac{\sqrt{\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}{\sqrt{\widetilde{\pi}_0(1-\widetilde{\pi}_0)}} + 1}, \tag{2.7}$$

which is equivalent to minimizing the variance of the mean estimate of Y when $p = 0.50$. Minimizing the variance of the risk difference does not explicitly depend on the distribution of X, and is only affected by X for cases where the optimal choice for r_0 is greater than the number of subjects in the Phase II sample where X = 0 or, equivalently, if the ideal value of r_1 is greater than the number of individuals with X = 1.

2.3.3 Estimation of the Log-Odds Ratio

If, instead, we were interested in the parameter estimate for β_1 from the logistic regression of Y on X, the estimate of the natural logarithm of the odds ratio, $log(\frac{\pi_1}{1-\pi_1}) - log(\frac{\pi_0}{1-\pi_0})$, would be of primary interest. The mle variance of the log-odds is given by

$$\frac{1}{r_1\pi_1(1-\pi_1)} + \frac{1}{r_0\pi_0(1-\pi_0)},$$

and for a fixed Phase II sample is minimized by choosing

$$r_0 = \frac{r}{1 + \sqrt{\frac{\pi_0(1-\pi_0)}{\pi_1(1-\pi_1)}}}. \tag{2.8}$$

The posterior distribution for the natural logarithm of the odds ratio (logOR) of π_0 and π_1 is given by

$$\pi(logOR|Y, r_0, r_1) = logit\{\pi(\pi_1|Y, r_1)\} - logit\{\pi(\pi_0|Y, r_0)\}, \text{ with}$$

$$E\{\pi(logOR|Y, r_0, r_1)\} =$$

$$\psi\left(\alpha_1+\sum^{r_1}Y_i\right)-\psi\left(\beta_1+r_1-\sum^{r_1}Y_i\right)-\psi\left(\alpha_0+\sum^{r_0}Y_i\right)+\psi\left(\beta_0+r_0-\sum^{r_0}Y_i\right),$$

and

$$V\{\pi(logOR|Y,r_0,r_1)\}=$$

$$\psi_1\left(\alpha_1+\sum^{r_1}Y_i\right)+\psi_1\left(\beta_1+r_1-\sum^{r_1}Y_i\right)+\psi_1\left(\alpha_0+\sum^{r_0}Y_i\right)+\psi_1\left(\beta_0+r_0-\sum^{r_0}Y_i\right),$$

with ψ denoting the digamma function and ψ_1 denoting the trigamma function. As with the risk difference, the variance does not explicitly depend on X. Computing the variance given the data is fairly straight forward, but finding which allocation produces the best variance for a fixed Phase II sample size given the parameter estimates is somewhat difficult. We could replace $\sum^{r_0}Y_i$ and $\sum^{r_1}Y_i$ with their expectation given the parameters and then compute the variance for each possible value of r_0 and r_1 to determine which sample allocation produces the smallest variance. Another possibility is using

$$\frac{1}{(r_1+\alpha_1+\beta_1+1)\pi_1(1-\pi_1)}+\frac{1}{(r_0+\alpha_0+\beta_0+1)\pi_0(1-\pi_0)}$$

to approximate the posterior variance. This equation comes from using the delta method on the posterior variances of π_0 and π_1. Note that when α_0, β_0, α_1, and β_1 are fairly large the posterior distributions of π_0 and π_1 are approximately normal and the delta method works well as an approximation for the variance. Using this approximation, we would use

$$r_0=\frac{r+\alpha_1+\beta_1+1-\frac{\sqrt{\widetilde{\pi}_0(1-\widetilde{\pi}_0)}}{\sqrt{\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}(\alpha_0+\beta_0+1)}{\frac{\sqrt{\widetilde{\pi}_0(1-\widetilde{\pi}_0)}}{\sqrt{\widetilde{\pi}_1(1-\widetilde{\pi}_1)}}+1} \tag{2.9}$$

to determine our sample size allocation. We recommend using this approximation for estimating the optimal allocation, as it is computationally simple, and then using the trigamma function for the variance calculation after sampling.

2.3.4 Adaptive Sample Selection

The variance and optimal design depend on π_0 and π_1, which will likely not be known before Phase II sampling. This makes it more difficult to design a strategy for selecting the

second phase subsample than in the continuous case. To overcome this, we propose using an adaptive approach where the Phase II sample is conducted in several batches. We can use the information obtained in earlier batches for preliminary estimates of π_0 and π_1, and choose the best allocation for the next batch based on both our current estimates and the previous allocations. A Bayesian approach is more suited for this method of sampling since after each batch we can compute the posterior estimates for our parameters and use our posterior distribution as the prior for our next batch, allowing us to update our parameter estimates at each step. Additionally, use of a vaguely informative prior is helpful for cases where either π_0 or π_1 are rare, as the maximum likelihood estimate of $\widehat{\pi}_0$ or $\widehat{\pi}_1$ could end up being zero. In which case, equations (2.4), (2.6), and (2.8) would not work for choosing the next allocation.

Our simulations have shown very little change in efficiency from increasing the number of batches beyond three, and two or three batches would be sufficient in most cases. For the first batch, we either select an equal sample size for r_0 and r_1 or sample sizes proportional to $\widetilde{p}$. After our first batch is collected, we compute our posterior distributions for π_0 and π_1, which become our prior distributions for the next batch. We then allocate the sample size in that batch by plugging our prior means for π_0 and π_1 in equations (2.5), (2.7), or (2.9). If r_0 is less than zero, we take $r_0 = 0$, similarly, if r_0 is greater than the next batch size, we take r_0 equal to the batch size. The sum of r_0 and r_1 always equals the batch size.

2.3.5 Simulations

We performed simulations to compare the performance of our proposed adaptive design against Bayesian estimation using proportional sampling in each strata, an equal sample size in each strata, and the optimal allocation. The optimal allocation is impossible unless π_0 and π_1 are known beforehand. We used a fixed Phase I sample size of 1,000 but varied π_0, π_1, p, and the total Phase II sample size in our simulations. We performed simulations, optimizing on the mean, risk difference, or log-odds. Table 2.3 displays the average variance for estimating the mean of Y under each of the proposed designs. Our adaptive approach always performs at least as well as proportional and equal allocation. When $\pi_0(1-\pi_0)$ equals $\pi_1(1-\pi_1)$, adaptive sampling will not improve the variance; however, when those values greatly differ, adaptive sampling can substantially lower the variance. For very large sample sizes or if π_0 and π_1 are not too small or large, the variance produced by our adaptive method approaches the variance under the best possible allocation. While for smaller sample sizes or when π_0 or π_1 are close to 0 or 1, the deviation in variance from our method to the

best allocation becomes more pronounced, likely due to the difficulty in estimating π_0 or π_1 in initial batches. However, our method still performs considerably better than proportional or equal sampling in these scenarios. We took α_0, β_0, α_1, and β_1 equal to 1 for a vaguely informative prior in the adaptive, proportional, and equal sampling designs. This prior was chosen as it did not provide much information about the parameters, and guaranteed the posterior for π_0 and π_1 would not be degenerate. For the optimal design, we fixed the sum of α_0 and β_0 and the sum of α_1 and β_1 at the same value as the other designs, but selected values so the prior mean would match the true means. The variance for our adaptive design would likely be improved from using priors distribution for π_0 and π_1 whose means are closer to the actual means.

Tables 2.4 and 2.5 display simulation results for the risk difference and log-odds ratio. Results are very similar to those from mean estimation. The only major discrepancy is the difference in the variance from our adaptive design and the optimal design does not improve as the sample size increases for the log-odds ratio. This is likely to due us having to approximate the posterior variance for allocation purposes, unlike when estimating the mean or risk difference. Since the value of p does not directly affect estimation and proportional allocation generally does not work well for these two quantities, they have been omitted from the table.

Table 2.3: Comparison of the variance of the estimated mean for adaptive sampling, proportional sampling, equal strata size, and optimal allocation. Phase I sample size is fixed at 1000 for each scenario.

r	p	π_0	π_1	Proportional	Equal	Adaptive	Best
300	0.50	0.01	0.50	0.000491	0.000491	0.000399	0.000352
300	0.50	0.05	0.50	0.000541	0.000542	0.000482	0.000469
300	0.50	0.10	0.50	0.000598	0.000599	0.000566	0.000560
300	0.50	0.30	0.70	0.000725	0.000726	0.000724	0.000722
300	0.65	0.01	0.50	0.000605	0.000753	0.000530	0.000475
300	0.35	0.01	0.50	0.000366	0.000297	0.000276	0.000241
200	0.50	0.01	0.50	0.000705	0.000707	0.000588	0.000492
400	0.50	0.01	0.50	0.000383	0.000384	0.000308	0.000281

2.3.6 Simultaneous Estimation for Multiple Quantities

Sample selection of Phase II samples for binary X and Y has focused on estimating either the mean, the risk difference, or natural logarithm of the odds ratio. In practice researchers may be interested in two or more of those estimates. Unfortunately, selecting an allocation

Table 2.4: Comparison of the variance of the estimated risk difference when using either adaptive sampling or equal size in each stratum with the variance under the optimal design. Phase I sample size is fixed at 1000 for each scenario.

r	π_0	π_1	Equal	Adaptive	Best
300	0.01	0.50	0.00173	0.00137	0.00117
300	0.05	0.50	0.00197	0.00173	0.00167
300	0.10	0.50	0.00223	0.00210	0.00208
300	0.50	0.50	0.00325	0.00325	0.00325
200	0.01	0.50	0.00259	0.00213	0.00173
400	0.01	0.50	0.00130	0.00100	0.00881

Table 2.5: Comparison of the variance for the estimated natural logarithm of the odds ratio when using either adaptive sampling or equal size in each stratum with the variance under the optimal design. Phase I sample size is fixed at 1000 for each scenario.

r	π_0	π_1	Equal	Adaptive	Best
300	0.01	0.50	0.758	0.545	0.484
300	0.05	0.50	0.177	0.148	0.145
300	0.10	0.50	0.102	0.096	0.095
300	0.50	0.50	0.053	0.053	0.053
200	0.01	0.50	1.002	0.802	0.726
400	0.01	0.50	0.602	0.404	0.363

that is optimal for all three estimates is impossible. We instead propose several methods for choosing a sample that will produce good results when estimating two or more quantities, again using adaptive sampling.

The first and simplest method is finding the best allocation under each estimate individually and then taking the average for r_0 and r_1. Each variance is parabolic across the values for r_0, and generally there is some region $r_0 < a$ or $r_0 > b$ that produces poor results for all three estimates. By averaging the best allocation for all three estimates, we can avoid regions that are suboptimal for all three estimates. We could use a weighted average if we are more interested in estimating certain quantities.

Two more possibilities are minimizing the average $^{se}/_{estimate}$ or minimizing the average percent increase in variance over the optimal allocation for each estimate. The $^{se}/_{estimate}$ places more weight on estimates with a smaller standard error compared to parameter estimate. This quantity can be thought of as analogous to $^{1}/_{Z}$, where Z is the Z-statistic for testing the null hypothesis that the estimate equals zero. This tends to down weight the mean in determining the allocation. Both quantities require computation of the variance across all possible Phase II allocations and calculation of the quantity to be minimized.

Again, we could use a weighted average if we place higher importance on certain estimates.

2.4 Categorical Z and Binary X and Y

Finally, for two-phase estimation with a binary outcome, we consider the case where we have a categorical variable, Z, with k categories and a binary variable, X, collected in Phase I. Note the categories for Z could be constructed from multiple continuous or categorical variables. The k categories could represent propensity score classes from the logistic regression of X regressed on numerous covariates, allowing us to apply this scenario to any two-phase sample where Y is binary and at least one Phase I covariate is binary.

Let q_j denote the probability of belonging to category j and p_j denote the probability that X equals 1 given Z equals j. Let π_{1j} denote the probability that Y equals 1 given X = 1 and Z = j and π_{0j} denote the probability that Y equals 1 given X = 0 and Z = j. Within category j, we can estimate the mean of Y and the variance of that estimate the way we estimated the mean and variance with a single binary X and Y. We have

$$\widehat{\mu}_{Yj} = \widehat{p}_j\widehat{\pi}_{1j} + (1-\widehat{p}_j)\widehat{\pi}_{0j}, \text{ and}$$

$$V(\widehat{\mu}_{Yj}) \approx \frac{p_j^2\pi_{1j}(1-\pi_{1j})}{r_{1j}} + \frac{(1-p_j)^2\pi_{0j}(1-\pi_{0j})}{r_{0j}} + \frac{p_j(1-p_j)(\pi_{1j}-\pi_{0j})^2}{n_j},$$

where n_j denotes the Phase I sample size where Z = j, r_{1j} denotes the Phase II sample size where Z = j and X = 1, and r_{0j} denotes the sample size where Z = j and X = 0. After computing the mean and variance within each category of Z, we can compute the overall mean as

$$\widehat{\mu}_Y = \sum_{i=1}^{k} \widehat{q}_i\widehat{\mu}_{Yi},$$

and the variance as

$$\begin{aligned} V(\widehat{\mu}_Y) &= \sum_{i=1}^{k} q_i^2 V(\widehat{\mu}_{Yi}) + \sum_{i=1}^{k} \mu_{Yi}^2 V(\widehat{q}_i) + \sum_{i\neq j} cov(\mu_{Yi}\widehat{q}_i, \mu_{Yj}\widehat{q}_j) \\ &= \sum_{i=1}^{k} q_i^2 V(\widehat{\mu}_{Yi}) + \sum_{i=1}^{k} \mu_{Yi}^2 \frac{q_i(1-q_i)}{n} + \sum_{i\neq j} \frac{\mu_{Yi}q_i\mu_{Yj}q_j}{n}. \end{aligned}$$

For a fixed Phase I and Phase II sample size, the variance is minimized by choosing

$$\frac{r_{0j}}{r_{1j}} = \frac{\sqrt{(1-p_j)^2\pi_{0j}(1-\pi_{0j})}}{\sqrt{p_j^2\pi_{1j}(1-\pi_{1j})}}, \text{ and}$$

$$\frac{r_{1j}}{r_{1i}} = \frac{\sqrt{q_j^2p_j^2\pi_{1j}(1-\pi_{1j})}}{\sqrt{q_i^2p_i^2\pi_{1i}(1-\pi_{1i})}}.$$

If our primary interest is in the relationship between X and Y, adjusted for Z, then we can instead calculate the risk difference or natural logarithm of the odds ratio in each category of Z and combine the estimates with a weighted sum to obtain the adjusted risk difference or log of the odds ratio. The overall variance is given by

$$\sum_{i=1}^{k} q_i^2 V(\widehat{RD_i}) + \sum_{i=1}^{k} RD_i^2 V(\widehat{q_i}) + \sum_{i\neq j} cov(RD_i\widehat{q_i}, RD_j\widehat{q_j})$$

for the risk difference, and

$$\sum_{i=1}^{k} q_i^2 V(\widehat{OR_i}) + \sum_{i=1}^{k} OR_i^2 V(\widehat{q_i}) + \sum_{i\neq j} cov(OR_i\widehat{q_i}, OR_j\widehat{q_j})$$

for the log-odds ratio, where RD_j and OR_j are the risk difference and log odds ratio when Z = j. The variance within each category can be computed from the variance formulas with a single binary X and Y. For fixed Phase I and Phase II sample sizes, we can minimize these variances by choosing

$$\frac{r_{0j}}{r_{1j}} = \frac{\sqrt{\pi_{0j}(1-\pi_{0j})}}{\sqrt{\pi_{1j}(1-\pi_{1j})}} \text{ and } \quad \frac{r_{1j}}{r_{1i}} = \frac{\sqrt{q_j^2\pi_{1j}(1-\pi_{1j})}}{\sqrt{q_i^2\pi_{1i}(1-\pi_{1i})}}$$

for the risk difference, and

$$\frac{r_{0j}}{r_{1j}} = \frac{\sqrt{\pi_{1j}(1-\pi_{1j})}}{\sqrt{\pi_{0j}(1-\pi_{0j})}} \text{ and } \quad \frac{r_{1j}}{r_{1i}} = \frac{\sqrt{q_j^2\pi_{1i}(1-\pi_{1i})}}{\sqrt{q_i^2\pi_{1j}(1-\pi_{1j})}}$$

for the log odds ratio.

Since π_{0j} and π_{1j} are likely unknown prior to Phase II sampling, we suggest obtaining

Phase II samples in a small number of batches and using a Bayesian approach to update our estimates for π_{0j} and π_{1j}. We then use this information to select our Phase II sample, as in section 2.3. Using a Dirichlet prior for $q_1, ...q_k$ and independent beta priors for $p_1, ..., p_k$, $\pi_{11}, ..., \pi_{1k}$, and $\pi_{01}, ..., \pi_{0k}$ give us conjugate priors. We can estimate the posterior mean, risk difference, and log odds ratio within each category for Z and combine the posterior estimates and variances the same way we did with the frequentist estimates to obtain the posterior mean and variance of the overall mean and the adjusted risk difference and log odds ratio. Estimates within each category are the same as those obtained with a single binary X and Y. For estimating the overall mean, we can minimize the variance by choosing

$$\frac{r_{0j} + \alpha_{0j} + \beta_{0j} + 1}{r_{1j} + \alpha_{1j} + \beta_{1j} + 1} = \frac{\sqrt{(1-p_j)^2 \pi_{0j}(1-\pi_{0j})}}{\sqrt{p_j^2 \pi_{1j}(1-\pi_{1j})}}$$

and

$$\frac{r_{1j} + \alpha_{1j} + \beta_{1j} + 1}{r_{1i} + \alpha_{1i} + \beta_{1i} + 1} = \frac{\sqrt{q_j^2 p_j^2 \pi_{1j}(1-\pi_{1j})}}{\sqrt{q_i^2 p_i^2 \pi_{1i}(1-\pi_{1i})}},$$

where α_{1j}, β_{1j}, α_{0j}, and β_{0j} are the prior parameters for π_{1j} and π_{0j}. We can similarly find the best sample size allocation for the risk difference and use the approximation for the best allocation for the log odds ratio, as we did in section 2.3.

2.5 Normal Y and a Single Binary X

In the following section, we show how to estimate the overall mean and difference in means for a two-phase study where Y is normal and X is binary. For this instance, if we assume that $Y_i|X_i = 0 \sim N(\mu_0, \sigma_0^2)$ and $Y_i|X_i = 1 \sim N(\mu_1, \sigma_1^2)$, then we can estimate the mean of Y as

$$\widehat{\mu}_Y = \widehat{p}\widehat{\mu}_1 + (1-\widehat{p})\widehat{\mu}_0,$$

where p is the probability that X = 1. The variance is given by

$$V(\widehat{\mu_Y}) \approx p^2 \frac{\sigma_1^2}{r_1} + (1-p)^2 \frac{\sigma_0^2}{r_0} + \frac{p(1-p)}{n}(\mu_1 - \mu_0)^2.$$

This variance is minimized for fixed Phase I and Phase II sample sizes by choosing

$$\frac{r_1}{r_0} = \frac{p\sigma_1}{(1-p)\sigma_0}.$$

If we are interested in the effect of X on Y, then we could look at the difference in means, $\mu_1 - \mu_0$, which is analogous to β_1 from the linear regression of Y on X. The variance of the difference in means is given by

$$\frac{\sigma_1^2}{r_1} + \frac{\sigma_0^2}{r_0},$$

and is minimized for a fixed Phase I and Phase II sample size with

$$\frac{r_1}{r_0} = \frac{\sigma_1}{\sigma_0}.$$

When σ_1 and σ_0 are unknown, we might consider using an adaptive approach for selecting our Phase II sample using Bayesian estimation. We suggest using conjugate priors for our parameters, $\pi(p) \sim Beta(\alpha_p, \beta_p)$, $\pi(\mu_0|\sigma_0^2) \sim N(\mu_{\phi 0}, \sigma_0^2/k_{\phi 0})$, $\pi(\sigma_0^2) \sim Inverse\,\chi^2(\nu_{\phi 0}, \sigma_{\phi 0}^2)$, $\pi(\mu_1|\sigma_1^2) \sim N(\mu_{\phi 1}, \sigma_1^2/k_{\phi 1})$, and $\pi(\sigma_1^2) \sim Inverse\,\chi^2(\nu_{\phi 1}, \sigma_{\phi 1}^2)$. The posterior distributions of μ_0 and σ_0 are independent of μ_1 and σ_1, which are independent of p, making calculating the posterior for μ_0 and σ_0 the same as in the case of a single normal variate with unknown mean and variance. The variance for the posterior expectation of μ_Y is given by

$$V(\widehat{\mu}_Y) \approx \widehat{p}^2 \frac{\widehat{\sigma}_1^2}{r_1 + k_{\phi 1}} + (1-\widehat{p})^2 \frac{\widehat{\sigma}_0^2}{r_0 + k_{\phi 0}} + \frac{\widehat{p}(1-\widehat{p})}{n + \alpha_p + \beta_p + 1}(\widehat{\mu}_1 - \widehat{\mu}_0)^2,$$

with $\widehat{p}$, $\widehat{\sigma}_0$, $\widehat{\sigma}_1$, $\widehat{\mu}_0$, and $\widehat{\mu}_1$ denoting the posterior mean corresponding to those parameters. This is minimized by solving

$$\frac{r_1 + k_{\phi 1}}{r_0 + k_{\phi 0}} = \frac{\widehat{p}\widehat{\sigma}_1}{(1-\widehat{p})\widehat{\sigma}_0}.$$

The variance for the posterior estimate of difference in means is given by

$$\frac{\widehat{\sigma}_1^2}{r_1 + k_{\phi 1}} + \frac{\widehat{\sigma}_0^2}{r_0 + k_{\phi 1}},$$

which is minimized by

$$\frac{r_1 + k_{\phi 1}}{r_0 + k_{\phi 0}} = \frac{\widehat{\sigma}_1}{\widehat{\sigma}_0}.$$

As in the case of binary X and Y, these methods for estimating the mean and difference in mean for a normal Y and binary X can easily be extended to the case of a binary X and categorical Z.

