Leveraging Decision Theory to Address Statistical Challenges and Regression Under Additive Nonignorable Missingness

Chloe Krakauer

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Kenneth Rice, Chair

Mauricio Sadinle, Chair

Marco Carone

Program Authorized to Offer Degree:
Biostatistics-Public Health

University of Washington

Abstract

Leveraging Decision Theory to Address Statistical Challenges and Regression Under Additive Nonignorable Missingness

Chloe Krakauer

Co-Chairs of the Supervisory Committee:
Kenneth Rice
UW Biostatistics

Mauricio Sadinle
UW Biostatistics

This dissertation is composed of three mutually exclusive projects, presented in individual chapters. The first chapter summarizes a novel method for regression estimation under nonignorable missingness. The latter two chapters use decision theory to address two ubiquitous statistical challenges: *post hoc* test assessment and shrinkage.

Chapter 1: Regression Estimation Under Additive Nonignorable Missingness (advised by Mauricio Sadinle, PhD) Existing methods to estimate associations between outcomes subject to missingness and covariates often depend on the assumption that missingness is ignorable. Chapter 1 lays out a novel approach to estimation that allows for nonignorable missingness. Assumptions are limited to the missingness mechanism falling in a broadly-defined class and availability of estimates of marginal means of the outcome from auxiliary data sources or expert opinion. Due to Sadinle and Reiter [2019], such assumptions ensure identifiability of parameters indexing the regression function, which are estimable using the generalized method of moments in all but pathological scenarios.

Chapter 2: Decision-Theoretic Approach to *Post Hoc* Significance Test Assessment (advised by Kenneth Rice, PhD) Binary decisions about parameters of interest—comparing a p-value to a pre-specified field-standardized threshold in a significance test—are ubiquitous in statistical analysis. These binary decisions are appealingly simple, and in some areas, reflect the binary nature of how knowledge will be applied (e.g. allowing a medication to be sold). However, the reliability of such tests is often poorly understood, and the inability of p-values to provide information about reliability is widely overlooked. Moreover, existing measures designed to accompany significance tests to assess their reliability are prone to paradoxical results, misinterpretation, dependence on further arbitrary thresholds, and/or have failed to gain popularity. As an alternative, we consider Bayesian decision-theoretic analogs to one- and two-sided significance tests, which may approximate frequentist Type I error rates and inference when desired [Rice et al., 2020]. Loss and risk arise naturally as measures of test reliability from the motivation of the test itself; in Chapter 2, we study feasibility, accuracy, precision, and usefulness of estimates of these metrics.

Chapter 3: Decision-Theoretic Justification for Winner's Curse Reduction (advised by Kenneth Rice, PhD) Default estimators often fail to retain their ideal properties (e.g. unbiasedness, admissibility, etc.) when used in non-regular settings. A classic example is the Winner's Curse bias, or upward bias in estimates when reports are limited to significant results. Improvements to default estimators, such as Zhong and Prentice [2008]'s amelioration of the Winner's Curse bias, often distill to *shrinkage*, or the movement of a complex decision to a simple or parsimonious one. Motivating estimators such as these using a single family of loss functions in a decision-theoretic framework allows for (1) intuitive understanding of the estimator, (2) incorporation of prior information, and (3) a natural framework to compare seemingly dissimilar estimators. We introduce such a family of loss functions in Chapter 3. Our particular focus is on decision-theoretic justification for Zhong

and Prentice [2008]'s estimators, their ability to reduce the Winner's Curse bias compared to other estimators in this newly-defined class, and how prior skepticism affects bias reduction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I owe thanks to countless individuals for their guidance and support throughout the doctoral process, but I will do my best to be concise. This dissertation will be completed in the midst of a collectively tumultuous time. The continuous support I received from all the following individuals during this time of upheaval deserves elevated thanks.

First, I would like to thank my advisor Kenneth Rice, PhD for his continued guidance, both through our research and the world of academia. His advising provided the opportunity for me to stretch my intellectual and philosophical muscles to extents I didn't previously think possible. I thank him profusely for his continued enthusiasm and patience.

Thanks to my advisor Mauricio Sadinle, PhD who stood by this project through a number of iterations and offered advice on writing I will take with me. I am particularly grateful for Dr. Sadinle's encouragement to pursue two research areas within the same dissertation.

A special thank you to Mary Lou Thompson, PhD, who guided me through my master's, PhD, and first professional position as my academic advisor and boss for seven years. Thank you as well to all faculty members who took the time to meet with me to discuss this research, including Drs. Marco Carone, Lianne Sheppard, Robyn McClelland, and Patrick Heagerty. The access I have been afforded to preeminent experts in numerous fields within this department is something I do not take for granted.