2.6 Example with NHANES Data

The National Health and Nutrition Examination Survey (NHANES) is a research program by the National Center for Health Statistics (CDC) examining health and nutrition in the United States (Curtin et al., 2012). The survey conducts both interviews and physical examinations, including laboratory tests. We used data from NHANES to examine how effective our methods for sample selection would be in practice for a two-phase study. Selected variables to represent Phase I included demographic information on race, gender, age, marital status, and income, as well as general health information on diet, self reported diabetes diagnosis, blood pressure, BMI, smoking status, minutes of activity, television watching, computer usage, and self-reported overall health status. Our Phase II outcome was the laboratory variable glycohemoglobin, a blood test measuring how much sugar is bound to hemoglobin. This test is used to diagnose diabetes and prediabetes. Our main interests in this study are estimation of the mean of glycohemoglobin for this population and the relationship between obesity and glycohemoglobin, adjusted for our other covariates.

We chose all 30,468 NHANES participants from 2010 through 2015 with any available data on selected variables. Missing values were singly imputed through predictive mean matching, creating our pseudo-population. From this population, we first selected 3,000 individuals as our Phase I sample. In our Phase I sample, we created three propensity score classes using the predicted probability of obesity given the other selected covariates. The propensity score classes form our categorical Z variable, as described in section 2.4, and the indicator for obesity is our binary X. Afterwards, we selected 500 individuals for our Phase II sample and obtained their glycohemoglobin as our Y, which was approximately normal. We first conducted our Phase II sampling with the goal of estimating the mean of glycohemoglobin in this population. We used several different methods for selecting our Phase II sample and computed the parameter estimates and standard errors over many iterations.

This experiment was latter repeated for estimating the difference in mean glycohemoglobin between individuals with obesity and those without.

Table 2.6 displays results for estimating the population mean using multiple methods to select the Phase II sample and estimate the mean. The estimates for the population mean are similar across all methods. For the first method, $\overline{Y}$, we select our Phase II sample using simple random sampling (SRS) and then use $\overline{Y}$ as our estimate for the population mean. This method does not make use of any Phase I variables for sample selection or estimation and produces the largest standard error. Stratification by obesity selects the Phase II sample using the optimal allocation when stratifying by obesity, and then estimates the population mean through that same stratification. Use of obesity on its own lowers the standard error over using $\overline{Y}$ by a small amount. The next method again selects the Phase II sample using SRS, but then uses obesity and the propensity score for stratification to estimate the population mean. This further lowers the standard error over using $\overline{Y}$ or only stratifying by obesity. It should be noted that when using simple random sampling to select the Phase II sample, we may not always be able to use the stratified estimator due to empty strata. For this analysis, we only looked at instances where it was possible to use the stratified estimator.

The other three methods used both obesity and the propensity score for stratification to choose the Phase II sample and estimate the population mean. The standard error from the proportional allocation was similar to using simple random sampling and then stratifying. This is unsurprising, as on average SRS will give us a sample size proportional to the size of each strata. Explicitly stratifying by obesity and the propensity score to select the Phase II sample guarantees that we can always use the stratified estimator, unlike SRS. Our adaptive sample selection method greatly outperformed proportional allocation in terms of variance. The average standard error from the adaptive design was close to the standard error from the optimal design, and the optimal design can only be achieved if the population parameters are known beforehand.

Table 2.7 compares several methods in estimating the difference in the average value of glycohemoglobin between individuals with obesity and those without, adjusted for the other Phase I variables through stratification on the propensity score classes. We selected the Phase II sample using either SRS, proportional allocation, an equal sample size in each strata, our adaptive approach, or the optimal allocation. In terms of standard errors, SRS again performed similarly to proportional allocation, having an equal sample size in each strata worked much better than both methods, and the adaptive design performed better than those

three methods. Our adaptive approach produced standard errors close to those obtained from the optimal design, which of course depended on unknown population parameters, and produced relatively similar estimates to the other approaches. Note that only the adaptive approaches used Bayesian estimation in these instances.

Table 2.6: Estimate and standard error for the mean of glycohemoglobin.

Method	Estimate	Standard Error
$\overline{Y}$	5.570	0.0403
Stratification by Obesity	5.571	0.0396
Stratifying by Obesity and PS after SRS	5.571	0.0383
Proportional Allocation using Obesity and PS	5.572	0.0383
Adaptive Allocation using Obesity and PS	5.568	0.0355
Optimal Allocation using Obesity and PS	5.571	0.0350

Table 2.7: Estimate and standard error for the difference in mean glycohemoglobin between those with obesity and those without, adjusted for propensity score classes.

Method	Estimate	Standard Error
Simple Random Sampling	0.207	0.1183
Proportional Allocation	0.204	0.1137
Equal Sample Size in Each Strata	0.206	0.0819
Adaptive Allocation	0.206	0.0766
Optimal Allocation	0.208	0.0754

2.7 Discussion

Two-phase survey sampling can be used to reduce costs for estimating a costly outcome in a study. Although two-phase sampling as proposed by Neyman (1938) is most frequently used in epidemiological studies on disease prevalence, it can be applied to any large-scale study where a disease, biomarker, or other outcome is too expensive to measure on the entire survey sample. Most literature on two-phase sampling focuses on estimating the population mean of the outcome measured in Phase II.

In this chapter, we discussed both mean estimation and estimating the relationship between our outcome in Phase II with the Phase I variables. We presented several methods for selecting our second phase sample, depending on the distribution of our variables. When Y is continuous, we can improve estimation of the mean of Y by only choosing Phase II samples where $\frac{r-p}{p(r-1)}\frac{(n-r)r}{n}(\overline{X}_{n-r}-\overline{X}_r)^T(S_r^2)^{-1}(\overline{X}_{n-r}-\overline{X}_r)$ falls below a pre-selected bound.

This method will always outperform simple random sampling, though how large of an improvement greatly depends on the sample sizes and number of X covariates. Also, there is an upper bound for the maximum possible percent reduction in variance. Similarly, estimation of the regression parameters for the conditional distribution of Y given X can be improved by selecting Phase II samples with large values of $log|(X_r - 1_{r\times 1}\overline{X}_r^T)^T(X_r - 1_{r\times 1}\overline{X}_r^T)|$, which only depend on X. Unlike for the mean, we do not have a formula to approximate how much this method will improve regression parameter estimation, but we can use simulations to estimate the reduction in variance. These criteria can be applied simultaneously to improve estimation of the mean and regression parameters. The increased difficulty in finding a suitable subsample is the only major drawback to applying these criteria.

We also presented methods for improving precision for two-phase estimation when Phase II variables are non-normal and Y is binary or normal, which typically depended on unknown parameters. To overcome this limitation, we proposed selecting our Phase II sample in batches and using Bayesian estimation to update our estimates after each batch. Using this adaptive method over other simpler methods can greatly improve the variance for estimating our quantity of interest. One potential downside of this approach is the use of Bayesian priors can introduce small bias in our parameter estimates. Also, this method only improves the precision of a single estimate.

The methods we have presented in this chapter can be used in a number of studies where an outcome is measured in only a subsample of total participants. By using propensity score classes to construct a categorical variable, Z, we can apply our method from section 2.4 in numerous instances. We have only considered cases where Y is continuous or binary, and may want to explore instances where Y follows other distributions. Additionally, we only considered cases with a single Y variable as our Phase II outcome. It may be worthwhile to examine designing two-phase studies where multiple variables are measured at the second phase.

Chapter 3

Split Questionnaire Design for Panel Surveys

3.1 Introduction

Longitudinal or panel surveys are essential for any study of change in the key outcome variables and their correlates. Given the cost of conducting and recruiting participants into such studies, researchers often try to get the most information they can out of study participants. Survey questions often are pooled from several investigators with multiple research interests, resulting in long questionnaires and increasing the burden on participants. These studies, therefore, are also subject to drop-out or nonresponse, which may lead to biased results. Furthermore, longer surveys can affect the quality of participants' responses. Several studies have shown that nonresponse rates tend to be high in surveys with long questionnaires (Adams & Darwin, 1982; Dillman et al., 1993; Roszkowski & Bean, 1990; Beckett et al., 2016), and item nonresponse is more frequent towards the end of a questionnaire (Raghunathan & Grizzle, 1995). Past a certain length participants become more likely to lose interest in the study, making responses less accurate (Herzog & Bachman, 1981; Gonzalez & Eltinge, 2007; Peytchev & Peytcheva, 2017). The problems caused by lengthy questionnaires are exacerbated in longitudinal studies due to calling on respondents to fill out the same questionnaires repeatedly. A shorter survey length alleviates these problems and potentially decreases the cost of data collection per subject. The goal, therefore, is to balance collecting a set of rich variables while not placing an undue burden on participants.

Meanwhile, advancements in software for handling missing data, especially the multiple imputation approach, have made missing data less problematic for data analysis, and, in fact, designing a survey to purposely include missing data could improve the quality of the study (Littvay, 2009). A planned missing data design provides an effective way to reduce questionnaire length while maintaining all relevant questions of interest. Furthermore, missing values resulting from the planned missing approach are by design either Missing

Completely at Random (MCAR) or Missing at Random (MAR). As a result, we can use multiple imputation, maximum likelihood, or fully Bayesian approaches to handle the missing data just by focusing on the model for variables in the survey. In fact, since planned missing data designs reduce the burden on participants, the probability of nonresponse decreases, making it less likely to observe values that are missing not at random (MNAR), and, as a result, multiple imputation and maximum likelihood approaches are more likely to be valid (Rhemtulla & Little, 2012; Jorgensen et al., 2014; Kaplan & Su, 2016). Planned missing data approaches have frequently been used for educational assessment where students are evaluated on several subjects (Shoemaker & Shoemaker, 1981). Evaluating a student's proficiency on every subject would take a great deal of time, making it disruptive for students and unlikely to be approved by administrators. For this reason, many assessments utilize multiple matrix sampling, where the questionnaire is divided into several sets. Each student responds to a common set of items, which are shared among all participants, as well as a randomly selected subset of remaining questions. The Kentucky Instructional Results Information System, the Massachusetts Comprehensive Assessment System, the National Assessment of Educational Progress, and the Dutch National Assessment Program have all used matrix sampling to reduce the testing burden on students (Childs & Jaciw, 2003).

Split questionnaire design, as proposed by Raghunathan & Grizzle (1995), is an extension of multiple matrix sampling that places constraints on item assignment so that all two-way associations are estimable. Split questionnaire design divides the survey questions into multiple components and each participant responds to a fraction of the total components. One common variant, the 3-form split questionnaire, divides the survey into four components (X,A,B,C). Each participant responds to all items in X and two of the three other components, resulting in three unique survey forms, (X,A,B), (X,A,C) and (X,B,C), which are administered in equal proportions (Graham et al., 2006; Rhemtulla & Little, 2012). This particular 3-form design reduces the survey length by approximately 25%, but modifications can be made to both the number of total components used and fraction that each participant answers, depending on the survey composition and desired reduction in length. The split questionnaire is simple to implement and has been used in a number of studies (Graham et al., 2006).

In cross-sectional studies, the split questionnaire design was found to produce estimates similar to those obtained in the absence of missing data (Raghunathan & Grizzle, 1995; Littvay, 2009), but with a decrease in power. The loss of statistical power can be somewhat mitigated by the increased sample size obtainable due to the decrease in cost per participant

(Littvay, 2009). Peytchev & Peytcheva (2017) demonstrated that responses to questions from a split questionnaire design more closely resembled responses to questions at the beginning of a lengthy survey than at the end of the survey. They also demonstrated that correlations between variables attenuated when those variables were collected at the end of a survey. This research indicated that data collected from a split questionnaire design could be more useful than data collected at the end of a long survey. Until recently, little research had been done on the implementation of split questionnaires for longitudinal studies (Jorgensen et al., 2014). In this chapter, we consider several longitudinal designs for split questionnaires and compare their performance through results obtained from both simulations and data from the Health and Retirement Study (HRS).

3.2 Longitudinal Split Questionnaire Survey Design

The administration of split questionnaires in longitudinal surveys is more complex due to repeated variable measurements on the same subject at different time points. Usually, these repeated measurements are highly correlated (Hardt et al., 2012). Prior to designing a study, we need to determine whether to administer the same form to each participant throughout the entire study, or rotate the form each participant receives from year to year, or employ some combination of those two designs where we administer the same form to some participants and rotate the forms for other participants. Which design works best likely depends on the correlation structure of the data, more specifically, how the correlation between components measured at the same wave (within-wave correlation) compares to the correlation of measurements on the same variable over time (autocorrelation) and the correlation between two separate components measured at different waves (between-wave correlation). When the autocorrelation is greater than the within-wave correlation, we expect it would be preferable for most estimates to not measure the same components at each wave. We also think that, when most correlations are non-zero, more complex form rotations can better measure these different correlations and will produce more precise estimates. Which quantity is of primary interest to investigators could also influence which design should be administered. If we are primarily interested in estimating the change in a variable over time, then it might be preferable to administer the same components throughout the study.

We consider six different design options for allocating a 3–form questionnaire in longitudinal studies. Table 3.1 displays how each component A, B, and C would be allocated under the first five proposed designs in a three wave study, but each design could be easily modified

depending on the number of waves and desired reduction in survey length. For the first two options, the participants are placed into three groups. With Option 1, each group receives either form (A,B), (A,C), or (B,C) through out the study, while the forms each group receives are rotated in a manner that allows each form to be given in equal proportions at each wave for Option 2. For Option 3, we again administer to participants a different form at each wave, but we instead create six groups, which provides more ways of cycling the forms. We believe that the more complex design of Option 3 may allow us to better model variable correlations than Option 2 and will outperform it when most correlations are fairly strong.

Combining aspects of the first two options may produce more robust split questionnaire designs that perform well for more estimates and under more correlation structures. As such, Option 4 represents a combination of the first and second options, with a total of six groups, half get the same form throughout and the other half follow the rotation for Option 2. Option 5, the most complex of the planned study designs, contains nine groups, three of which receive the same forms as in Option 1 and the other six are rotated like in Option 3. The final design considered, Option 6, randomly assigns forms at each wave, making it simple to administer. This option allows every possible rotation to occur, which is beneficial when most correlations are non-zero, but makes the form design unbalanced. Our hypothesis is that the more complex designs would prove to be more robust, performing well under many types of correlation structures and for multiple estimands, but might not necessary be optimal under any scenario or for any estimate.

3.3 Analysis of Split Questionnaire Surveys

Due to the presence of missing data, maximum likelihood, fully Bayesian approaches, or multiple imputation are necessary for data analysis. We focus on results from data analysis using both maximum likelihood and multiple imputation in this chapter.

3.3.1 Multiple Imputation

The multiple imputation (MI) approach is particularly convenient for analysis done with split questionnaire surveys. In multiple imputation, we fill in missing values to create several complete data sets and perform standard statistical analysis on each data set. The estimated parameters and standard errors from each analysis are combined using Rubin's rules (Rubin, 1987) to obtain our final estimates. Often multiple imputation is done using a Sequential Regression Multiple Imputation (SRMI) framework where missing values are drawn using

Table 3.1: Form allocation for each design option.

Option	Wave 1	Wave 2	Wave 3
1	AB	AB	AB
	AC	AC	AC
	BC	BC	BC
2	AB	AC	BC
	AC	BC	AB
	BC	AB	AC
3	AB	AC	BC
	AB	BC	AC
	AC	BC	AB
	AC	AB	BC
	BC	AB	AC
	BC	AC	AB
4	AB	AB	AB
	AB	AC	BC
	AC	AC	AC
	AC	BC	AB
	BC	BC	BC
	BC	AB	AC
5	AB	AB	AB
	AB	AC	BC
	AB	BC	AC
	AC	AC	AC
	AC	BC	AB
	AC	AB	BC
	BC	BC	BC
	BC	AB	AC
	BC	AC	AB

Gibbs sampling from the posterior predictive distribution of a regression model, with each missing variable regressed on all other variables (Raghunathan et al., 2001).

Since imputation uses information from other variables to estimate missing values, having observed values highly correlated with missing values will better predict those missing values and improve imputations (Collins et al., 2001; Thomas et al., 2006; Hardt et al., 2012). Ideally, we would design split questionnaire surveys that take advantage of this attribute. For this reason, Raghunathan & Grizzle (1995) and Thomas et al. (2006) assign variables using correlations, where variables with high partial correlations are placed in different components. Thus, for most cross-sectional studies, we simply administer forms AB (X,A,B), AC (X,A,C), and BC (X,B,C) in equal proportions and focus our attention on how to place the variables

into each component; however, in a longitudinal study not all variables are collected at the same time, adding an extra dimension to consider.

3.3.2 Maximum Likelihood Estimation

Missing values resulting directly from the split questionnaire survey implementation are typically MAR by design. As a result, we can ignore the missing data mechanism and base our inference on only the observed data likelihood. For most instances, we can compute the maximum likelihood estimator (MLE) using an iterative method, such as Newton−Raphson or the EM algorithm, and asymptotic standard errors can be obtained by inverting the Fisher information matrix or using other methods (Little & Rubin, 2002).

In the case where complete data from each individual follows a p dimensional multivariate normal distribution with mean μ and variance Σ, we can obtain closed form solutions for the information matrix. Hartley & Hocking showed how to compute the Fisher information of the joint likelihood of $Y_1, ...Y_n$ for any arbitrary missing data pattern (Hartley & Hocking, 1971).

For each observation, Y_i, we can construct a $p \times p$ matrix such that each entry $a_{j,j} = 1$ if variable j was observed and 0 otherwise, where $1 \leq j \leq p$. We can then take d_i equal to this constructed matrix after all rows of zeros have been deleted, creating a $o_i \times p$ matrix, where o_i denotes the number of variables observed on subject i. Then each Y_i follows a multivariate normal distribution with mean $d_i\mu$ and variance $d_i\Sigma d_i^T$.

In the multivariate normal case, the information matrix is block diagonal, allowing us to compute and invert the information for μ and Σ separately. For μ, we can conveniently write the Fisher information for the total sample as

$$I_\mu = \sum_{i=1}^{n} d_i^T (d_i \Sigma d_i^T)^{-1} d_i. \tag{3.1}$$

Now, let σ_{jk} denote the element of Σ located at row j and column k. For the information matrix of Σ, it is convenient to create a vector of length $\binom{p+1}{2}$, δ, which contains all unique elements of Σ, or all σ_{jk} such that $j \leq k$. The information matrix based on δ is a $\binom{p+1}{2} \times \binom{p+1}{2}$ matrix. Let $\Sigma_i = d_i \Sigma d_i^T$ denote the variance matrix for subject i and let $I_{\Sigma i}$ denote the information matrix for Σ from subject i. The expected information for subject

i, corresponding to the negative partial derivative,

$$-E\left[\frac{\partial^2 logL}{\partial\sigma_{jk}\partial\sigma_{lm}}\right] = \frac{1}{2}tr(\Sigma_i^{-1}\Sigma_{ijk}\Sigma_i^{-1}\Sigma_{ilm}), \tag{3.2}$$

where Σ_{ijk} is a $o_i \times o_i$ matrix with entries corresponding to elements σ_{jk} and σ_{kj} in Σ_i equal to one and all other elements equal zero. Thus, $-E\left[\frac{\partial^2 logL}{\partial\sigma_{jk}\partial\sigma_{lm}}\right] = 0$ if σ_{jk} or σ_{lm} are not elements of Σ_i. From this formula, we can calculate the information matrix for each observation. Due to the additive property of the Fisher information, the total information for $Y_1, ...Y_n$ is equal to the sum of the information for each individual observation, hence

$$I_\Sigma = \sum_{i=1}^{n} I_{\Sigma i}, \tag{3.3}$$

where I_Σ is the Fisher information for Σ. We can then obtain the asymptotic variances for our MLE estimates of μ and Σ by inverting their information matrices, assuming that all elements of μ and Σ are estimable from our study.

3.4 Simulation and Analysis with Proposed Split Questionnaire Designs

3.4.1 Simulation Set Up

We performed simulations to examine how well each proposed design performs under a number of different correlation structures when data follow a multivariate distribution. We compared the performance of each design in estimating three primary quantities of interest: variable means, variance-covariance components, and the linear change in a variable's mean over time. The means and covariance are the sufficient statistics for the multivariate normal distribution, and, for longitudinal studies, change in a variable over time is likely one of the primary interests. We estimated our quantities of interest using both maximum likelihood estimation and multiple imputation.

For our data, we took three variables at each wave to represent the A, B, and C components of the split questionnaire, with the X component omitted for simplicity, using a sample size of 108. The nine variables came from a multivariate normal distribution with the variance-covariance structure displayed in Table 3.2. We used five parameters for the covariances, ρ_1 denotes the within-wave correlation, ρ_2 and ρ_3 are the autocorrelations, and

ρ_4 and ρ_5 represent the correlation between two separate components measured at different waves. The means of each variable changed linearly over time.

By varying the values of ρ_1, ρ_2, ρ_3, ρ_4, and ρ_5, we can create numerous correlation structures. Table 3.3 displays the correlation values we used to test the performance of the different options. For the first three correlation structures, only one of the within-wave, autocorrelation, or between-wave correlations is non-zero, enabling easy comparisons between the performance of each design under extremely different conditions. These correlations are rather unrealistic for repeated measures data, especially structure 3 where the between-wave correlation is the only non-zero correlation. They were chosen because they allow us to test the performance of each design at the boundary conditions. The last three structures more accurately reflect correlations for longitudinal data, where within-wave and autocorrelations are greater than between-wave correlations. In structure 4, within-wave correlation is largest, while for structure 5 autocorrelation is greatest. Structure 6 follows an autoregressive structure, where correlations decrease over time, with two main correlations, the correlation between different variables and the autocorrelation. In addition, we tested the performance of each design under a random correlation structure, with the correlation drawn from a Wishart distribution. This allows us to test which design will perform best on average across all possible correlations.

Table 3.2: Structure of the variance-covariance matrix used in simulations.

		Wave 1			Wave 2			Wave 3		
		A	B	C	A	B	C	A	B	C
Wave 1	A	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4	ρ_3	ρ_5	ρ_5
	B	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4	ρ_5	ρ_3	ρ_5
	C	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2	ρ_5	ρ_5	ρ_3
Wave 2	A	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4
	B	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4
	C	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2
Wave 3	A	ρ_3	ρ_5	ρ_5	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1
	B	ρ_5	ρ_3	ρ_5	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1
	C	ρ_5	ρ_5	ρ_3	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1

We then compared performance of each design option under the proposed correlation structures by computing the variance for the variable means, variance-covariance components, and the linear change in means over time using both MLE and mutliple imputation. For maximum likelihood estimation, we calculated the variance of the means, and variance-covariance components from inverting the Fisher Information using the true underlying co-

Table 3.3: Different correlation values used in simulations.

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
Structure 1	0.50	0.00	0.00	0.00	0.00
Structure 2	0.00	0.50	0.50	0.00	0.00
Structure 3	0.00	0.00	0.00	0.50	0.50
Structure 4	0.80	0.50	0.50	0.40	0.40
Structure 5	0.50	0.70	0.70	0.30	0.30
Structure 6	0.50	0.70	0.49	0.25	0.125

variance matrix. We estimated the linear change in means over time for a variable using contrasts. The variance of these linear combinations of means from the contrasts can be computed from the variances obtained by inverting the Fisher Information. For MLE, the variances of our estimates of interest did not require data simulation to compute; however, we used many iterations to assess the overall performance of Option 6, the random form assignment.

For multiple imputation, we simulated complete data from the multivariate normal distribution. Values were then set to missing so observed data matched what would have been obtained under each study design option and we performed multiple imputation. We then estimated the mean, variance-covariance, and linear change in mean over time. We performed a log transformation on the sample variance for each variance component and a Fisher z-transformation for the correlations to make the estimates approximately follow an normal distribution in order to apply Rubin's rules. We used a linear mixed model of the variable regressed on time for estimating the change in mean and stored the parameter estimates and standard errors for the slope and intercept.

Due to the transformations needed to normalize the distributions for the variance-covariance components and the use of repeated measures regression for multiple imputation, the estimated variances for MLE and MI are not directly comparable; however, the relative performance, rankings, and conclusions are largely the same when using either MLE or MI. We mostly focus on simulation results using multiple imputation, as MI will more likely be used in practice. The simulation results for MI can be found in 3.4.2 and the results for MLE can be found in section 3.4.3.

3.4.2 Simulation Results fo Multiple Imputation

Table 3.4 displays the average percent increase in variance for the mean and variance-covariance components from using the proposed split questionnaire design versus complete

data for correlation structures 1, 2, and 3 and the random correlation. Since relative rankings for the mean and variance-covariance were very similar, we combined them into one category in the table. Similarly, table 3.5 displays the variance increase for repeated measure regression using the same correlation structures. All of the design options performed very similarly for structure 1, where only variables within the same wave are correlated. The longitudinal selection of split question forms did not matter very much due to the lack of correlation across waves, though, Option 5 and Option 6 did perform slightly better at estimating variable correlations due to the additional form rotations, even if most correlations were zero.

Results from structure 2, where only the autocorrelations are non-zero, show that Option 2 and Option 3 performed best in terms of estimating variable means and variance-covariance components while Option 1 performed the worst, but the opposite occurred for estimating the change in mean over time. In this scenario, a variable is predictive of its values measured at different time points for the same subject. Since Option 2 and Option 3 rotate split questionnaire forms for all individuals, we measure each variable on a subject during the study, which enables us to better estimate missing values for that variable and leads to a better estimate for the mean and variance. On the other hand, Option 1 measures the same variables on an individual for every wave, which provides a larger sample of individuals with a variable measured at all time points than the other options, allowing a better estimate for how a variable changes over time.

For the unusual correlation structure 3, where a variable is only correlated with other variables measured at different waves, Option 6, the random form assignment, performed the best overall by far. Option 5 was the second best under this correlation while Option 1 and Option 2 performed terribly, especially Option 2. This indicates that extra form rotations and more complex designs are beneficial when between-wave correlations are large. Finally, for the random variable correlation, Option 5 and Option 6 again performed the best and Option 1 and Option 2 did the worst. Based on this, it appears that the more complex designs perform better across all possible variable correlations; however, all correlation structures are not equally likely for longitudinal studies.

After testing each design option at the boundary conditions, we examined the capability of each design across the more realistic correlation structures 4, 5, and 6. Table 3.6 displays the average percent increase in variance over complete data for those three structures, broken down by mean, variance, covariance, and repeated measures change in mean. We averaged results across all three correlations, since the relative performance of the designs were similar

Table 3.4: Average percent increase in variance from complete data for mean and variance-covariance components using MI.

	Structure 1	Structure 2	Structure 3	Random Structure
Option 1	50.95	82.62	57.41	70.59
Option 2	50.68	51.72	188.49	72.34
Option 3	48.13	47.75	32.18	62.84
Option 4	48.68	59.19	18.57	61.84
Option 5	46.30	53.87	17.77	57.53
Option 6	47.47	52.73	6.52	59.87

Table 3.5: Average percent increase in variance from complete data for repeated measures regression change in mean over time using MI.

	Structure 1	Structure 2	Structure 3	Random Structure
Option 1	52.44	74.92	83.02	45.23
Option 2	51.24	100.15	149.79	45.21
Option 3	52.81	99.15	32.53	36.48
Option 4	51.81	85.07	25.92	36.97
Option 5	52.85	88.72	21.63	33.95
Option 6	53.18	92.86	7.25	33.70

for each correlation. Option 3 performed the best in terms of estimating the mean, variance, and covariance, while Option 5 and Option 6 were close behind. For these structures, within-wave and autocorrelation were greater than between-wave correlation, which is likely why Option 3 outperformed Option 5 and Option 6 in estimating the sufficient statistics. Option 3 likely outperformed Option 2 because the extra rotations allowed us to better estimate the between-wave correlations. Option 1 was by far the worst at estimating the mean and variance-covariance, but was the best at estimating the change in mean over time. Option 4 and 5 were fairly close to Option 1 for repeated measures, probably because both of those options measure the same variables at each wave on a subset of study participants. Those two options are also quite a bit better at estimating the sufficient statistics than Option 1.

3.4.3 Simulation Results for MLE

Tables 3.7 and 3.8, analogous to tables 3.4 and 3.5 from multiple imputation analysis, display the average percent increase in variance for the mean, variance-covariance components, and linear change over using complete data for correlation structures 1, 2, and 3 and the random correlation using MLE to compute variance. Once again, design options performed very similarly for structure 1, but Option 5 and Option 6 performed slightly better in estimation

Table 3.6: Average percent increase in variance from complete data for means, variance-covariance components, and repeated measures regression parameters averaged over correlation structures 4, 5, and 6 using MI.

	Mean	Variance	Covariance	Repeated Measures
Option 1	20.03	50.83	11.02	38.07
Option 2	13.23	39.70	2.63	59.34
Option 3	10.75	33.05	1.87	48.60
Option 4	13.53	37.35	3.18	38.62
Option 5	11.74	33.51	2.27	39.22
Option 6	11.18	33.30	2.05	45.89

for variance-covariance components.

For structure 2, Option 3 again had the most precision for estimating the variable means and variance-covariance components; however, both Option 5 and Option 6 performed slightly better than Option 2 for MLE estimation due to a small improvement in estimating the covariance components, unlike in the MI case. Option 1 again performed worst in estimating the mean and variance-covariance. Under structure 2, the performance order for the change in mean over time was identical to the results from MI, with Option 1 performing the best and Option 2 and Option 3 doing the worst.

For correlation structure 3, Option 6 again does best with all estimates, while Option 4 performed the second best instead of Option 5, which was third best overall. The rankings of Option 4 and Option 5 are switched from MI. In addition, Option 1 performed slightly better than Option 3 using MLE, but Option 3 was superior for MI. Option 2 was again the worst under this correlation structure. Finally, the relative performance under the random correlations was consistent with what we saw for multiple imputation, with the more complex designs performing better.

Table 3.7: Average percent increase in variance from complete data for mean and variance-covariance components using MLE.

	Structure 1	Structure 2	Structure 3	Random Structure
Option 1	80.79	120.59	38.50	65.13
Option 2	80.79	80.79	53.93	65.45
Option 3	72.87	72.87	46.76	39.89
Option 4	72.87	85.36	15.07	39.89
Option 5	70.48	77.62	25.33	34.79
Option 6	71.87	76.34	4.17	34.17

Table 3.9, analogous to table 3.6, displays the average percent increase in variance over

Table 3.8: Average percent increase in variance from complete data for change in mean over time using MLE.

	Structure 1	Structure 2	Structure 3	Random Structure
Option 1	35.00	50.00	25.00	23.79
Option 2	35.00	80.00	33.33	27.58
Option 3	35.00	80.00	25.00	22.15
Option 4	35.00	63.64	12.50	20.69
Option 5	35.00	68.75	15.39	19.66
Option 6	35.49	73.65	4.66	19.66

complete data for correlation structures 4, 5, and 6 using MLE. Option 3 again performed best in terms of estimating the mean, variance, and covariance, with Options 6 close behind, followed by Option 5, with Option 1 doing the worst. For the change in means over time, the rankings for each option is identical to the multiple imputation case, with Option 1 performing the best, followed by Option 4 and Option 5.

Table 3.9: Average percent increase in variance from complete data for means, variance-covariance components, and change in means over time averaged across correlation structures 4, 5, and 6 using MLE.

	Mean	Variance	Covariance	Repeated Measures
Option 1	27.66	38.88	67.99	20.01
Option 2	15.10	26.70	35.94	39.63
Option 3	12.12	21.71	22.11	31.59
Option 4	17.20	26.79	33.05	25.58
Option 5	14.63	23.63	26.73	26.98
Option 6	12.89	21.86	22.79	31.00

3.4.4 Conclusions from Simulation Results

From the analysis, we see that the optimal design for a longitudinal study with planned missingness depends on the structure of the data and the estimates of interest. If the variables have high autocorrelations, we prefer Option 2 or Option 3 where participants change forms every year for estimating the mean and variance-covariance parameters, but Option 1 would be better for estimating the change in mean over time. If the data follow a more unusual structure, it would be better to use Option 5 or Option 6 which are more complex, but the added form rotation better allows them to estimate covariance parameters and makes them more robust to the correlation structure. For more realistic correlations, Option 3 appears to perform best for estimating the mean and variance-covariance and Option 1 is best at

estimating the change in mean. Option 4 and Option 5 can measure the mean and variance-covariance, as well as the change in mean, fairly well. While our simulations are useful for evaluating how each option performs, our data structures are more simplistic than what we would observe in practice. Next, we examine which planned missing design option works best using data from a longitudinal survey.

3.5 Analysis of HRS Data

The Health and Retirement Study (HRS), beginning in 1992, is an ongoing longitudinal study of U.S. adults age 50 and older (Juster & Suzman, 1995). HRS collects information pertaining to the health, income, and job status of participants. New samples of individuals enter the study after turning 50, while some older participants exit due to death or loss to follow up. For our purposes, we selected seven components from the survey data collected in the 2004, 2006, 2008, and 2010 waves of the study. The first component (X) contains basic demographic information (age, gender, race, height, and education) and health behavior information (smoking status and drinking history). The next five components contain questions related to health: diabetes (D), hypertension and blood pressure (B), heart disease and stroke (H), cancer (C), and weight (W). The final component contains information regarding income and wealth (I). We consider all participants with no missing values in each of the seven modules for all four study waves as our complete data (ideal), a total of 3059 subjects. The survey questions from the X module were taken from baseline only. We then set values to missing, mimicking the data pattern that would have been observed had we used a planned missing data design. Table 3.10 displays the five structured design options we used, corresponding to a similar design option from the simulations. In addition to the designs listed in the table, we included a random form assignment at each wave, Option 6.

We performed multiple imputation on the missing values using IVEWare (Raghunathan et al., 2002). To assess the validity of the imputations, we examined the marginal distribution of the imputed values compared to the observed values and checked for any major discrepancies between the two. Since data are MCAR by design, we would not expect the marginal distribution of imputed values to differ from the observed values. Furthermore, we performed diagnostics on the regression models through goodness of fit testing and diagnostic plots to examine the validity of our imputation models. For linear regression models, we plotted the residuals versus the fitted values, shown in figure 3.1. If the imputation model was correctly specified, we would expect to see a random scatter centered around the X-axis, with no

discernible difference between observed and imputed values, like in figure 3.1.

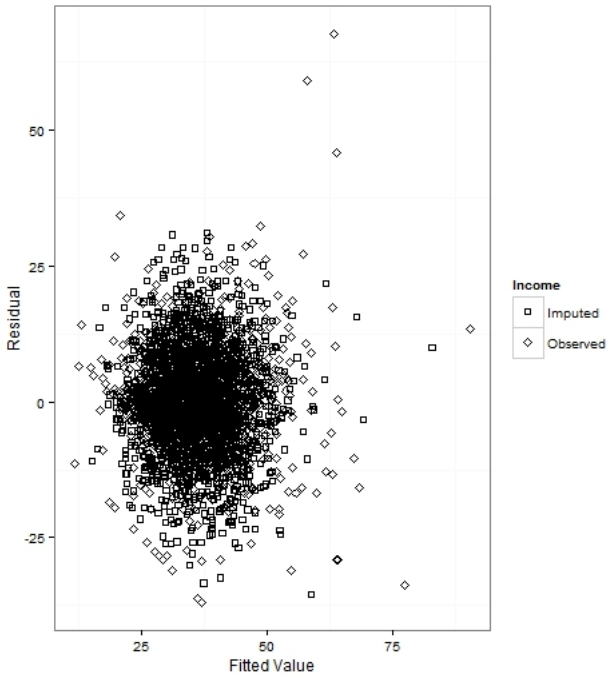

Figure 3.1: Imputation diagnostic plot of residual versus fitted values.

3.5.1 Univariate Analysis of HRS Variables

After performing multiple imputation, we analyzed the mean, median, and quartiles of the continuous variables under the different options. We estimated π_j, the probability that variable j is equal to one, for categorical variables. For each population parameter, we took the difference between the estimates under the planned missing designs and the estimates from the complete data. The differences for the mean and quantile estimates were standardized by dividing by the complete data standard deviation of the variables, while the difference in π_j was standardized by dividing by $\pi_j(1-\pi_j)$ under complete data. Figure 3.2 plots the distribution of the differences under the six options. Option 3 produced the best estimate for univariate parameters with the bias distributed tightly around zero, followed closely by Option 6 and Option 2, while Option 1 performed the worst. Most of the bias in Option 1 came from poor estimation of the quartiles for continuous variables. Figure 3.3 displays the ratio of the standard error of the mean estimates for each option to the standard error from the complete data. Once again, Option 3 performed best, generally producing smaller standard errors. Most of the other designs performed fairly similarly, but Option 1 and

Option 4 performed considerably worse than the others.

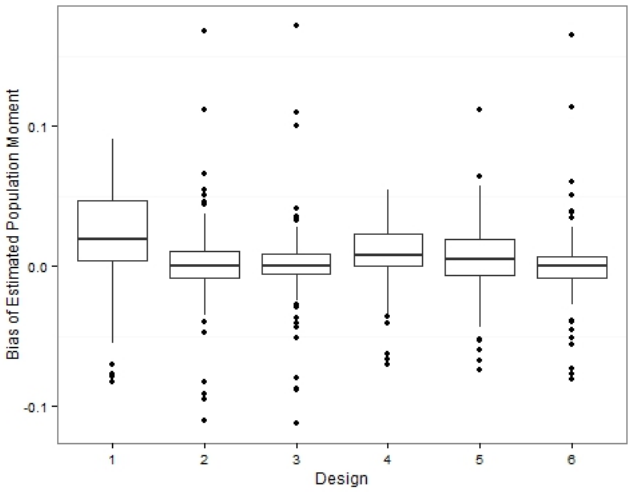

Figure 3.2: Distribution of standardized bias for univariate estimates (a single outlier was removed for Option 1).

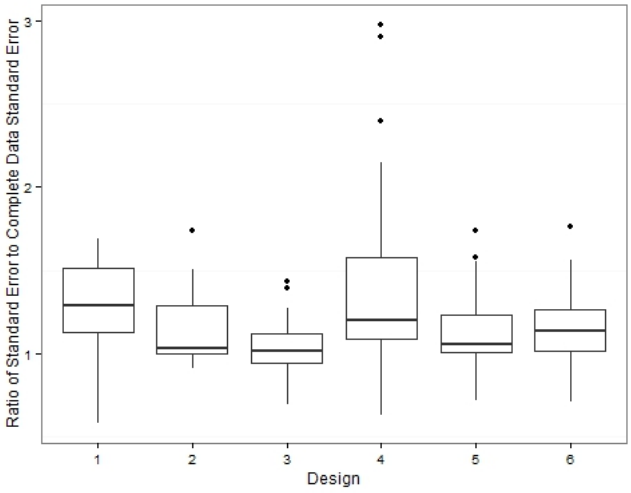

Figure 3.3: Ratio of standard errors for the estimated variable means under each design to complete data standard errors.