Gitana Garofalo was the first member of the department I spoke with, and has provided support on countless occasions since, for which I am eternally grateful. Thank you as well to my GSR, Alison Fohner, PhD for her support.

This degree would not have been possible without friends and family to whom I often turned to for in-person and virtual support. I can't wait to (safely) thank you in person.

DEDICATION

To my cheerleaders: my parents Patti and Joel, and my sister Hannah.

Chapter 1

REGRESSION UNDER ADDITIVE NONIGNORABLE MISSINGNESS

Abstract

Scientific questions are often answerable by estimating the association between some baseline covariates and outcome collected after a period of follow-up; while the former are often completely observed, the latter are inevitably subject to missingness. Estimating the regression without accounting for the joint behavior of the response indicator and the outcome may lead to bias. Default methods assume missingness is "ignorable", despite the ubiquity of situations when this assumption is not scientifically plausible. The approach introduced in this chapter is applicable to nonignorable missingness, so long as the model for missingness falls into a broad class defined by Sadinle and Reiter [2019]. We build the final regression function by using this model for missingness to relate the dependence of unobserved outcomes on baseline covariates to the dependence of observed outcomes on baseline covariates. The final regression model is governed by parameters indexing the model for missingness. Thanks to findings by Sadinle and Reiter [2019], identifiability of these parameters is assured, given estimates of the marginal means of some functions of the outcome in the target population. The use of this class of missingness models and dependence on readily available estimates of marginal means is novel and very flexible. Parameters indexing the outcome's mean model's dependence on baseline covariates may be estimated using the generalized method of moments. We illustrate the methodology using a simulation study and a re-assessment of an analysis by Blasi et al. [2018] assessing characteristics associated with seeking professional dental care among low-income smokers.

1.1 Introduction

Applied research questions can often be answered using *regression*—by which we mean, studying how the expected value of an outcome depends on specific covariates. This method is relatively formulaic with *fully observed* covariate and outcome data. Typically, an investigator starts by defining the functional form of the regression. Then, the parameters indexing the regression may be estimated using a variety of default approaches, e.g. solving estimating equations that may depend on specifying other features of the outcome distribution, such as the variance in quasi-likelihood. With these estimates in hand, necessary assumptions about the outcome distribution for the estimation technique may be verified, typically including some verification and/or building of the regression model itself.

However, particularly with data from human subjects there is ubiquitously *missing data*, especially for studies with covariates collected at baseline and the outcome collected after a period of follow-up. Baseline data are commonly essentially fully-collected among participants, with loss to follow-up causing most missingness in the outcome. With missing outcome data, the functional form of the regression model cannot be empirically verified, requiring some strong assumptions about how the missingness arises to define it.

Typically, data are assumed to be *missing at random*, so that the missingness mechanism is *ignorable* [Little and Rubin, 2002, pg 119-120]. If, after accounting for baseline covariates, the probability of being lost to follow-up depends on the potentially missing outcome—i.e. if *missingness is not at random*—then estimation assuming missingness at random will be biased. Missingness will generally not be ignorable when the outcome is socially unacceptable, such as alcohol or drug use [Rotnitzky and Robins, 1997], or when the outcome itself leads to lack of follow-up, such as measures of self-efficacy or motivation. The development of methods allowing for missingness not at random—a.k.a. *nonignorable missingness*—is therefore critical.

Generally, under nonignorable missingness, we must consider how missingness impacts the outcome or vice versa, and therefore model joint behavior of the outcome and response

indicator. This joint distribution is not identifiable without some strong assumptions, which are often made about the *missingness mechanism*: a model for how the probability of being lost to follow-up depends on baseline covariates and the potentially-missing outcome [Daniels and Hogan, 2008, pg 90]. Parameters indexing the missingness mechanism may either directly govern the final regression model or govern the likelihood of parameters indexing the final regression model.

As a consequence, the form of the missingness mechanism dictates if parameters indexing the final regression model are identifiable using observed data. Therefore, existing methods are developed for specific functional dependence of the missingness mechanism on baseline covariates and outcomes, limiting true missingness mechanisms to which existing methods apply. Among applicable missingness mechanisms, the chosen modeling technique dictates how much of the functional form of the missingness mechanism must be specified, and which parameters indexing the missingness mechanism must be fixed to ensure identifiability. Some methods completely fix functional dependence on the outcome and baseline covariates, estimating parameters using observed data [Rotnitzky and Robins, 1997]. Other methods do not require specification of the functional dependence on baseline covariates, but require completely-specified functional dependence of the missingness mechanism on outcomes, governed by fixed parameter values [Scharfstein et al., 1999, Shardell et al., 2013]. These approaches are quite restrictive, especially given lack of empirical verifiability of the missingness mechanism.