3.5.2 Regression Analysis from Previous Publications on HRS

We also evaluated the performance of each option using three different regression analyses based on three previous publications with HRS data: a paper by Bowen (2010), a paper by Best et al. (2005), and a paper by Avendano & Glymour (2008).

The paper by Bowen (2010) studied associations with coronary heart disease (CHD). For our model, CHD was the outcome variable, and, using logistic regression, we examined how hypertension, diabetes, obesity, income, wealth, age, years of schooling, and being born in the United States affect CHD. The variable CHD was an indicator for whether a subject was ever told by a doctor that they had coronary heart disease. We specified a subject level random intercept to account for repeated measures. Table 3.11 shows the parameter estimates and standard errors obtained from the complete data (CD) and six design options. Hypertension and Diabetes are indicator variables for whether the subject was ever diagnosed with either condition by a doctor. Obese and US Born are dummy variables indicating if the subject had a BMI greater than 30 or was born in the United States, respectively. Income measures the amount in U.S. dollars that the participant and his/her spouse earned for the year, while Wealth measures their total net worth in dollars. Both Income and Wealth were transformed by taking the cube root divided by 1,000 to diminish the influence of outliers in the regression model and make coefficients easier to distinguish between designs. Age and School Yrs are the subject's age during the 2004 wave and years of school. Option 1 performed the worst out of all options, producing the worst estimates for the effects of hypertension, diabetes, age, years of schooling, and being born in the US on the effects of CHD. Option 4 was the second worst in terms of parameter estimation, while Option 3 produced the closest parameter estimates and performed the best overall, though Option 2 and Option 6 also performed fairly well.

Our next model, similar to the Best et al. (2005) paper, used a mixed logistic regression model with a random intercept to examine the effect that age, race, gender, weight, income, wealth, and years of schooling have on diabetes. The results of this regression model are shown in table 3.12. Black and Female are indicators for whether the subject is of African ancestry and female, respectively. An Age and Age^2 term were included in the model. We again used the cube root of Income and Wealth divided by 1,000. Weight represented the subject's weight in units of ten pounds. In this regression model, Option 1 again performed the worst, followed by Option 5. Options 2 and 3 achieved the best overall parameter estimates for this model, producing similar estimates for Age, Age^2, Black, Female, and Wealth. Option 6 also performed fairly well.

Our final model was based on the paper by Avendano & Glymour (2008), which used survival analysis to study the association of income, wealth, and education on the incidence of stroke, stratified by age. Since our data cover a smaller time frame and there were very few new stroke cases from 2004 to 2010, we instead used a mixed logistic regression model

to measure the association between stroke and age, income, wealth, and years of schooling. Table 3.13 shows the results, we should note that the prevalence of stroke was around 3% and analysis is based on a small number of cases. Once again, we took the cube root of Income and Wealth divided by 1,000. In this instance, Option 1 was the worst at estimating the intercept and the effect of age, but did comparatively well at estimating income and wealth effects. Option 5 performed the best, producing the best estimates for Age, Income, Wealth, but was one the worst for estimating School Years. Option 2 and Option 4 performed similarly in this instance. Option 6 performed the worst for this regression model. Surprisingly, both Income and Wealth were only moderately correlated with themselves across waves, having a correlation between 0.3 and 0.6 in most cases, in contrast to hypertension, diabetes, and heart disease, which had correlations greater than 0.8 across waves. This may explain why Option 3 did not perform the best overall in this instance and why all six options were fairly similar in terms of performance.

After recreating the analysis from the three previously published papers, we combined the parameter estimates and standard errors across the different models to determine how the options performed overall. Figure 3.4 shows how each option performed in terms of parameter estimation bias across the three regression models, by taking the difference in parameter estimates under the design option and the complete data and dividing by the complete data standard error. Option 3 performed the best at estimating regression coefficients while Option 1 performed the worst, though it was only slightly worse than Option 4. Option 2 was second best, followed by Option 5. In Figure 3.5, we display the ratio of the MI standard errors under each option to the complete data standard errors. Ratios greater than 2 were removed from the plot, removing a single point from options 1, 2, 3, 4, and 6, while two points were removed from Option 5. From the figure, we see that Option 3 produced the smallest standard errors, followed by Option 2 and Option 6. The other three options performed fairly similar to each other.

3.5.3 Conclusions from HRS Data

Overall, Option 3 performed the best with the HRS data, followed by Option 2, while Option 1 performed the worst in terms of mean and regression analysis. The strong performance of Option 2 and Option 3 in this instance is not surprising since the autocorrelation was quite large for most variables, and we saw from simulations that Option 2 and Option 3 performed better in terms of mean and variance estimation when the autocorrelation is much greater than the other correlations. Had the within-wave and between-wave correlations been

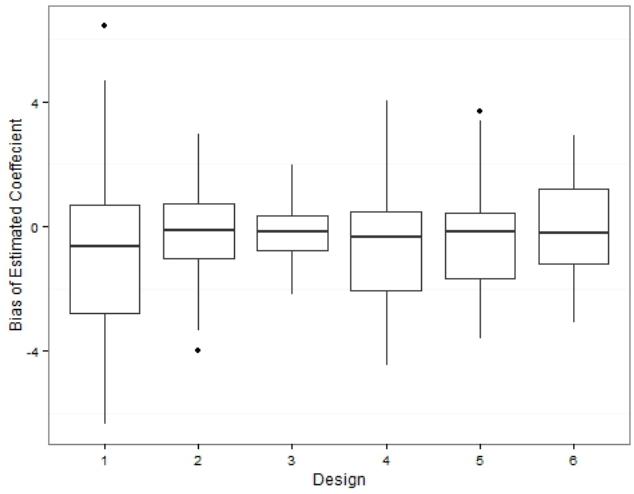

Figure 3.4: Distribution of regression parameter estimation bias.

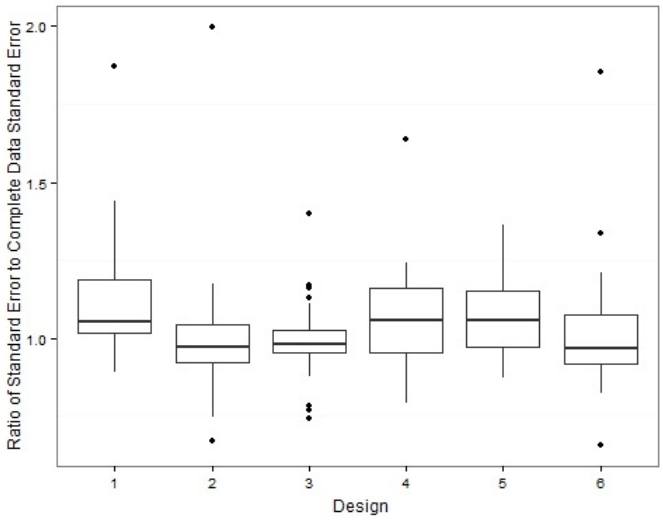

Figure 3.5: Distribution of the ratio of regression parameter standard errors under each design to complete data standard errors.

stronger, Option 5 and Option 6 would likely have performed better.

3.6 Discussion

The use of a split questionnaire in a survey can reduce the burden and fatigue on participants, leading to an increase in the quality of the data and a reduction in unplanned missing values. This could also reduce the drop out rate in longitudinal studies. There is a good deal of literature on implementing planned missing designs in cross-sectional studies, commonly

using a 3-form design, but, until recently, there has been little research on implementing planned missing data designs in longitudinal studies. One exception was Jorgensen et al. (2014), which considered multiple longitudinal assignment methods for 3-form survey designs in a latent variable setting.

In this chapter, we examined six different design options for allocating forms in a longitudinal study. We saw from our simulation study that the performance of each design option depended on the correlation structure of the data and which estimates were of interest. We later tested these designs on survey data from the HRS. From the simulations and HRS data, it seems that if investigators are primarily concerned with estimating the mean or regression parameters for a variable regressed on a subset of study variables, then Option 3 will likely perform the best due to high autocorrelations. Option 2 performed similarly to Option 3 when autocorrelations were very strong, but was usually slightly worse. Option 2 did a lot worse when autocorrelations were weak. One reason to use Option 2 over Option 3 is its simplicity to implement. If interest lied in the change in a variable over time, Option 1 would be preferable in most instances. Option 4 generally performed fairly similar to Option 1 in terms of estimating change in mean over time, but performed better in estimating the mean and variance-covariance.

Prior information regarding the data structure would help in determining the optimal design option for a study and can help in the imputation. We used seven modules from the HRS study, and our choice for dividing the questions was based on simplicity and not based on any method for determining an optimal design. Several papers have considered dividing the survey questions into forms based on correlation structures or other methods to improve estimation, but require prior information on the variables (Raghunathan & Grizzle, 1995; Adigüzel & Wedel, 2008). Using these methods to create the form design could affect which option performs best. We should also beware that our results when using planned missing designs will be biased if we do not specify a correct imputation model, which by no means is a trivial matter.

Table 3.10: Longitudinal design options for the Health and Retirement Study.

	Wave 1	Wave 2	Wave 3	Wave 4
Option 1	DBH	DBH	DBH	DBH
	DWC	DWC	DWC	DWC
	IWH	IWH	IWH	IWH
	IBC	IBC	IBC	IBC
Option 2	DBH	IWH	DWC	IBC
	DWC	DBH	IBC	IWH
	IWH	IBC	DBH	DWC
	IBC	DWC	IWH	DBH
Option 3	DBH	IWH	DWC	IBC
		DWC	IBC	IWH
	DWC	DBH	IBC	IWH
		IBC	IWH	DBH
	IWH	IBC	DBH	DWC
		DBH	DWC	IBC
	IBC	DWC	IWH	DBH
		IWH	DBH	DWC
Option 4	DBH	DBH	DBH	DBH
		IWH	DWC	IBC
	DWC	DWC	DWC	DWC
		DBH	IBC	IWH
	IWH	IWH	IWH	IWH
		IBC	DBH	DWC
	IBC	IBC	IBC	IBC
		DWC	IWH	DBH
Option 5	DBH	DBH	DBH	DBH
		IWH	DWC	IBC
		DWC	IBC	IWH
		IBC	IWH	DWC
	DWC	DWC	DWC	DWC
		DBH	IBC	IWH
		IBC	IWH	DBH
		IWH	DBH	IBC
	IWH	IWH	IWH	IWH
		IBC	DBH	DWC
		DBH	DWC	IBC
		DWC	IBC	DBH
	IBC	IBC	IBC	IBC
		DWC	IWH	DBH
		IWH	DBH	DWC
		DBH	DWC	IWH

Table 3.11: Parameter estimates and standard errors for regression 1: coronary heart disease.

	Parameter	CD	Opt 1	Opt 2	Opt 3	Opt 4	Opt 5	Opt 6
Estimate	Intercept	-5.142	-2.616	-3.981	-4.373	-3.565	-4.078	-4.003
	Hypertension	0.474	0.124	0.486	0.462	0.272	0.291	0.376
	Diabetes	0.513	0.187	0.371	0.449	0.223	0.394	0.438
	Obesity	0.052	-0.006	-0.004	0.009	-0.008	- 0.014	0.055
	Income*	0.772	-1.091	-6.281	-2.184	-1.509	-2.396	1.476
	Wealth*	-0.933	-0.671	0.307	-0.317	-1.164	-0.975	0.210
	Age	0.058	0.028	0.043	0.047	0.039	0.046	0.044
	School Yrs	-0.002	0.011	0.003	0.004	-0.003	0.001	-0.016
	US Born	-0.100	-0.033	-0.081	-0.087	-0.083	-0.095	-0.087
Standard Error	Intercept	0.392	0.396	0.360	0.369	0.439	0.366	0.356
	Hypertension	0.085	0.075	0.082	0.084	0.076	0.087	0.070
	Diabetes	0.106	0.117	0.081	0.092	0.096	0.100	0.097
	Obesity	0.054	0.051	0.036	0.040	0.044	0.047	0.058
	Income*	1.773	2.264	2.018	1.968	2.066	1.947	2.082
	Wealth*	0.648	0.641	0.682	0.655	0.678	0.802	0.568
	Age	0.005	0.005	0.004	0.005	0.005	0.005	0.004
	School Yrs	0.014	0.015	0.014	0.013	0.014	0.014	0.014
	US Born	0.042	0.044	0.039	0.040	0.044	0.043	0.041

* indicates variable transformation through diving the cube root by 100

Table 3.12: Parameter estimates and standard errors for regression 2: diabetes.

	Parameter	CD	Opt 1	Opt 2	Opt 3	Opt 4	Opt 5	Opt 6
Estimate	Intercept	-21.712	-8.383	-18.662	-17.842	-14.565	-11.259	-17.918
	Age	0.582	0.203	0.498	0.475	0.383	0.279	0.477
	Age^2	-0.004	-0.001	-0.003	-0.003	-0.003	-0.002	-0.003
	Black	0.861	0.467	0.735	0.759	0.637	0.578	0.694
	Female	-0.179	-0.139	-0.158	-0.149	-0.119	-0.070	-0.200
	Income*	1.626	-1.324	-1.730	0.519	-1.389	-1.554	-1.490
	Wealth*	-0.197	-0.385	-0.268	-0.230	-1.038	-0.485	-0.616
	Weight	0.012	0.001	0.005	0.004	0.005	0.006	0.000
	School Yrs	-0.071	-0.042	-0.060	-0.073	-0.067	-0.051	-0.048
Standard Error	Intercept	2.842	2.892	2.778	2.753	3.298	2.727	2.605
	Age	0.085	0.087	0.083	0.084	0.099	0.086	0.078
	Age^2	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	0.0006
	Black	0.135	0.144	0.131	0.132	0.132	0.143	0.133
	Female	0.106	0.117	0.102	0.103	0.100	0.122	0.102
	Income*	1.006	1.880	2.010	1.409	1.646	2.724	1.866
	Wealth*	0.189	0.452	0.609	0.586	0.516	0.537	0.492
	Weight	0.004	0.004	0.003	0.004	0.003	0.005	0.005
	School Yrs	0.017	0.018	0.017	0.017	0.018	0.018	0.017

* indicates variable transformation through diving the cube root by 100

Table 3.13: Parameter estimates and standard errors for regression 3: stroke

	Parameter	CD	Opt 1	Opt 2	Opt 3	Opt 4	Opt 5	Opt 6
Estimate	Intercept	-6.222	-5.581	-6.521	-6.392	-6.774	-6.693	-6.195
	Age	0.059	0.043	0.057	0.056	0.060	0.058	0.052
	Income*	-14.54	-10.47	-8.23	-4.84	-9.07	-12.36	-5.31
	Wealth*	-4.902	-6.397	-2.185	-4.657	-2.528	-4.914	-0.680
	School Yrs	-0.044	-0.058	-0.048	-0.057	-0.044	-0.027	-0.070
Standard Error	Intercept	0.775	0.948	0.911	0.899	0.962	0.953	0.899
	Age	0.009	0.011	0.010	0.010	0.011	0.011	0.011
	Income*	7.425	10.680	5.564	5.816	5.898	10.110	7.196
	Wealth*	2.266	2.751	1.866	1.741	2.537	2.429	1.488
	School Yrs	0.032	0.033	0.033	0.031	0.032	0.033	0.034

* indicates variable transformation through diving the cube root by 100

Chapter 4

Optimal Variable Allocation for Split Question Designs

4.1 Introduction

Split questionnaire designs, where participants only respond to a subset of total questions, are used to reduce survey questionnaire length and lower the burden and fatigue placed on participants without reducing the total number of variables measured in a survey (Raghunathan & Grizzle, 1995). The reduction in burden and fatigue should improve the quality of responses given in surveys. There is a good deal of literature regarding the use and performance of split questionnaire designs (Raghunathan & Grizzle, 1995; Graham et al., 2006; Littvay, 2009; Rhemtulla & Little, 2012; Peytchev & Peytcheva, 2017), and details on how to create and reasons to use a split questionnaire have been discussed in chapters 1 and 3 of this paper.

In the previous chapter, we examined the performance of several design options for administering a split questionnaire in the longitudinal setting. For our analysis, we fixed the split questionnaire variable assignments beforehand, using the same variable allocation at each time point, and examined how to best administer the questionnaire longitudinally by rotating the forms. We did not consider how to assign variables to splits within the cross–sectional split questionnaire forms.

For this chapter, we revisit split questionnaire design to focus on variable allocation within splits. To do this, we first must establish a criterion for comparing the performance of proposed split questionnaire designs, with the goal of minimizing the impact of missing data on our final analysis. There is less literature on how to assign variables into splits.

Raghunathan & Grizzle (1995) discussed using partial correlation coefficients for allocating variables. They proposed placing variables with high partial correlation into separate splits. Our simulations from chapter 3 also indicated the importance of using variable correlations in designing optimal splits. Generally, we would like observed variables to be highly

predictive of missing values to minimize the impact of missing data. While this information is useful and can be used as a guideline, it still does not establish an objective criterion for comparing proposed split questionnaire designs.

Thomas et al. (2006) developed an algorithm for creating split questionnaires based on improving the variance of the estimated marginal means of each variable. This is a useful design method when the population means are of primary interest; however, for many settings the main interest lies in estimating the relationship between variables. Adigüzel & Wedel (2008) proposed minimizing the Kullback-Leibler (KL) divergence between the complete data likelihood obtained when there is no missing data and the observed data likelihood obtained using a split questionnaire design as a method for determining the optimal variable allocation. The KL divergence provides a measure of discrepancy between two probability distributions and a smaller divergence would indicate less loss of information from the missing data. This method accounts for the full parameter space when determining the optimal design, not just the variable means.

All the described methods at the very least require preliminary estimates of the joint variable correlations, if not all parameters, in order to implement. Adigüzel & Wedel (2008) recommended to first conduct a pilot study to obtain estimates for the likelihood parameters and then treat those estimates as the true parameter values to determine the KL divergence under the proposed splits. Preliminary estimates for some parameters may also be obtained from previous publications.

There is no objective way to determine which split questionnaire design is going to perform best without some knowledge of the parameter values, but taking the preliminary point estimates for the true parameter values ignores any variance associated with those estimates. There may be more information available for some parameter estimates than for others, in which case it is probably beneficial to design a split questionnaire that allows for better estimation of parameters with more uncertainty. A Bayesian framework will better allow us to incorporate prior information and uncertainty through the prior distribution, instead of using point estimates, and can account for differing levels of uncertainty in each parameter.

An informative prior is necessary in this instance as uninformative priors provide no help in determining the best split questionnaire. We will need to use previously collected data to construct our prior distributions, since it makes little sense to use any method for constructing an optimal split questionnaire design without having estimates or prior distributions based on actual data. We would most likely use data from a single pilot study to construct our prior, though we could incorporate several sources with data on some of

the variables being used in our survey, including previous publications. The use of a prior distribution can incorporate differing levels of uncertainty in our parameters, such as when certain parameters have been measured on a larger sample size. When using a pilot study, we would use the posterior distribution from our pilot study as our prior distribution.

We propose using a Bayesian framework to determine the optimal variable allocation in a split questionnaire design by minimizing the KL divergence between the complete data and the observed data posterior distributions. In section 4.2, we discuss our method for determining the optimal split questionnaire design using the KL divergence. Section 4.3 focuses on methods for computing the KL divergence between posterior distributions and provides specific examples. Section 4.4 describes how to use the KL divergence to select split questionnaire designs through a search algorithm. We present simulation results from using our method in section 4.5. In section 4.6, we revisit the HRS data and longitudinal designs from chapter 3 using the methods presented in this chapter. Finally, in section 4.7, we provide our conclusions and discussion.

4.2 Determining Optimal Split Questionnaire Designs

4.2.1 Kullback-Leibler Divergence

For the purpose of creating the best possible split questionnaire design, we wish to determine the variable allocation that would result in the smallest loss of information from missing data or make our inference from the split questionnaire as close to what it would have been had we used the full questionnaire. In order to find the best variable allocation, we require some a priori information on the variables being used in the survey. As a result, we propose using a Bayesian framework for determining the optimal allocation to better account for uncertainty in our prior knowledge. Within this framework, all inference will be based on the posterior distribution. Our goal then is to minimize the difference or loss of information from the posterior distribution under the split questionnaire design compared to the posterior distribution under the full questionnaire. In order to minimize this difference, we first need a metric that measures the discrepancy between the split questionnaire and the full questionnaire posterior distribution for any given split questionnaire design. We can then apply this metric to every potential split questionnaire design and use it to determine which design achieves the smallest loss in information.

For a metric to determine the difference in the two posterior distributions, we turn to the Kullback-Leibler (KL) divergence. The Kullback-Leibler divergence first described by

Kullback & Leibler (1951) is a measure of discrepancy between two probability distributions. The KL divergence from distribution Q to distribution P for two continuous distributions is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx.$$

The KL divergence is always greater than or equal to zero, and the KL divergence equaling zero implies that P = Q almost surely. In addition, $D_{KL}(P||Q) < D_{KL}(P||R)$ implies that Q is closer in distribution to P than R is to P. Although it is not a true distance metric due to the lack of symmetry in the KL divergence from Q to P and the KL divergence from P to Q, it is still a useful quantity for measuring the discrepancy of distributions. The KL divergence is often used in information theory and $D_{KL}(P||Q)$ can be thought of as a measurement for how much information is lost if the distribution Q is used as an approximation for P, where P is an ideal measure (Adigüzel & Wedel, 2008).

Returning to our problem for measuring the discrepancy between the complete data posterior distribution under the full questionnaire and the observed data posterior distribution under a split questionnaire design, we let Y denote the complete data matrix with n rows and p columns that would have been obtained had the full questionnaire been answered. We let A denote the assignment matrix under a split questionnaire design. A has the same dimensions as Y and we take A_{ij} equal to 0 if variable Y_j is not assigned to individual i and 1 if Y_j is assigned to individual i. Each possible split questionnaire design has indicator matrix A, which can be determined from the form assignments. Let Y_{obs} denote the data that would be observed given A. For any given Y and proposed split questionnaire design, we can use the KL divergence between the complete data posterior distribution, $\pi(\theta|Y)$, and the observed data posterior distribution, $\pi(\theta|Y_{obs})$, to evaluate the performance of that split questionnaire design. Note that the missing data mechanism under split questionnaire design is ignorable and we can base our inference on the observed data posterior distribution in the presence of missing data. This KL divergence for the posterior distributions is given by

$$\int_{-\infty}^{\infty} \pi(\theta|Y) log \frac{\pi(\theta|Y)}{\pi(\theta|Y_{obs})} d\theta. \tag{4.1}$$

For now we assume that this integral is solvable when both Y and Y_{obs} are known. We

discuss methods to approximate this integral when no analytical solution exists in section 4.3. Unfortunately, when administering a split questionnaire design, we will only observe Y_{obs} and not Y. After getting Y_{obs} from a split questionnaire design, our goal is to estimate the KL divergence between the Y_{obs} obtained and the data we would have observed using the complete questionnaire. To do this, we partition Y into (Y_{obs}, Y_{miss}), where Y_{miss} denotes the values of Y that are unobserved due to the split questionnaire design. Our goal should be to minimize the KL divergence across plausible values of Y_{miss} given our prior knowledge of the data before conducting the split questionnaire design. Since Y_{miss} is unobserved, we need to incorporate uncertainty on Y_{miss} into our calculation for the KL divergence between the complete data and observed data posterior distributions. To do this, we draw values of Y_{miss} from the prior predictive distribution of $f(Y_{miss}|Y_{obs}, \theta)$ for a given Y_{obs}. This is done by first drawing a value of θ from the prior $\pi(\theta)$ and then drawing Y_{miss} conditional on Y_{obs} and the draw of θ. Taking $Y^{(l)} = (Y_{obs}, Y_{miss}^{(l)})$, where $Y_{miss}^{(l)}$ is a draw of Y_{miss} from the prior predictive distribution, we can compute the KL divergence at each value of $Y^{(l)}$. This gives us a prior predictive distribution of the KL divergence for a given Y_{obs}. We can use this method to obtain the prior predictive distribution of the KL divergence for any potential split questionnaire design conditional on our observed data.

4.2.2 Criteria for Determining the Optimal Split Questionnaire Design

We will not be able to obtain Y_{obs} before administering the split questionnaire, but, ideally, we want to determine which split questionnaire design will perform best before beginning the study. As a result, simulations will be necessary to determine which design does best at minimizing the loss of information due to missing data. We want to minimize the loss of information based on our prior distribution in order to determine which design performs best based on our current knowledge of the data structure. We first generate underlying data from our prior distribution. From that data, we can determine Y_{obs} and the distribution of the KL divergence for every potential split questionnaire design. Due to sampling uncertainty, we should repeat this process, generating multiple datasets and determining the prior predictive distribution of the KL divergence under each Y_{obs}. Figure 4.1 displays how the prior predictive distributions of the KL divergence might look for m complete datasets and under d different designs, where we have a unique prior predictive distribution for each generated dataset and split questionnaire design.

Once we have the prior predictive distributions for the KL divergence of each proposed

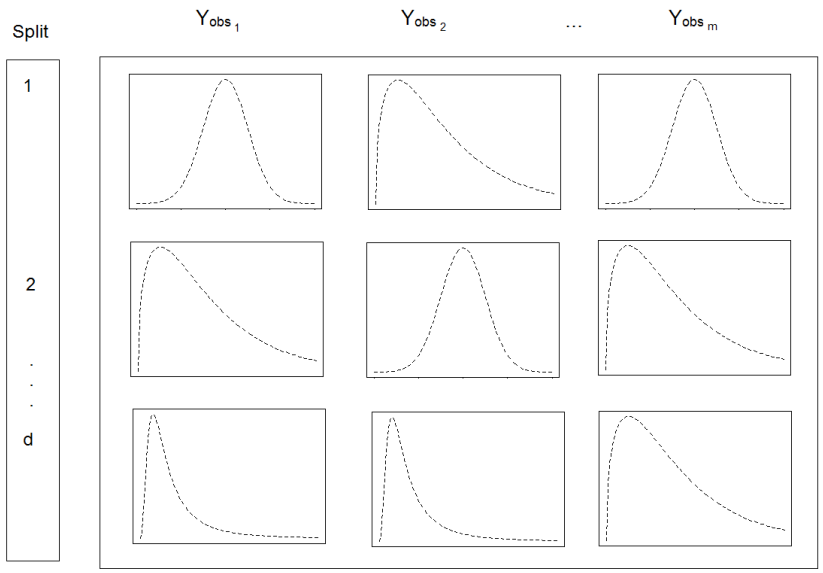

Figure 4.1: Illustration of prior predictive distributions of the KL divergence for m complete datasets and d designs.

questionnaire design under multiple datasets, we need a criterion for determining which one performed the best. There are several possible choices, the most straight forward being to first take the mean of the simulated prior predictive distribution of the KL divergence for each proposed design within each generated dataset. With m complete datasets and under d different designs, we would have a matrix with d row and m columns containing the prior predictive means. From the $d \times m$ matrix, we can take the average for each row and take the design with the smallest average as our preferred split questionnaire design. In this instance, we are choosing the design with the lowest average KL divergence across all simulated data from our prior distribution.

Another possibility is, after taking the mean of the prior predictive distributions of the KL divergence and obtaining the $d \times m$ matrix, to determine the rank for each design within a generated complete dataset. Essentially, we are ranking each row within a single column of the $d \times m$ matrix. We can then determine which split questionnaire design had the lowest average ranking across the different datasets, giving us the design with best ranking across our plausible data. This method also gives us a better idea for how often a certain design performs better than other designs for a given dataset. There could be other additional criteria for selecting a split questionnaire design based on the simulation output, but these two are the only options we considered. In most instances, the rankings based on both criteria were quite similar, so we opted to use the average KL divergence across the simulations.

Since determining the optimal split questionnaire design requires simulating the underlying data from the prior, one might consider simulating complete data using the prior distribution and then directly computing the KL divergence between the observed and complete data posterior distributions for that generated dataset and repeating the process. While it is possible to use that method, we recommend using the proposed method as it was more computationally efficient in the examples that we examined. In many instances, we are likely to have a closed form for $\pi(\theta|Y)$, but have to approximate $\pi(\theta|Y_{obs})$ using MCMC. In which case, estimating $\pi(\theta|Y_{obs})$ is the most time consuming step for determining the KL divergence. By drawing Y_{miss} conditional on Y_{obs} and draws of θ from our prior distribution, we only have to estimate $\pi(\theta|Y_{obs})$ using a MCMC once for each design and dataset and can quickly approximate the prior predictive distribution for each Y_{obs}. In addition, simulations indicate that both methods will lead to very similar conclusions, but the proposed method produces a much smaller standard error in the average KL divergence across the same number of generated datasets. This makes sense as in both instances we are generating our data from the prior distribution and averaging across generated values. For our method, we first generate Y_{obs} from the prior distribution and then generate values of Y_{miss} conditional on Y_{obs}. More details on how these two methods compare can be found in section 4.5.1.

4.3 Computing the KL Divergence

For this section we focus on how to compute the KL divergence for a given value of Y_{obs} and Y.

4.3.1 Approximating the Integral when there is no Analytical Solution

The KL divergence between the complete data and observed data posterior distribution is given by equation (4.1). If $\pi(\theta|Y)$ and $\pi(\theta|Y_{obs})$ are known, we can directly solve the integral and it is fairly straight forward to calculate the KL divergence for a given value of Y. For some cases this integral may not have an analytical solution even when $\pi(\theta|Y)$ and $\pi(\theta|Y_{obs})$

have known distributions. In which case we can see that,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \pi(\theta|Y) log \frac{\pi(\theta|Y)}{\pi(\theta|Y_{obs})} d\theta \\
&= \int_{-\infty}^{\infty} \pi(\theta|Y) log\pi(\theta|Y) - \pi(\theta|Y) log\pi(\theta|Y_{obs}) d\theta \\
&= E_{\pi(\theta|Y)} \left\{ log\pi(\theta|Y) \right\} - E_{\pi(\theta|Y)} \left\{ log\pi(\theta|Y_{obs}) \right\}.
\end{aligned}$$

The KL divergence is equivalent to the difference in the expectation, under the complete data posterior, of the natural logarithm of both posterior distributions. We can approximate the KL divergence using

$$\frac{1}{g} \sum_{j=1}^{g} \left\{ log\pi(\theta^{(j)}|Y) - log\pi(\theta^{(j)}|Y_{obs}) \right\}, \tag{4.2}$$

where $\theta^{(j)}$ are draws of θ from $\pi(\theta|Y)$, yielding a simple way to approximate the KL divergence when both posterior distributions have a closed form.

Example: KL Divergence for Multivariate Normal Y with Known Covariance and Unknown Mean

We now examine estimating the KL divergence when data follow a multivariate normal distribution with a known covariance matrix. Let $Y_i|\mu, \Sigma \sim N_p(\mu, \Sigma)$, where N_p is a p dimensional multivariate normal distribution. Y is a matrix containing the collection of Y_i. We will assume that Σ is a fixed constant and take our prior for μ as the conjugate prior when we have no missing data, $\pi(\mu) = N_p(\mu_0, \Lambda_0)$. In which case, we have $\pi(\mu|Y, \Sigma) = N_p(\mu_n, \Lambda_n)$, where $\Lambda_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$ and $\mu_n = \Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\overline{Y})$ (Hoff, 2009).

When Σ is a known constant, the observed data posterior, $\pi(\mu|Y_{obs}, \Sigma)$, has a closed form solution as well. For each observation, we can take $Y_{i_{obs}} = d_i Y_i$, where d_i is a matrix constructed by taking the identity matrix, I_p, and deleting all rows j corresponding to where variable y_j is missing for subject i, with $1 \leq j \leq p$. The matrix d_i is used to marginalize the multivariate normal distribution over the missing values. We then have $Y_{i_{obs}}|\mu, \Sigma \sim N_{p_i}(d_i\mu, d_i\Sigma d_i^T)$, where p_i is the number of observed variables for individual i.

We have

$$f(Y_{obs}|\mu,\Sigma) \propto exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^T d_i^T(d_i\Sigma d_i^T)^{-1}d_i(Y_i-\mu)\right\},$$
$$\pi(\mu) \propto exp\left\{-\frac{1}{2}(\mu-\mu_0)^T\Lambda_0^{-1}(\mu-\mu_0)\right\},$$

and the posterior

$$\pi(\mu|Y_{obs},\Sigma) \propto$$
$$exp\left\{-\frac{1}{2}\mu^T[\Lambda_0^{-1}+\sum_{i=1}^{n}d_i^T(d_i\Sigma d_i^T)^{-1}d_i]\mu+\mu^T[\Lambda_0^{-1}\mu_0+\sum_{i=1}^{n}d_i^T(d_i\Sigma d_i^T)^{-1}d_iY_i]\right\}.$$

So, we have

$$\pi(\mu|Y_{obs},\Sigma) \sim N_p(\mu_{n_{obs}},\Lambda_{n_{obs}}),$$

where

$$\Lambda_{n_{obs}} = \left\{\Lambda_0^{-1}+\sum_{i=1}^{n}d_i^T(d_i\Sigma d_i^T)^{-1}d_i\right\}^{-1} \text{ and}$$
$$\mu_{n_{obs}} = \Lambda_{n_{obs}}\left\{\Lambda_0^{-1}\mu_0+\sum_{i=1}^{n}d_i^T(d_i\Sigma d_i^T)^{-1}d_iY_i\right\}.$$

Therefore, given Y, we can compute both $log\pi(\mu|Y,\Sigma)$ and $log\pi(\mu|Y_{obs},\Sigma)$ and estimate the KL divergence using (4.2) or we can directly solve (4.1), with the KL divergence being

$$\frac{1}{2}\left\{tr(\Lambda_{n_{obs}}^{-1}\Lambda_n)+(\mu_{n_{obs}}-\mu_n)^T\Lambda_{n_{obs}}^{-1}(\mu_{n_{obs}}-\mu_n)-p+ln\frac{|\Lambda_{n_{obs}}|}{|\Lambda_n|}\right\}.$$

4.3.2 Observed Data Posterior has no Known Distribution

Unfortunately, due to missing data, we will not be able to obtain a closed form for $\pi(\theta|Y_{obs})$ in many instances, even if we use a conjugate prior for $\pi(\theta|Y)$. For cases where $\pi(\theta|Y_{obs})$ does not have a closed form, we can approximate the posterior distribution through Gibbs sampling by adding a data augmentation step. Data augmentation is performed by taking Y and partitioning it into Y_{obs} and Y_{miss} and running a Gibbs sampler where, at iteration t, we draw $Y_{miss}^{(t)}$ from the conditional distribution $p(Y_{miss}|Y_{obs},\theta^{(t-1)})$ and then draw $\theta^{(t)}$ from

$\pi(\theta|Y_{obs}, Y_{miss}^{(t)})$ (Tanner & Wong, 1987; Little & Rubin, 2002). If we are already running a MCMC to approximate the full data posterior, $\pi(\theta|Y)$, then we only need to add a extra step to each iteration of the MCMC where we draw Y_{miss} conditional on Y_{obs} and our last sampled value of θ. This allows us to approximate the distribution of $\pi(\theta|Y_{obs})$ in most instances where $\pi(\theta|Y)$ has a known distribution or can be approximated using MCMC. Note that the draws of Y_{miss} from this Gibbs sampler are not the same as the prior predictive draws of Y_{miss} discussed in the previous section. In this section, when we refer to drawing Y_{miss}, we are referring to Gibbs sample draws for data augmentation.

After performing MCMC to obtain samples from $\pi(\theta|Y_{obs})$, we still need a method to compute the KL divergence. By definition we have

$$\pi(\theta|Y_{obs}) = \frac{f(Y_{obs}|\theta)\pi(\theta)}{f(Y_{obs})},$$

and if we substitute this value into (4.2), we get

$$\frac{1}{g}\sum_{j=1}^{g}\left\{log\pi(\theta^{(j)}|Y) - log\frac{f(Y_{obs}|\theta^{(j)})\pi(\theta^{(j)})}{f(Y_{obs})}\right\}, \tag{4.3}$$

where both the likelihood, $f(Y_{obs}|\theta)$, and prior, $\pi(\theta)$, should have known distributions. We only need to estimate the distribution of the marginal likelihood, $f(Y_{obs})$, to obtain an approximation for the KL divergence. Chib (1995) and Chib & Jeliazkov (2001) proposed methods to approximate the marginal likelihood when the posterior distribution is obtained through Gibbs sampling output and the Metropolis–Hastings algorithm, respectively. If we used data augmentation to approximate $\pi(\theta|Y_{obs})$, it should be fairly simple to estimate $f(Y_{obs})$.

In general, the observed data marginal likelihood can be estimated from

$$log\widehat{f}(Y_{obs}) = logf(Y_{obs}|\theta^*) + log\pi(\theta^*) - log\widehat{\pi}(\theta^*|Y_{obs}), \tag{4.4}$$

where the ordinate θ^* is a value of θ that corresponds to high density point in the posterior distribution $\pi(\theta|Y_{obs})$. It is easy to compute both $f(Y_{obs}|\theta^*)$ and $\pi(\theta^*)$, as those distributions are known. We do need to approximate $\pi(\theta^*|Y_{obs})$, which can be done from the MCMC. The exact steps to estimate $\pi(\theta^*|Y_{obs})$ can differ. When $\pi(\theta^*|Y_{obs}, Y_{miss})$ has a closed distribution

then we can use

$$\widehat{\pi}(\theta^*|Y_{obs}) = \frac{1}{g}\sum_{j=1}^{g}\pi(\theta^*|Y_{obs}, Y_{miss}^{(j)}) \tag{4.5}$$

to approximate $\pi(\theta^*|Y_{obs})$, where $Y_{miss}^{(j)}$ are the draws of Y_{miss} from the data augmentation Gibbs sampler. Here we are treating the missing data as a latent variable and we integrate out Y_{miss} by averaging across the imputed values to get the marginal distribution $\pi(\theta|Y_{obs})$.

Example: KL Divergence of Multivariate Normal Y with Unknown Mean and Covariance

We will now examine estimating the KL divergence when data are multivariate normal with both the mean and covariance unknown. We take $Y_i|\mu, \Sigma \sim N_p(\mu, \Sigma)$ and use conjugate priors for μ and Σ. Take $\pi(\mu|\Sigma) = N_p(\mu_0, \frac{1}{m}\Sigma)$ and $\pi(\Sigma) = W^{-1}(\nu_0, \Phi_0)$, where W^{-1} is an inverse–Wishart distribution, ν_0 is the prior degrees of freedom, and Φ_0 is a positive definite matrix. Then, we have $\pi(\mu|Y, \Sigma) = N_p(\mu_n, \frac{1}{n+m}\Sigma)$, where $\mu_n = \frac{n\overline{Y}+m\mu_0}{n+m}$ and $\pi(\Sigma|Y) = W^{-1}(\nu_n, \Phi_n)$, with $\nu_n = \nu_0 + n$ and $\Phi_n = \Phi_0 + S + \frac{nm}{n+m}(\overline{Y} - \mu_0)^T(\overline{Y} - \mu_0)$ (Murphy, 2007). S denotes the sum of squares for Y, $S = (Y - \overline{Y})^T(Y - \overline{Y})$. The posterior distribution under complete data is equal to $\pi(\mu|Y, \Sigma)\pi(\Sigma|Y)$ and we can sample directly from the posterior by first drawing $\Sigma|Y$ and then drawing $\mu|Y, \Sigma$.

With the presence of missing data and both μ and Σ unknown, we will not be able to find an analytical solution to $\pi(\mu, \Sigma|Y_{obs})$ and will instead need to approximate the posterior through Gibbs sampling using data augmentation. Let us partition each Y_i into Y_{obs_i} and Y_{miss_i} and for each subject partition μ into $\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}$ and Σ into $\begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2},\end{bmatrix}$, where μ_1 corresponds to the missing variables, μ_2 corresponds to the observed variables on subject i, and the partitions of Σ correspond to the variance–covariance components of the partitioned μ.

Now, for a given draw of $\mu^{(j)}$ and $\Sigma^{(j)}$ from a Gibbs sampler, the distribution of $Y_{miss_i}|Y_{obs_i}, \mu^{(j)}\Sigma^{(j)} \sim N_{p_{miss_i}}(\mu_{miss_i}, \Sigma_{miss_i})$, where p_{miss_i} is the number of missing variables on subject i.

$$\mu_{miss_i} = \mu_1^{(j)} + \Sigma_{1,2}^{(j)}\Sigma_{2,2}^{(j)-1}(Y_{obs_i} - \mu_2^{(j)})$$
$$\text{and } \Sigma_{miss_i} = \Sigma_{1,1}^{(j)} - \Sigma_{1,2}^{(j)}\Sigma_{2,2}^{(j)-1}\Sigma_{2,1}^{(j)}.$$

Using this conditional distribution for every subject, we draw $Y_{miss}|Y_{obs}, \mu^{(j)}\Sigma^{(j)}$ in the Gibbs

sampler. The distribution of μ and Σ given Y_{miss_i} and Y_{obs_i} is the same as the posterior distribution with no missing data.