The most flexible approach fixes only the functional dependence of the missingness mechanism on the outcome indexed by unknown parameters, e.g. the probability of response is linear in the outcome on the logit scale. In this case, identifiability of parameters indexing the regression function requires some other information beyond that obtained from the study variables at hand. Previously-developed methods use outcome data collected from initial non-respondents [Kim and Yu, 2011] or additional covariates collected without missingness [Shao and Wang, 2016]. Some existing methods also use estimates of parameters indexing the dependence of the missingness mechanism on the outcome from an independent

sample [Tang et al., 2014, Zhao et al., 2013, Kim and Yu, 2011]. Acquiring these resources requires planning before the main study and may not be feasible. Additionally, the aforementioned methods (with the exception of methods developed by Shao and Wang [2016]) have been developed specifically for logit-linear dependence of the missingness mechanism on the outcome, limiting true missingness mechanisms to which the method is applicable; even on the logit scale, expanding how the missingness mechanism may depend on the outcome (e.g. quadratic) would expand suitable applications.

We propose a novel approach that retains the flexibility of only specifying the functional dependence of the missingness mechanism on the outcome, indexed by unknown parameters, supposing additivity in baseline covariates and the outcome on a given scale. The additional data we use to ensure identifiability are estimates of marginal means of some functions of the outcome in the target population. Collecting these estimates is often feasible from expert opinion, administrative databases, or surveys. Availability of these marginal means dictates missingness mechanisms to which the methodology applies. Hence, in theory, the missingness mechanism may have any functional dependence on the outcome.

This method also provides an interpretable framework for *tipping point* analyses: we can evaluate how extreme marginal means must be to change inference in the regression model. Tipping point analyses using this novel approach and existing methods can also show how strongly the missingness mechanism must depend on the outcome for inference to change. However, for continuous outcomes, defining a range for the difference in probability of missingness per unit difference in outcome can be difficult; finding a reasonable range for marginal means of the outcome, the quantity that may be directly manipulated using our method, may be simpler.

Our (novel) method has two stages. First, we specify the joint distribution of the outcome and indicator of missingness using the missingness mechanism. We will suppose the missingness mechanism falls in the class defined by Sadinle and Reiter [2019], ensuring parameters indexing the joint distribution of the response indicator and outcome (and therefore the regression function) are identifiable given availability of marginal means of the outcome. As a

consequence of defining the joint distribution, the functional form of the regression function of interest is also specified, which we use as the framework for our approach. The regression function may be decomposed into empirically verifiable and unverifiable components. Given functional forms of the former, the functional form of the latter is identified using the missingness mechanism. Second, with the functional form of the regression function in hand, we estimate parameters indexing the regression function using the Generalized Method of Moments (GMM), ensuring consistent and asymptotically-normal estimates of parameters indexing the regression function and values of the regression function itself.

The text is structured as follows. After introducing some notation and preliminaries in Section 1.2, we will summarize existing methods in further detail in Section 2.2. Our approach to jointly model the response indicator and outcome is described in Section 1.4, starting with a detailed discussion of the properties of joint distributions identified using the class of missingness mechanisms defined by Sadinle and Reiter [2019] in Section 1.4.1. We then consider empirically verifiable components of the regression function in Sections 1.4.2.1 and 1.4.2.2, to be used (along with the missingness mechanism) to identify unverifiable components in Section 1.4.2.3. Section 1.5 outlines use of GMM to estimate parameters indexing the newly-identified regression function. The novel methodology is then applied to a simulation study with Poisson-distributed outcomes and a continuous covariate in Section 1.6. We also used the methodology to assess the sensitivity of findings due to Blasi et al. [2018] to the assumed missingness mechanisms; Blasi et al. [2018] used a single-imputation procedure when finding determinants of seeking professional dental care among low-income smokers. The chapter is completed with a summary of future research directions.

1.2 Notation and Preliminaries

Let X represent the vector of baseline covariates, which we assume to be collected without missingness, and Y the outcome that is subject to missingness. Unless stated otherwise, we suppose the regressions have an intercept, i.e. that the first element of X, and any named subsets of X, is 1. The *response indicator* is denoted R, with $R = 0/1$ when Y is

missing/observed respectively.

Our aim is to estimate the regression function $\mathbb{E}[Y \mid X = x]$ indexed by parameters θ; estimation of θ and inference on the regression function evaluated at a given $X = x$ will be discussed separately. We will define multiple conditional expected values and densities throughout, for which the name of the random variable being conditioned on will be dropped if the meaning remains clear, e.g. $\mathbb{E}[Y \mid X = x]$ will be denoted by $\mathbb{E}[Y \mid x]$. Rather than defining separate functions for each density, we use the common notation of $f(\cdot)$ for any density, defined specifically for random variables passed to it as arguments. For example, we denote the conditional density of $[Y \mid x]$ by $f(y \mid x)$.