After running the Gibbs sampler to obtain draws from $\pi(\mu, \Sigma|Y_{obs})$, we can approximate the marginal likelihood, $f(Y_{obs})$, using the method described by Chib (1995). We first take a fixed value for μ^* and Σ^* that corresponds to a high density point in the posterior distribution (we used the posterior means). We take

$$\widehat{\pi}(\mu^*, \Sigma^*|Y_{obs}) = \frac{1}{g}\sum_{j=1}^{g} \pi(\mu^*|Y_{obs}, Y_{miss}^{(j)}, \Sigma^*)\pi(\Sigma^*|Y_{obs}, Y_{miss}^{(j)})$$

to approximate the observed data posterior density, where $Y_{miss}^{(j)}$ are draws of Y_{miss} from the Gibbs sampler. We then estimate the marginal likelihood using

$$log\widehat{f}(Y_{obs}) = logf(Y_{obs}|\mu^*, \Sigma^*) + log\pi(\mu^*|\Sigma^*) + log\pi(\Sigma^*) - log\widehat{\pi}(\mu^*, \Sigma^*|Y_{obs}),$$

and we can use equation (4.3) to approximate the KL divergence for any split questionnaire given Y.

Example: KL Divergence when Y is a Collection of Binary or Categorical Variables using the Saturated Model

We now move onto to estimating the KL divergence when the complete data are a collection of p binary and/or categorical variables, $y_1...y_p$. In this instance, there are a finite number of combinations of $y_1...y_p$ that we can observe. In the case where the variables are all binary, there are 2^p unique covariate profiles. We can classify and group individuals by which of the unique covariate profiles they belong to and model the joint likelihood of the binary variables using the multinomial distribution,

$$f(Y|\theta) = \frac{n!}{\prod_{j=1}^{m} n_j!}\prod_{j=1}^{m}\theta_j^{n_j},$$

where n is the total sample size, m is the total number of unique covariate profiles, n_j is the number of individuals that belong to group j, and θ_j represents the probability of belonging to group or cell count j. Alternatively, we can treat each cell count as an independent

Poisson distribution and take

$$f(Y|\mu) = \exp(-\sum_{j=1}^{m} \mu_j) \frac{\prod_{j=1}^{m} \mu_j^{n_j}}{\prod_{j=1}^{m} n_j!},$$

where μ_j represents the expected count in cell j. It has been demonstrated that inference is identical for both likelihood formulations when using conjugate priors for θ and μ in the saturated model (Forster, 2010). In this example, we illustrate computing the KL divergence using the saturated model with the multinomial likelihood and a conjugate prior for θ.

If we take $\pi(\theta) \sim Dir(\gamma)$, a Dirichlet distribution with hyperparameter $\gamma = (\gamma_1, ..\gamma_m)$, then we have $\pi(\theta|Y) \sim Dir(\gamma_n)$, where $\gamma_n = (\gamma_1 + n_1, ..\gamma_m + n_m)$. Thus, computing the posterior distribution is very straight forward when we have no missing data.

For Y_{obs}, we first note that we can group observations based on their missing data pattern, and there should only be a handful of unique missing data patterns resulting from the split questionnaire design. For each missing data pattern, we only have a finite number of covariate profiles which we can use to group observations. Thus, we can model the likelihood in each unique missing data pattern using a multinomial distribution with the probability obtained by marginalizing over the missing values. The observed data likelihood is obtained by multiplying the likelihoods from each missing data pattern.

For a simple example demonstrating how to construct the likelihoods, suppose we have two binary variables y_1 and y_2 and a total sample size of n. With no missing data, there are four unique covariate profiles from the observed values (y_1, y_2) to classify subjects, groups (0,0), (1,0), (0,1), and (1,1). We will let n_{00}, n_{10}, n_{01}, and n_{11} denote the total number of subjects and θ_{00}, θ_{10}, θ_{01}, and θ_{11} denote the probabilities corresponding to groups (0,0), (1,0), (0,1), and (1,1). The data likelihood is given by the multinomial distribution

$$f(Y|\theta) = \frac{n!}{n_{00}!n_{10}!n_{01}!n_{11}!} \theta_{00}^{n_{00}} \theta_{10}^{n_{10}} \theta_{01}^{n_{01}} \theta_{11}^{n_{11}}.$$

Suppose all n subjects are missing the variable y_2. Instead of groups (0,0), (1,0), (0,1), and (1,1), we now only observe a total of two groups: a group of size n_{0*}, where $y_1 = 0$, and a group of size n_{1*}, where $y_1 = 1$. We have sample sizes $n_{0*} = n_{00} + n_{01}$ and $n_{1*} = n_{10} + n_{11}$. The probability $y_1 = 0$ ($\theta_{0*} = \theta_{00} + \theta_{01}$) and the probability $y_1 = 1$ ($\theta_{1*} = \theta_{10} + \theta_{11}$) are obtained by marginalizing over y_2. The observed likelihood in this case is given by a binomial distribution

$$f(Y_{obs}|\theta) = \frac{n!}{n_{0*}!n_{1*}}\theta_{0*}^{n_{0*}}\theta_{1*}^{n_{1*}}.$$

After obtaining the observed data likelihood, we need to approximate the observed data posterior distribution using Gibbs sampling. As we are using conjugate priors, sampling from the posterior conditional on the imputed data is straight forward, $\pi(\theta|Y_{obs}, Y_{miss}^{(g)}) \sim Dir(\gamma_n^{(g)})$, where $\gamma_n^{(g)} = (\gamma_1 + n_1^{(g)}, ..\gamma_m + n_m^{(g)})$ and $n_j^{(g)}$ is the sample size for cell j given the imputed data at iteration g. So, we just need to impute $Y_{miss}|Y_{obs}, \theta$ to complete the Gibbs sampler.

To obtain the observed data likelihood, we had to determine the multinomial probabilities within each missing data pattern by marginalizing over the missing values. These probabilities are obtained by summing the probabilities from the cells in the full data likelihood that are collapsed to create the observed data cells. We can reverse these steps to find the probabilities for mapping the collapsed cells back to the full data cells. If we have draw $\theta^{(g)}$ from the posterior, then, for each missing data pattern, we can impute the observed data into the full data cells by sampling from multinomial distributions. The multinomial distributions have a sample size equal to the sample size in the observed data cell and probability given by the cell probabilities needed to map the collapsed cell back to the full data.

We return to our earlier example where we have variables y_1 and y_2 but only y_1 is observed. Say we draw $\theta^{(g)}$ from the Gibbs sampler. Now, we want to map n_{0*} and n_{1*} back to n_{00}, n_{10}, n_{01}, and n_{11}, but we do not observe those sample sizes. We know that n_{0*} has to belong to either n_{00} or n_{01}. Given $\theta^{(g)}$, we know that the probability of a single individual with $y_1 = 0$ belonging to (0,0) is given by $\theta_{00}^{(g)}/\theta_{0*}^{(g)}$. Thus, we have $n_{00}|n_{0*}, \theta^{(g)} \sim Bin(n_{0*}, \theta_{00}^{(g)}/\theta_{0*}^{(g)})$, and we can impute n_{00} by sampling from a binomial distribution of size n_{0*} and impute n_{01} as $n_{0*} - n_{00}$. Similarly, we can impute n_{10} and n_{11} from n_{1*}.

Thus, we can impute our full data cells from our observed data using the cell probabilities obtained from $\theta^{(g)}$ for any missing data pattern. We then sum the imputed cell counts across all missing data patterns to obtain imputed data for the joint distribution and sample θ conditional on the imputed data. From this Gibbs sampler, we approximate the marginal likelihood, $f(Y_{obs})$, using equations (4.5) and (4.4). We then use (4.3) to estimate the KL divergence.

4.3.3 Complete Data Posterior has no Known Distribution

Extra work is required to obtain the estimate $\widehat{\pi}(\theta^*|Y_{obs})$ when $\pi(\theta|Y)$ does not have a known distribution. Based on the methods discussed in Chib (1995), we can develop an approach for approximating the distribution of $\pi(\theta^*|Y_{obs})$ when samples from the distribution $\pi(\theta|Y_{obs}, Y_{miss})$ are obtained using Gibbs sampling. Let us first consider the case where we can partition θ into (θ_1, θ_2) and the conditional distributions $\pi(\theta_1|\theta_2, Y_{obs}, Y_{miss})$ and $\pi(\theta_2|\theta_1, Y_{obs}, Y_{miss})$ have closed forms. We then have

$$\pi(\theta^*|Y_{obs}) = \pi(\theta_2^*|Y_{obs}, \theta_1^*)\pi(\theta_1^*|Y_{obs})$$

and we can estimate $\pi(\theta_1^*|Y_{obs})$ with

$$\widehat{\pi}(\theta_1^*|Y_{obs}) = \frac{1}{g}\sum_{j=1}^{g} \pi(\theta_1^*|\theta_2^{(j)}, Y_{obs}, Y_{miss}^{(j)}),$$

where $\theta_2^{(j)}$ and $Y_{miss}^{(j)}$ are draws of θ_2 and Y_{miss} from the Gibbs sampler used to approximate the distribution of $\pi(\theta|Y_{obs}, Y_{miss})$. This marginalizes over θ_2 and Y_{miss} to estimate $\pi(\theta_1|Y_{obs})$. To estimate $\pi(\theta_2^*|Y_{obs}, \theta_1^*)$, we first fix θ_1 at the ordinate θ_1^* and run a new Gibbs sampler using the conditional distributions $\pi(\theta_2|\theta_1^*, Y_{obs}, Y_{miss})$ and $p(Y_{miss}|Y_{obs}, \theta_1^*, \theta_2)$. We then estimate $\pi(\theta_2^*|Y_{obs}, \theta_1^*)$ using

$$\widehat{\pi}(\theta_2^*|Y_{obs}, \theta_1^*) = \frac{1}{m}\sum_{k=1}^{m} \pi(\theta_2^*|\theta_1^*, Y_{obs}, Y_{miss}^{(k)}),$$

where $Y_{miss}^{(k)}$ are draws of Y_{miss} from the new Gibbs sampler. We can then take

$$\widehat{\pi}(\theta^*|Y_{obs}) = \widehat{\pi}(\theta_2^*|Y_{obs}, \theta_1^*)\widehat{\pi}(\theta_1^*|Y_{obs})$$

to estimate $\pi(\theta^*|Y_{obs})$ and plug that value into (4.4) to obtain an estimate for $f(Y_{obs})$. In general, if there are B full conditional distributions that we use in our Gibbs sampler and θ can be broken into B blocks with $\theta = (\theta_1, \theta_2, ..., \theta_B)$, then we have

$$\pi(\theta^*|Y_{obs}) = \pi(\theta_B^*|Y_{obs}, \theta_{B-1}^*, ...\theta_1^*)...\pi(\theta_2^*|Y_{obs}, \theta_1^*)\pi(\theta_1^*|Y_{obs}).$$

Once again, we can obtain an estimate for each density evaluated at our posterior ordinate through B Gibbs sampling chains. For each chain, we fix the values of the θ blocks that

we condition on at the posterior ordinate. We then average the density over the Gibbs sampler draws for all parameters that are not explicitly conditioned on to estimate that density. After estimating the B densities in our decomposition of $\pi(\theta^*|Y_{obs})$, we multiply the estimated densities. Using this method we can approximate the marginal likelihood, $f(Y_{obs})$, in any instance where $\pi(\theta|Y_{obs}, Y_{miss})$ is estimated through Gibbs sampling.

We must rely on another method in instances where we cannot use Gibbs sampling to approximate $\pi(\theta|Y_{obs}, Y_{miss})$. Chib & Jeliazkov (2001) provided a method for approximating the marginal likelihood when the posterior distribution is estimated using Metropolis–Hastings. We can adapt this method in order to estimate $f(Y_{obs})$ when the distribution of $\pi(\theta|Y_{obs}, Y_{miss})$ is simulated using the Metropolis–Hastings algorithm.

The Metropolis–Hastings algorithm is used to sample from a distribution when the exact distribution is unknown, but a function proportional to that distribution is known (Hastings, 1970). Here, for a fixed value of Y_{miss}, we have $\pi(\theta|Y_{obs}, Y_{miss}) \propto f(Y_{obs}, Y_{miss}|\theta)\pi(\theta)$. We will assume the sampling of $\pi(\theta|Y_{obs}, Y_{miss})$ is done in a single block. The algorithm works by starting at some initial value θ_0. From θ_0, we sample a value, θ', from density $q(\theta'|\theta_0)$. The function $q(\theta'|\theta_0)$ represents the proposal density or jumping distribution for proposing a movie to θ' given we are at θ_0. We then accept a move to θ' with probability

$$\alpha(\theta'|\theta_0, Y_{obs}, Y_{miss}) = min\left\{1, \frac{f(Y_{obs}, Y_{miss}|\theta')\pi(\theta')}{f(Y_{obs}, Y_{miss}|\theta_0)\pi(\theta_0)}\frac{q(\theta_0|\theta')}{q(\theta'|\theta_0)}\right\},$$

where $\alpha(\theta'|\theta_0, Y_{obs}, Y_{miss})$ denotes the probability of moving from θ_0 to θ'. If we do not accept the proposed moved then we remain at θ_0.

We repeat this over many iterations to generate samples from $\pi(\theta|Y_{obs}, Y_{miss})$ when Y_{miss} is known, but, with Y_{miss} unknown, we instead simulate $\pi(\theta|Y_{obs})$ using data augmentation. We use a Gibbs sampler and generate an instance of Y_{miss} from $p(Y_{miss}|\theta, Y_{obs})$ and then use a single step from the Metropolis–Hastings algorithm to sample a value of θ from $\pi(\theta|Y_{obs}, Y_{miss})$. We repeat this process to simulate the distribution of $\pi(\theta|Y_{obs})$.

We still need to estimate $f(Y_{obs})$ after generating the samples. Using the methods described in Chib & Jeliazkov (2001), for a fixed value of Y_{miss} we can estimate

$$\widehat{\pi}(\theta^*|Y_{obs}, Y_{miss}) = \frac{\frac{1}{z}\sum_{g=1}^{z}\alpha(\theta^*|\theta^{(g)}, Y_{obs}, Y_{miss})q(\theta^*|\theta^{(g)})}{\frac{1}{w}\sum_{h=1}^{w}\alpha(\theta^{(h)}|\theta^*, Y_{obs}, Y_{miss})}, \tag{4.6}$$

where $\theta^{(g)}$ are draws from $\pi(\theta|Y_{obs}, Y_{miss})$ using the Metropolis–Hastings algorithm for a fixed value of Y_{miss}, and $\theta^{(h)}$ are draws from $q(\theta'|\theta^*)$, where θ^* is fixed at a high density

point. We can then estimate $\pi(\theta^*|Y_{obs}, Y_{miss})$ for each value of Y_{miss} generated from the data augmentation Gibbs sample and take the average of those values using (4.5) to obtain an estimate for $\pi(\theta^*|Y_{obs})$, which we can use to approximate $f(Y_{obs})$.

A computationally more efficient method when using data augmentation is to instead fix θ at θ^* and take new samples of Y_{miss} from $p(Y_{miss}|Y_{obs}, \theta^*)$. Then we sample the same number of θ from the proposal distribution $q(\theta'|\theta^*)$. From that, we can take

$$\widehat{\pi}(\theta^*|Y_{obs}) = \frac{\frac{1}{z}\sum_{g=1}^{z} \alpha(\theta^*|\theta^{(g)}, Y_{obs}, Y_{miss}^{(g)}) q(\theta^*|\theta^{(g)})}{\frac{1}{w}\sum_{h=1}^{w} \alpha(\theta^{(h)}|\theta^*, Y_{obs}, Y_{miss}^{(h)})}, \tag{4.7}$$

where both $\theta^{(g)}$ and $Y_{miss}^{(g)}$ are draws from the data augmentation Gibbs sample with a nested Metropolis–Hastings algorithm, $\theta^{(h)}$ are the draws from $q(\theta'|\theta^*)$, and $Y_{miss}^{(h)}$ are draws from $p(Y_{miss}|Y_{obs}, \theta^*)$. From this, we can approximate $f(Y_{obs})$.

In these instances, since $\pi(\theta|Y)$ does not have a known distribution, we also have to approximate $\pi(\theta|Y)$ through MCMC. In which case we would use

$$\pi(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{f(Y)}$$

and substitute into equation (4.3) to obtain

$$\frac{1}{g}\sum_{j=1}^{g} \left\{ log\frac{f(Y|\theta^{(j)})\pi(\theta^{(j)})}{f(Y)} - log\frac{f(Y_{obs}|\theta^{(j)})\pi(\theta^{(j)})}{f(Y_{obs})} \right\}. \tag{4.8}$$

Since $f(Y|\theta)$ has a know distribution, we only need to estimate the marginal likelihood, $f(Y)$. This can be done using the same methods described for approximating $f(Y_{obs})$. We would then use our MCMC draws as samples from our full data posterior distribution and use equation (4.8) to approximate the KL divergence. This gives us a fairly straightforward method to approximate the KL divergence for a given split questionnaire design when Y is known, provided $\pi(\theta|Y)$ can be approximated using MCMC.

Example: KL Divergence when Y is a Collection of Binary or Categorical Variables using a Generalized Linear Model

Previously, we demonstrated how to compute the KL divergence when our data were a collection of binary and categorical variables. We opted to use the saturated model to avoid imposing any assumptions on the joint distribution of the data; however, the number of

parameters will increase exponentially as we add more y variables. It then becomes more difficult to use the saturated model and to place an informative prior on our parameters based on a pilot study or previous publications, as many cells will be empty. Placing certain restrictions on the cell probabilities to reduce the number of parameters is preferable in these instances. A natural way of restricting the cell probabilities would be to use the Poisson likelihood and take $\mu_j = x_j^T \beta$, where x_j is the covariate profile for individuals in group j. The covariate profile is the collection of values for relevant variables, $y_1 ... y_p$, their interactions, and the constant 1, corresponding to the intercept parameter. We can substantially reduce the number of parameters by eliminating higher order interactions of the y variables. This approach is the same as using Poisson regression or a log–linear model.

We can place a multivariate normal prior on β and compute the complete data posterior distribution using the Metropolis–Hastings (MH) algorithm. We can then modify equation (4.6) to estimate $\widehat{\pi}(\beta^*|Y)$ and approximate $f(Y)$ by modifying equation (4.4).

We can simulate the observed data posterior distribution by adding a Gibbs sampling step to our Metropolis–Hastings to impute $Y_{miss}^{(g+1)}|Y_{obs}, \beta^{(g)}$ at each iteration and then draw $\beta^{(g+1)}|Y_{obs}, Y_{miss}^{(g+1)}$ from the MH. After selecting β^*, we can approximate $\pi(\beta^*|Y_{obs})$ using equation (4.7) and obtain our estimate of $f(Y_{obs})$. We would then use equation (4.8) to estimate the KL divergence.

4.4 Search Algorithm

Our method gives us a criterion to measure the performance of specific split questionnaire designs. Even so, there are often many potential designs to consider, making it no easy task to identify the best possible. With an equal number of blocks per split, a total of b blocks of variables, s splits, q blocks per split, and when each questionnaire form is administered in approximately equal numbers, there are

$$\frac{\binom{b}{q}\binom{b-q}{q}...\binom{2q}{q}\binom{q}{q}}{s!}$$

unique split questionnaire designs. We consider there to be only one unique design in a split questionnaire with 3 blocks and 3 splits, even though we have several permutations we could assign blocks 1, 2, and 3 to splits A, B, and C. This is because there is no effective difference in the assignment (A,B,C) versus (C,B,A) when each form is administered in equal numbers.

If the forms are not given in equal numbers, we would have

$$\binom{b}{q}\binom{b-q}{q}...\binom{2q}{q}\binom{q}{q}$$

possible designs. For our purposes, we will assume that the forms will be administered in approximately equal numbers.

If the number of splits is not a divisor of the total number of variable blocks, we have $b = qs + r$, where r an integer in the interval $[1, s)$. If we only consider designs where we allocate blocks so that each split has as close to the same number of blocks as the other splits as possible, then r splits will have $q+1$ blocks and $s-r$ splits will have q blocks. There will be

$$\frac{\binom{s}{r}\binom{b}{q+1}\binom{b-q-1}{q+1}...\binom{b-(r-1)(q+1)}{q+1}\binom{b-r(q+1)}{q}...\binom{2q}{q}\binom{q}{q}}{s!}$$

unique designs to consider.

So, with 3 splits and 9 variable blocks, there are a total of 280 allocations, and the number of possible allocations increases to 2,858,856 with only 18 blocks and 3 splits. As the number of blocks increases, it becomes computationally difficult, if not impossible, to enumerate all possible allocations. As a result, a search algorithm is necessary for finding a very good variable allocation, if not the optimal allocation, in many instances.

Adigüzel & Wedel (2008) used a Modified Federov Algorithm to search for good designs. For our purposes, we consider using simulated annealing, a probabilistic search algorithm that works similarly to Metropolis-Hastings. Any jump that improves the criterion we are trying to optimize is accepted and a move to a less optimal allocation is accepted with a probability related to how much worse it performs and how long the algorithm has been running (Kirkpatrick et al., 1983). This method differs from Metropolis-Hastings as it is used for optimization, not for the approximation of a distribution, and the probability of accepting a worse allocation should decrease as a function of how long the algorithm has been running. Potentially accepting worse designs allows this algorithm to escape from local maxima. It was proven that this algorithm will converge to the global optimum if run for long enough (Granville et al., 1994).

The ease of implementing simulated annealing makes it an appealing search algorithm to use in our case. All we really need to specify is a method for moving to neighboring designs given our current design and the probability of accepting a worse design. In our case, we start from a random split questionnaire design and jump to a neighboring design by

exchanging two variable blocks with each other. We accept a move from our current design, d, to neighbor, d^*, with probability

$$exp\left(\frac{KL_d - KL_{d^*}}{t}\right)$$

when the average of the prior predictive distribution means for the KL divergence of d^* is larger than for d and a probability of 1 when it is less. Here KL_d represents the average of the prior predictive distribution means for the KL divergence of d, and the temperature, t, decreases as a function of the number of iterations the algorithm has run. As the algorithm runs longer and t decreases, the probability of accepting a worse design will decrease. When applying this algorithm, t is usually decreased by multiplying it by a constant between 0 and 1 after each iteration. We are essentially choosing two designs to compare and computing the prior predictive distribution of the KL divergence across multiple simulated datasets and determining which design to choose based on the average of the expected value for each KL divergence distribution across datasets and then repeating the same process over and over again. Each time, we are essentially selecting two rows from figure 4.1 and using them to choose which design to pick based on the overall average. We could instead look at how many datasets that one design performed better than the other. We evaluate our search algorithm in the simulation section.

4.5 Simulation Results

The simulations are broken into three parts corresponding to three examples for calculating the KL divergence from section 4.3. We use the instance of multivariate normal data with a known covariance to compare two methods for computing the average KL divergence over our prior distributions and for examining the effectiveness of our search algorithm. For the KL divergence of multivariate normal data with an unknown mean and covariance and with multinomial data, we perform simulations to determine how well the best designs, as selected using the KL divergence criterion, perform. For the purpose of the simulations, each variable block contains only a single variable.

4.5.1 Multivariate Normal with Known Covariance

Instances where a multivariate normal distribution has a known covariance matrix but unknown mean are rare; however, it is one of the few instances where the complete data and

observed data posterior distributions have closed forms. This makes it much easier computationally to estimate the average KL divergence across datasets and allows us to compare how the proposed method for taking the average of the prior predictive distribution means for the KL divergence across different simulated datasets (method 1) compares to generating complete data from the prior distribution and computing the average KL divergence across the simulated data (method 2).

We examined a case where we have 9 variables and 3 splits, for a total of 280 possible split questionnaire designs. We computed the average KL divergence for each of the 280 designs using both methods. Figure 4.2 displays a scatter plot comparing the average KL divergence from these methods. The plot suggests an extremely strong linear relationship between the two methods. Furthermore, the Pearson correlation is 0.99. This indicates using either method for determining which designs performed best will lead to a similar conclusion. In addition, method 1 produced a smaller standard error for the average KL divergence with the same number of simulated datasets. This makes method 1 more efficient when there is no closed form for the observed data posterior distribution but the complete data posterior has a known distribution.

We then evaluated our search algorithm with our 9 variables and 3 splits scenario, as well as a scenario with 15 variables and 3 splits. With 15 variables, there are 126,126 possible designs to consider. After computing the average KL divergence for each possible design, we apply our search algorithm to both scenarios. We should note that the search algorithm is intended to be used when there are far too many designs to enumerate all possibilities. In these instances, we would typically run the algorithm until it meets some convergence criterion, but we will never know how well that chosen design performs compared to all possible designs. We aim to demonstrate that the algorithm can perform very well by just searching over a small fraction of all possible designs. For this reason, we specified a maximum number of iterations to run the algorithm and return whatever the current design is at that iteration, instead of waiting for convergence. We used a maximum of 90 and 500 iterations for our search algorithm when the number of possible designs were 280 and 126,126, respectively.

We applied the search algorithm multiple times and recorded which split questionnaire design was selected each time and how that design compared to other possible designs, the results of which can be found in Figure 4.3. The percentile rankings of split questionnaire designs and the KL divergence at that ranking are on the X axis, while the cumulative percentage is on the Y axis. The height of the curve indicates the percentage of times that

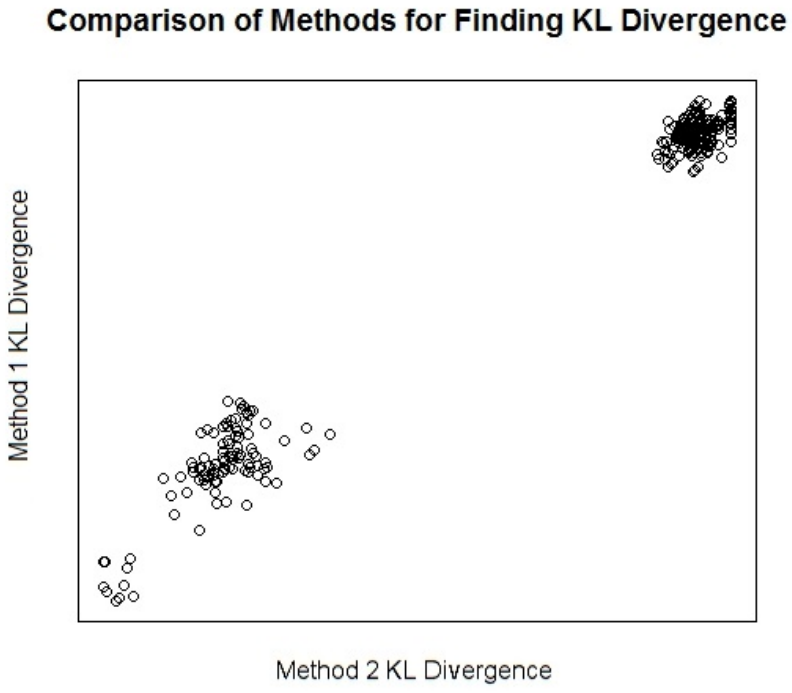

Figure 4.2: Scatter plot of the average KL divergence computed by two different methods for all 280 possible designs.

the design chosen by the search algorithm was better than a certain percent of total designs. The proposed search algorithm chose a split design with a KL divergence in the top five percentile over 80% of the time when there were only 280 possible designs. The best design was chosen 10% of the time, and one of the top 10 designs was selected 60% of the time. For the case with 280 possible designs, we used the same simulation scenario displayed in figure 4.2. There is a very small difference in average KL divergence among the top 10 designs.

In our simulation with 126,126 possible variable allocations, the difference in KL divergence between the best and worst design was nowhere near as pronounced as in the instance with 280 designs. Additionally, the algorithm was run for a much smaller number of iterations compared to the total amount of possible designs, so we would not expect the search algorithm to do as well in this instance. Even so, the search algorithm does very well. The selected design ranked in the top 1 percentile 16% of the time, in the top 10 percentile 44% of the time, and in the top 20 percentile 64% of the time. The top design chosen by the algorithm ranked 31 out of all 126,126 designs in KL divergence and the search algorithm selected one of the top 500 designs 10% of the time. The algorithm likely would have worked even better had it been allowed to run longer or run until convergence; however, even with a small number of iterations this algorithm generally selected good designs according to our KL divergence criterion.

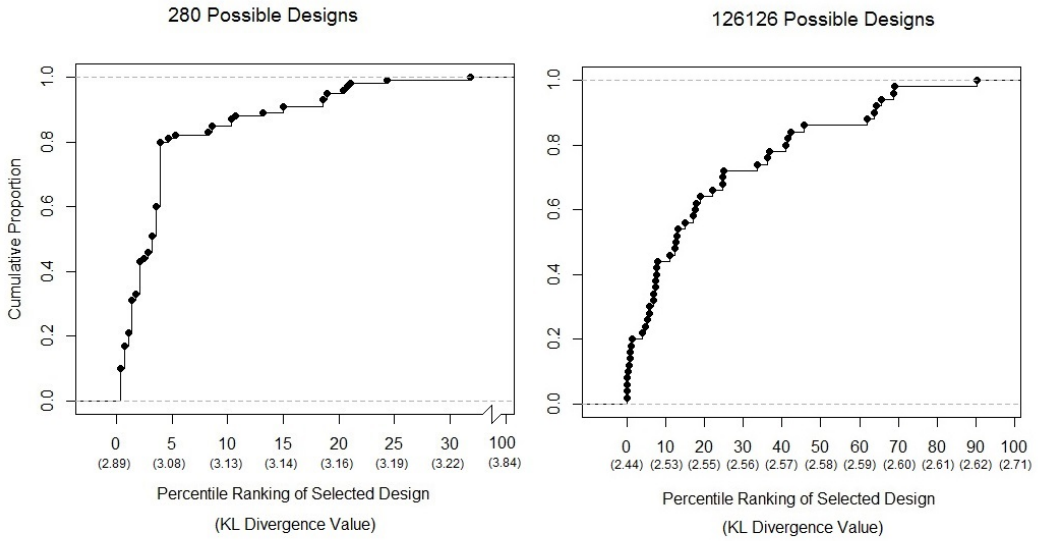

Figure 4.3: Plot of the cumulative percentage of times that the design chosen by the search algorithm was better than a certain percentage of total possible designs for instances with 280 possible designs (left) and 126,126 possible designs (right).

4.5.2 Multivariate Normal with Unknown Covariance

We will now focus on how well our method of using KL divergence to select split questionnaire designs performs in cases where the joint distribution of our data follows a multivariate normal distribution with an unknown mean and covariance. We looked at three different correlation structures when there were a total of 8 variables, 4 splits, and 105 possible split questionnaire designs. Structure 1

$$\begin{bmatrix} 1.0 & 0.8 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.8 & 1.0 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.4 & 0.4 & 1.0 & 0.8 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.8 & 1.0 & 0.4 & 0.4 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 & 1.0 & 0.8 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.8 & 1.0 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 1.0 & 0.8 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.8 & 1.0 \end{bmatrix},$$

structure 2

$$\begin{bmatrix} 1.0 & 0.3 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.3 & 1.0 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1.0 & 0.3 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.3 & 1.0 & 0.6 & 0.6 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 1.0 & 0.3 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.3 & 1.0 & 0.6 & 0.6 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 1.0 & 0.3 \\ 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.3 & 1.0 \end{bmatrix},$$

and structure 3

$$\begin{bmatrix} 1.00 & 0.78 & 0.72 & 0.38 & 0.31 & 0.46 & 0.47 & 0.58 \\ 0.78 & 1.00 & 0.45 & 0.15 & 0.38 & 0.68 & 0.56 & 0.46 \\ 0.72 & 0.45 & 1.00 & 0.74 & 0.33 & 0.22 & 0.45 & 0.63 \\ 0.38 & 0.15 & 0.74 & 1.00 & 0.57 & 0.09 & 0.35 & 0.26 \\ 0.31 & 0.38 & 0.33 & 0.57 & 1.00 & 0.79 & 0.32 & 0.38 \\ 0.46 & 0.68 & 0.22 & 0.09 & 0.79 & 1.00 & 0.42 & 0.58 \\ 0.47 & 0.56 & 0.45 & 0.35 & 0.32 & 0.42 & 1.00 & 0.75 \\ 0.58 & 0.46 & 0.63 & 0.26 & 0.38 & 0.58 & 0.75 & 1.00 \end{bmatrix}$$

represent the underlying correlations for the population. We used the same mean in each correlation structure. The average KL divergence of the prior predictive distribution for each scenario was computed using the method described in section 4.2.2.

How well each design performs in the multivariate normal case is determined by the correlation structure, since the Kullback-Leibler divergence is invariant to transformations and we can always change the variables means by re-centering. As a result, there are only four unique designs for correlation structures 1 and 2, since most of the covariance parameters are identical. We ran the MCMC and stored the observed data posterior parameter means and variances, in addition to computing the KL divergence. The average posterior variance based on the KL divergence ranking for each correlation structure can be seen in table 4.1. The table displays the variance for each unique allocation for structures 1 and 2, while we created four groups based on the computed KL divergence and calculated the average variance for each design within the groups for structure 3. We can see from the table that the designs with a smaller KL divergence produced a smaller average variance. The differences are especially noticeable for structure 1, where a handful of correlations are much greater than the others. In this instance, the larger correlation can make a huge difference in imputation uncertainty. Structure 2 is the opposite, where most correlations are fairly large and only a handful are smaller than the others. In this instance, there is very little difference in variance from each design. For structure 3, there is also a fairly small difference in variance between the groups with the best KL divergence and the groups with the worst KL divergence. Most correlations were fairly moderate, which led to only a slight difference in KL divergences and MCMC variance for most designs.

Table 4.1: Average observed data posterior variance for split questionnaire designs grouped by KL divergence for multivariate normal data.

KL Divergence Rank	Structure 1	Structure 2	Structure 3*
1	0.0212	0.0199	0.00552
2	0.0224	0.0200	0.00553
3	0.0240	0.0200	0.00554
4	0.0269	0.0207	0.00563

* Indicates that MCMC variance is averaged across multiple designs grouped by KL divergence.

4.5.3 Data are a Collection of Binary Variables

Finally, we examined the performance of using the KL divergence to determine variable allocation in a split questionnaire design when our data are a collection of binary variables. In our simulations, we took a total of 6 variables and 3 splits, giving us 15 unique designs and a total of 64 parameters for the saturated model. We used the method described in section 4.2.2 to compute the average KL divergence of the prior predictive distribution. Afterwards, we placed the 15 scenarios into 5 groups based on their KL divergence and examined the average MCMC variance within those groups. The results are displayed in table 4.2. The designs with a lower average KL divergence produced smaller parameter variances, similar to the multivariate normal case. In this particular instance, there was not a huge difference in the variance between the 15 designs.

Table 4.2: Average observed data posterior variance for split questionnaire designs grouped by KL divergence for a collection of binary variables.

KL Divergence Group	Average MCMC Variance*
1	2.902
2	2.925
3	2.928
4	2.938
5	2.971

* Indicates that MCMC variance is multiplied by 10^6

4.6 Health and Retirement Study

We will now revisit the Health and Retirement Study (HRS) and longitudinal split questionnaire design that we explored earlier. In chapter 3, we used HRS to determine which of our proposed longitudinal split questionnaire designs would work best on survey data based on

estimating the variable means and regression parameters. Now, we would like to revisit the longitudinal split questionnaire using the KL divergence to compare our proposed designs. To do this, we used the same components as before: diabetes (D), hypertension and blood pressure (B), heart disease and stroke (H), cancer (C), weight (W), and income and wealth (I). However, this time, we took only a single variable from each component and ignored the demographic variables. For simplicity, we will assume that the variables are jointly distributed as multivariate normal, in which case we only need the variable correlations for determining the best split.

We first took the complete case data from 2004, which we treated as a pilot study, and used it to estimate our population parameters. Using this information, we then created a cross-sectional form with 6 variables and 3 splits based the KL divergence criterion. Table 4.3 displays the chosen cross-sectional form from our method. The correlations between the selected variables were fairly weak, with values ranging from -0.24 to 0.29. As a result, there was not a huge difference in the KL divergence between the possible variable allocations for the cross-sectional forms. The correlations between variables within the same splits as chosen by our method were especially weak, with the strongest within split correlation having a magnitude of 0.09.

Table 4.3: Optimal cross-sectional split questionnaire variable assignments based on KL divergence.

HRS Component	Assigned Split Questionnaire Component
D	B
B	A
H	C
C	A
W	C
I	B

After choosing our cross-sectional form, we determined how to administer that form to the 2006, 2008, and 2010 years of the study, using the five proposed longitudinal designs from table 3.1 in chapter 3. We used the data from those years to generate our priors for the population parameters needed to compute the KL divergence. In practice, this data would not be available beforehand; however, if the complete data from 2004 and the split questionnaire data for 2006 were available, we could use that information to approximate the correlations of the variables over time. Longitudinal correlations between variables across waves did not change very much throughout the study, indicating this is a reasonable approximation.

We applied our method for determining the KL divergence, based the prior information

for the correlation parameters, on the proposed longitudinal designs. Figure 4.4 displays the distributions of the average KL divergence for each generated complete dataset under the five proposed longitudinal designs. We can see from the figure that Option 2 produced the smallest KL divergence across our prior distribution, followed by Option 3. The remaining options all performed similarly in terms of KL divergence. In chapter 3, we found Option 3 to perform the best overall, followed closely by Option 2. These conclusions were based on estimating the means, quantiles, and several regression model parameters. Since we have shortened the number of variables collected and number of years, altered the questionnaire forms, assumed a normal distribution for each variable, and are not looking at the exact same parameter estimates, our results here are not directly comparable to those in chapter 3. Still, the higher ranking of the options where all individuals receive a different split questionnaire form at each year (Option 2 and Option 3) compared to the other options is consistent with what we had seen previously. Interestingly, using simulations to generate multivariate normal data with the correlation structure from the HRS data and analyzing that data using our posterior distributions showed the average MCMC variance to be roughly the same for Option 2 and Option 3, with an average variance of 7×10^{-4} for Option 2 versus 6.97×10^{-4} for Option 3. However, the Mahalanobis distance between the posterior means of the complete data and observed data posterior distributions was smaller for Option 2 (14.3 vs 14.6). This seems to indicate that while the marginal estimates are quite similar, Option 2 better preserves the joint distribution of the parameter space. For the most part, the KL divergence criterion agrees with our conclusions from Chapter 3 and rotating the split questionnaire forms for every individual is preferable.

4.7 Discussion

In this chapter, we presented a new method for ranking and selecting variable allocations in a split questionnaire design. We suggested averaging the means of the prior predictive distributions of the KL divergence between the complete data and observed data posterior distributions. All previously proposed methods for allocating variables in split questionnaire designs require some prior estimates of parameters for the joint distribution of the variables, usually obtained from a pilot study, but often did not incorporate the uncertainty of those estimates in the design. We felt the best way to incorporate the information and uncertainty of our prior knowledge in our design and estimation was to perform Bayesian analysis and use a proper prior distribution.

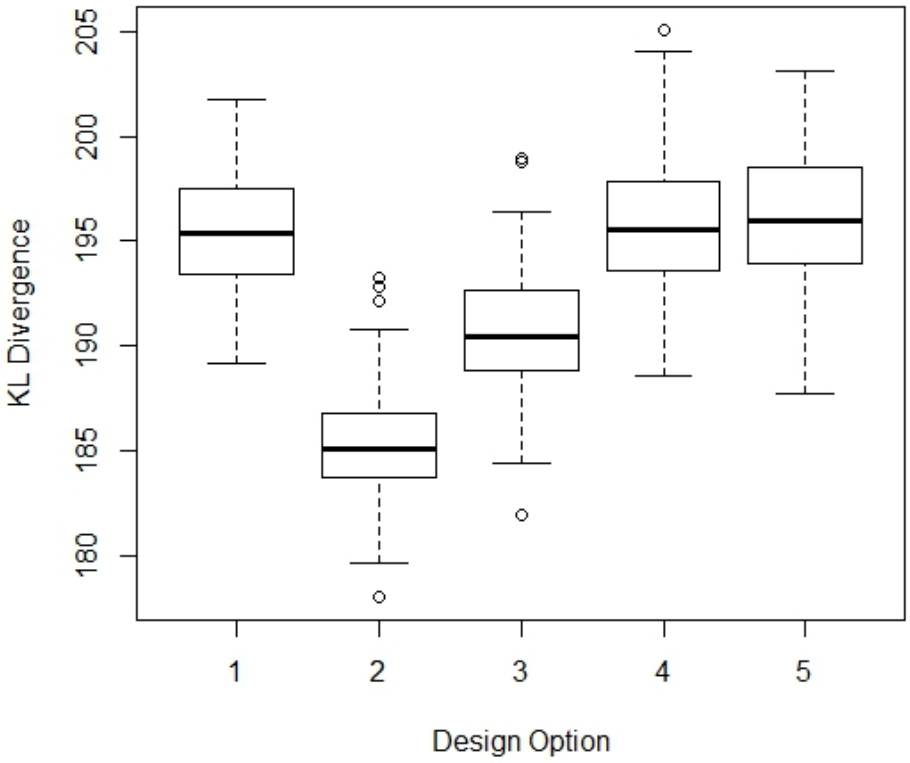

Figure 4.4: Box plots of the distributions of the average prior predictive KL divergence for multiple complete datasets under the five proposed longitudinal designs.

We provided methods for computing the KL divergence, which can be used in almost any type of Bayesian analysis, provided we can approximate our posterior distribution through MCMC methods. We also presented several examples for how to calculate the KL divergences when the joint distribution can be expressed as either a multivariate normal or multinomial distribution. We also proposed a potential search algorithm, using simulated annealing, for locating a good variable allocation in split questionnaire designs when there are a large number of potential variable allocations.

The results from our simulations demonstrated that the search algorithm generally produced good results, even with a small number of iterations. We also demonstrated in our simulations that, overall, the variable designs with better KL divergences produced smaller MCMC variances in instances where the data were distributed as either multivariate normal with unknown mean and variance or multinomial. Furthermore, we showed that using the KL divergence as a criterion to determine which longitudinal split questionnaire design to select leads to a similar conclusion as in chapter 3, with regards to the HRS data. This provides evidence that our KL divergence method could be useful in determining optimal split questionnaire designs. One drawback to using our method is the amount of computational time it takes to estimate the KL divergence in the Bayesian setting.

Naturally, one may wonder why should we use the KL divergence criteria when we could instead calculate which design has the smallest average MCMC variance for each parameter. We used the average MCMC variance as an easy and clear way for demonstrating the usefulness of our method in designing split questionnaire designs, but there is more to characterizing the joint distribution than just looking at the marginal variance of each parameter. There could be instances where the average marginal MCMC variance of two designs are similar, but the observed joint distribution of parameters is much closer to the complete data joint distribution for one of the designs. We saw an example of this in our simulations with the HRS data where Option 2 and Option 3 had similar marginal variances, but Option 2 better preserved the joint distribution of the parameter space. This would indicate that the KL divergence is more complex than simply examining the marginal variances of parameters. Note that the KL divergence between the complete data and observed data posterior distributions for the multivariate normal distribution with a known covariance has a closed form solution for a fixed value of Y. This quantity directly relates to both the Mahalanobis distance of the observed data parameter means compared to complete data means and the variance-covariance matrices for those posterior distributions. Our method may not always produce the smallest marginal variance of each parameter, though it did in many instances,

but should in theory produce the observed data posterior distribution which is closest to the posterior distribution that would have been obtained without any missing data.

Most of the proposed methods for allocating variables in a split questionnaire design require using a pilot study beforehand to obtain information on the population parameters of interest. Presumably, the pilot study uses the full questionnaire, yet one of the major reason for using a split questionnaire design was to reduce respondent fatigue and improve data quality in surveys with long questionnaires. In these instances, it may not be the best idea to use the full questionnaire in a pilot study. It might be a better idea to use a split questionnaire for the pilot study, and, based on those results, alter the split questionnaire as the study progresses. We might even want to consider an adaptive design, where the split questionnaire forms are altered throughout the study based on which design is projected to perform best according to our current data. We could even eliminate the split questionnaire forms and choose a subset of questions to ask the next few individuals entering the study based on our KL divergence criteria and frequently update which subset of questions will be answered, yielding a completely adaptive design. These adaptive designs are worth exploring in the future and should be easier to apply under our Bayesian framework through the use of prior and posterior distributions.

Chapter 5

Limitations and Future Directions

In this paper, we discussed implementing planned missing data designs as a way to either improve efficiency or data quality. We focused on two particular planned missing data designs: two-phase sampling and split questionnaire designs. Our overarching goal was to determine ways of implementing these designs in order to improve their overall efficiency.

In chapter 2, we proposed several methods for selecting a subsample in two-phase sampling to improve the variance of our estimate of interest. The preferred method for sample selection depended on the quantity of interest and the distribution of our variables. For instances where our Phase II variable was continuous, we could directly use the distribution of our Phase I variables to select our subsample. For most other instances, we required information regarding the conditional distribution of Y given X in order to determine the optimal sample size allocation for our subsample. As a result, we proposed conducting the second phase sample in batches in order to obtain information on the conditional distribution.

As mentioned earlier, this paper only examined instances where we had a single outcome variable in Phase II and that outcome was continuous or binary. One natural extension of this work would be to explore cases where we have multiple variables collected in Phase II and our primary interest lies in the estimation of this collection of variables. Another natural extension would be to explore instances where our outcome variable followed another distribution not considered in this paper. We could also examine instances where we do not assume a parametric distribution for our outcome variable beforehand.

Additionally, in chapter 2, we assumed that Phase I of our study had completed prior to starting Phase II. Several of our methods required having Phase I variables available to use in selecting our Phase II sample, which may not always be the case. There are several instances where it is feasible that all Phase I variables would be available to investigators before Phase II, such as when our Phase II sample is subsample of a larger study that has already completed, when our Phase II sample is an expensive biomarker and can be

obtained from a blood sample which is already available for all participants from measuring other biomarkers, or when individuals are sampled from a preexisting database. For some studies, it may be difficult to wait until Phase I completes before beginning the subsample. This could be due to Phase I requiring a long time to complete or because it is cheaper to obtain Phase I and Phase II variables at the same visit.

If sub-sampling begins before the completion of Phase I, several of our proposed methods would have to be modified. When we have a single binary X variable, we can still apply our method for adaptive sampling, though we would have to make a few minor modifications. This time, we would have to estimate and update the parameter values for X as well as Y given X. For instances where Y is continuous, we would need to use another method for selecting or Phase II sample, though we still could use the values of our Phase I variables for selecting our sample. We would also have to modify our method when we are constructing a Phase I categorical variable using propensity scores.

Furthermore, for each of the methods in chapter 2, sub-sampling was available on the individual level. When using cluster sampling to select our initial sample (such as sampling hospitals, schools, etc.), it might be difficult or expensive to select individual participants for our Phase II sample. In these instances, we might prefer selecting specific clusters for our subsample, where each cluster could have a different number of participants and cost for being included in Phase II. In addition to the distribution of our Phase I and Phase II variables and estimand of interest, determining which clusters to select would depend on available sample size and costs.

For chapter 3, we explored several methods for assigning split questionnaire forms in a longitudinal or panel study. We proposed a total of six designs and compared their performances in estimating means, variances, and changes over time under several data structures. The best design depended on the correlation structure and quantity of interest, though we were able to establish some useful guidelines for which design to apply based on the situation. For each design, the split questionnaire forms were unchanged at each visit, but were rotated in a different manor throughout the study. It might be worthwhile to consider changing the variable assignment in the split questionnaire forms throughout the study, instead of rotating the forms. Our method from chapter 4 allows us to determine variable assignments and compare the performance of different designs. We can determine the variable allocation at each wave based on what we have previously assigned and compare it to using the optimal rotation of the fixed forms.

In chapter 3, we evaluated the proposed design options using simulations and HRS data.

Although the HRS data contained both binary and normal variables, we did not examine any joint distributions of variables other than multivariate normal for our simulations. It would be useful to examine whether the conclusions from our simulations are affected by different variable distributions. We might want to consider instances where we have binary, categorical, or count variables in addition to the continuous multivariate normal distribution. Although, in principal, we would not expect a huge difference from the multivariate normal case, the distribution of our variables affects the imputation models and could result in different conclusions.

Finally, in chapter 4, we developed a method for determining optimal split questionnaire designs based on minimizing the impact of missing data on our inference. We proposed using the KL divergence to determine which split questionnaire design performed best and a search algorithm for locating good split questionnaire designs. This was done in a Bayesian framework to better incorporate the initial uncertainty for our population parameters into our design. One drawback is the amount of computation time require to compute the KL divergence using MCMC. The long computation time is worsened by the amount of possible split questionnaire designs needed to consider in many instances. The prior distribution can account for differing levels of uncertainty in each parameter, such as when some parameters were measured on more subjects than other parameters prior to the study. However, in our examples, we only considered instances where the prior sample size is the same for each parameter. This was done in order to use conjugate priors. It might be worth examining the impact using more flexible prior distributions, in terms of variance structure, might have on selecting optimal designs.

This method also requires using an informative prior, which means we need prior information on our population parameters in order to select the optimal split questionnaire design. This specific drawback is present in other approaches for determining variable allocations in split questionnaire designs. Typically, a pilot study which uses the full questionnaire needs to be conducted beforehand. In the future, we might want to determine the impact that using a split questionnaire design for the pilot study might have.

As mentioned earlier, we should also consider an adaptive design for updating the variable assignments to split questionnaire forms throughout the study and see how that compares to fixing the split questionnaire forms after the pilot study. We could even make the design fully adaptive by eliminating split questionnaire forms and simply selecting a subset of variables to be measured on each subject entering the study. The fully adaptive design would be difficult to implement for pen and paper questionnaires, but could prove useful for online or

computer based surveys.

For our chapters on split questionnaire design, we did not explicitly discuss how context effects impact survey design. The ordering of questions within a questionnaire and where a question appears in relation to other questions could have an impact on responses (Schuman & Presser, 1981; Sudman et al., 1996). Context effects must be considered when designing a split questionnaire survey. We can intentionally design variable blocks to account for context effects, but we should also consider what impact the order of blocks and splits within a split questionnaire might have on responses.

Finally, the designs explored in this paper did not consider the effect that complex survey designs might have on our proposed methods. Most of our proposed methods implicitly assumed simple random sampling; however, most large-scale studies use more complex survey designs such as stratified multistage sampling (Zhou et al., 2016). In chapter 2, we discussed estimation in the presence of a categorical Z variable obtained in Phase I. Although we did not explicitly relate this to complex survey designs, we could use this variable to account for strata, clusters, and other design features in the initial sampling and can incorporate weighting into our estimates. For the instances where we have a continuous outcome variable, we could on apply our proposed methods for selecting a Phase II sample within each strata and/or cluster from the original sample.

It is more difficult to directly incorporate complex survey designs into our methods for split questionnaire designs in chapters 3 and 4. Chapter 3 focuses on methods for rotating split questionnaire forms in longitudinal studies and provides guidelines on which method works best for certain estimands under different correlation structures. This focuses specifically on the data obtained in the study and not how it generalizes to the population of interest. As a result, the features of complex survey designs should not have much of an impact on the conclusions for which longitudinal design to use. For chapter 4, where we attempt to determine the optimal split questionnaire design to use based on pilot data and previous studies, the survey design features of the studies conducted beforehand could impact our conclusions for which design to use. One possible way of dealing with survey design features is to create a synthetic population from the prior data. Dong et al. (2014) provided a nonparametric method to create synthetic populations that account for stratified, clustered, and unequal probability sampling by using the Bayesian bootstrap. We can ignore the complex sampling design and treat the generated synthetic population as a simple random sample from the population of interest. We could then use the synthetic data as our pilot data and apply our method for determining the optimal split questionnaire design.

The two types of planned missing data designs explored in this paper can be used to improve certain studies. In this paper, we provided methods for selecting a subsample in two-phase studies, guidance for implementing a split questionnaire design in a longitudinal study, and ways to assign variables to different splits within a split questionnaire design. These methods can be used to further improve the usefulness of planned missing data designs in practice; however, there is still a good deal more to explore within each of these topics.

REFERENCES

Adams, L., & Darwin, G. (1982). Solving the quandary between questionnaire length and response rate in educational research. *Research in Higher Education*, *17*(3), 231–240.

Adigüzel, F., & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608-617.

Alonzo, T. A. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics*, *4*(2), 313-326.

Avendano, M., & Glymour, M. M. (2008). Stroke disparities in older americans: Is wealth a more powerful indicator of risk than income and education? *Stroke*, *39*(5), 1533–1540. doi: 10.1161/STROKEAHA.107.490383

Beckett, M. K., Elliott, M. N., Gaillot, S., Haas, A., Dembosky, J. W., Giordano, L. A., & Brown, J. (2016). Establishing limits for supplemental items on a standardized national survey. *Public Opinion Quarterly*, *80*(4), 964-976.

Best, L. E., Hayward, M. D., & Hidajat, M. M. (2005). Life course pathways to adult-onset diabetes. *Social biology*, *52*(3-4), 94–111. doi: 10.1080/19485565.2005.9989104

Bowen, M. E. (2010). Coronary heart disease from a life-course approach: Findings from the health and retirement study, 1998-2004. *Journal of aging and health*, *22*(2), 219–241. doi: 10.1177/0898264309355981

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11–20.

Cai, T. T., Liang, T., & Zhou, H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, *137*, 161-172.

Cain, K. C., & Breslow, N. E. (1988). Logistic regression analysis and efficient design for two-stage studies. *American Journal of Epidemiology*, *128*(6), 1198–1206.

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

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

Childs, R. A., & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research and Evaluation*, *8*(16).

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330–351. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/11778676` doi: 10.1037/1082-989X.6.4.330

Costello, E. J., Angold, A., Burns, B. J., Stangl, D. K., Tweed, D. L., Erkanli, A., & Worthman, C. M. (1996). The great smoky mountains study of youth: Goals, design, methods, and the prevalence of dsm-iii-r disorders. *Archives of General Psychiatry*, *53*(12), 1129-1136. Retrieved from `+ http://dx.doi.org/10.1001/archpsyc.1996.01830120067012` doi: 10.1001/archpsyc.1996.01830120067012

Crimmins, E., Guyer, H., Langa, K., Ofstedal, M. B., Wallace, R., & Weir, D. (2008a). *Documentation of biomarkers in the Health and Retirement Study* (Report). Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Crimmins, E., Guyer, H., Langa, K., Ofstedal, M. B., Wallace, R., & Weir, D. (2008b). *Documentation of physical measures, anthropometrics and blood pressure in the Health and Retirement Study* (Report). Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Curtin, L. R., Mohadjer, L. K., Dohrmann, S. M., Montaquila, J. M., Kruszan-Moran, D., Mirel, L. B., ... Johnson, C. L. (2012). The national health and nutrition examination survey: Sample design, 1999-2006. *Vital Health Stat*, *2*(155), 1-39.

Dillman, D., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*, 289–304. Retrieved from `http://poq.oxfordjournals.org/content/57/3/289.abstract` doi: 10.1086/269376

Dong, Q., Elliott, M. R., & Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, *40*(1), 29–46.

Erkanli, A., Soyer, R., & Stangil, D. (1997). Bayesian inference in two-phase prevalence studies. *Statistics in Medicine*, *16*(10), 1121-1133.

Forster, J. J. (2010). Bayesian inference for Poisson and multinomial log-linear models. *Statistical Methodology*, *7*(3), 210-224.

Gao, S., Hui, S. L., Hall, K. S., & Hendrie, H. C. (2000). Estimating disease prevalence from two-phase surveys with nonresponse at the second phase. *Statistics in Medicine*, *19*(16), 2101-2114.

Gonzalez, J., & Eltinge, J. (2007). Multiple matrix sampling: A review. *Proceedings of the Section on Survey ...*(December), 3069–3075. Retrieved from http://www.amstat.org/sections/srms/Proceedings/y2007/Files/JSM2007-000494.pdf

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*, 197-218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. doi: 10.1037/1082-989X.11.4.323

Granville, V., Krivanek, M., & Rasson, J. (1994). Simulated annealing: A proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(6), 652-656.

Haan, M. N., Mungas, D. M., Gonzalez, H. M., Ortiz, T. a., Acharya, A., & Jagust, W. J. (2003). Prevalence of dementia in older Latinos: The influence of type 2 diabetes mellitus, stroke and genetic factors. *Journal of the American Geriatrics Society*, *51*, 169–177. doi: 10.1046/j.1532-5415.2003.51054.x

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC medical research methodology*, *12*, 184. Retrieved from http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-184 doi: 10.1186/1471-2288-12-184

Hartley, H., & Hocking, R. (1971). The Analysis of Incomplete Data. *Biometrics*, *27*(4), 783-823.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549-559.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods* (First ed.). New York: Springer.

Jorgensen, T. D., Rhemtulla, M., Schoemann, a., McPherson, B., Wu, W., & Little, T. D. (2014). Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies. *International Journal of Behavioral Development*, *38*, 397–410. Retrieved from http://jbd.sagepub.com/cgi/doi/10.1177/0165025414531094 doi: 10.1177/0165025414531094

Juster, F. T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, *30*, S7–S56. doi: 10.2307/146277

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, *41*(1), 57–80. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84955470010&partnerID=tZOtx3y1 doi: 10.3102/1076998615622221

Kirkpatrick, S., Gelat, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671-680.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Li, B., & Martin, E. B. (2002). An approximation to the F distribution using the chi-square distribution. *Computational Statistics & Data Analysis*, *40*(1), 21-26.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Littvay, L. (2009). Questionnaire design considerations with planned missing data. *Review of Psychology*, *16*(2), 103–113.

Mayeda, E. R., Haan, M. N., Kanaya, A. M., Yaffe, K., & Neuhaus, J. (2013). Type 2 diabetes and 10-year risk of dementia and cognitive impairment among older Mexican Americans. *Diabetes Care*, *36*(9), 2600–2606. doi: 10.2337/dc12-2158

McNamee, R. (2003). Efficiency of two-phase designs for prevalence estimation. *International Journal of Epidemiology*, *32*(6), 1072-1078. doi: 10.1093/ije/dyg230

McNamee, R. (2004). Two-phase sampling for simultaneous prevalence estimation and case detection. *Biometrics*, *60*(3), 783-792.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *33*(2), 1263-1282. doi: 10.1214/12-AOS1008

Murphy, K. P. (2007). *Conjugate bayesian analysis of the gaussian distribution* (Tech. Rep.).

Neyman, J. (1938). Contribution to the theory of sampling human populations. *The Annals of Statistics*, *33*(201), 101-116.

Patton, G. C., Coffey, C., Posterino, M., Carlin, J. B., & Wolfe, R. (2000). Adolescent depressive disorder: A population based study of ICD-10 symptoms. *Australian and New Zealand Journal of Psychiatry*, *34*(5), 741-747.

Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, *11*(4), 361-368.

Pierce, B. L., & Burgess, S. (2013). Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, *178*(7), 1177-1184.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical AssociationStatistical Association*, *90*(429), 54–63. doi: 10.1080/01621459.1995.10476488

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberg, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Raghunathan, T. E., Solenberger, P. W., & Van Hoewyk, J. (2002). IVEware : Imputation and variance estimation software user guide. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan*(March).

Rhemtulla, M., & Little, T. (2012). Tools of the trade: Planned missing data designs for research in cognitive development. *Journal of Cognition and Development : Official Journal of the Cognitive Development Society*, *13*(4). doi: 10.1080/15248372.2012.717340

Roszkowski, M. J., & Bean, A. G. (1990). Believe it or not! Longer questionnaires have lower response rates. *Journal of Business and Psychology*, *4*(4), 495–509. doi: 10.1007/BF01013611

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys : Experiments on question form, wording, and context* (First ed.). New York: Academic Press.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

Shoemaker, D. M., & Shoemaker, J. S. (1981). Applicability of multiple matrix sampling to estimating effectiveness of educational programs. *Evaluation and Program Planning*, *4*(2), 151 - 161. Retrieved from `http://www.sciencedirect.com/science/article/pii/0149718981900057` doi: https://doi.org/10.1016/0149-7189(81)90005-7

Shrout, P. E., & Newman, S. C. (1989). Design of two-phase prevalence surveys of rare disorders. *Biometrics*, *45*(2), 549-555.

Sitter, R. (1977). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, *92*(438), 780-787.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology* (First ed.). San Francisco: Jossey-Bass.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Survey Methodology*, *32*(2), 217–231.

Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *J Off Stat*, *32*(1), 231–256. doi: 10.1515/JOS-2016-0011