We will occasionally define expected values and densities with parameter values indexing them, either for clarity or to denote the value of a parameter to use when evaluating the density or expected value. These parameters will be placed after semicolons. For instance, if the distribution of $[Y \mid x]$ is indexed by parameters θ, then the density may be written as $f(y \mid x; \theta)$. If the function is to be evaluated at a specific value of θ, e.g. an estimate $\hat{\theta}$ of θ, rather than writing $f(y \mid x; \theta = \hat{\theta})$, we write $f(y \mid x; \hat{\theta})$ when the use of such parameters is clear. With expected values, we reserve subscripts to denote which variables to take the expected value with respect to for the sake of clarity. For instance, the law of iterated expectation would be denoted by $\mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y \mid X]]$.

In addition to baseline covariates and the outcome, response is also a random variable. Denote complete data for individual $i \in \{1, ..., n\}$ by (X_i, Y_i, R_i). We suppose $(X_1, Y_1, R_1), ..., (X_n, Y_n, R_n)$ are independent and identically distributed (i.i.d). With i.i.d. data, we will denote the joint density of complete data for an arbitrary individual (X, Y, R) by $f(x, y, r)$; the joint density evaluated for specific individual i will be denoted by $f(x_i, y_i, r_i)$. Since the target of $\mathbb{E}[Y \mid x]$ depends on fixed values of X, existing methods often focus specifically on jointly modeling Y and R conditional on X, i.e. $f(y, r \mid x)$ instead of $f(y, r, x)$. However, with the density of $f(x)$ (which depends on complete data only), recovering density $f(y, r, x)$ with the definition of $f(y, r \mid x)$ is trivial. The complete data density of individual

i is

$$f(y_i, x_i, r_i) = f(y_i, r_i \mid x_i) f(x_i).$$

With iid data, this density is the contribution of individual i to the complete data likelihood of all parameters indexing the distribution of (X, Y, R), so that the contribution to the log of the complete data likelihood for the same parameters is

$$\log f(y_i, r_i \mid x_i) + \log f(x_i).$$

When $R_i = 1$, the *observed data* for individual i is $(X_i, Y_i, R_i = 1)$. When $R_i = 0$, observed data are limited to $(X_i, R_i = 0)$. The observed data density may generally be derived by integrating $f(y, x, r)$ with respect to missing outcomes Y. More specifically, using some factorizations, the contribution of individual i to the observed data likelihood of parameters indexing the observed data distribution is

$$\begin{aligned} & f(y_i, r_i, x_i)^{r_i} f(x_i, r_i)^{1-r_i} \\ =& \left[f(y_i, r_i \mid x_i) f(x_i)\right]^{r_i} \left[f(r_i \mid x_i) f(x_i)\right]^{1-r_i} \\ =& f(x_i) \left[f(y_i, r_i \mid x_i)^{r_i} f(r_i \mid x_i)^{1-r_i}\right], \end{aligned}$$

so that the contribution to the log of the observed data likelihood is

$$\log f(x_i) + r_i \log f(y_i, r_i \mid x_i) + (1 - r_i) \log f(r_i \mid x_i). \tag{1.1}$$

The conditional density $f(r \mid x, y)$—called the *missingness mechanism* [Daniels and Hogan, 2008, pg 90] or *selection model*—defines how the probability of missingness depends on both baseline data X and the potentially-missing outcome Y. We suppose this model is indexed by parameters $\boldsymbol{\beta}$. Due to dependence on Y, functional forms of this density are not empirically verifiable. Assumptions about the functional form may be categorized using familiar names. If data are *missing completely at random* (MCAR), then $f(r \mid x, y)$ depends on neither

X nor Y; if data are *missing at random* (MAR), then $f(r \mid x, y)$ may depend on X, but not on Y; finally, if data are *missing not at random* (MNAR), then $f(r \mid x, y)$ depends on outcome Y and may also depend on X. (Hand [2020, pg 227] re-named these assumptions for easier interpretability, which are *not data dependent* (NDD), *seen data dependent* (SDD), and unseen data dependent (UDD) respectively.) The final assumption of MNAR—a.k.a. unseen data dependent—is the most flexible and the most difficult to account for.

1.2.1 Full Data Density Factorizations

Since our aim is to estimate some subset of parameters indexing $f(y \mid x)$, which does not appear to depend at all on the response indicator, it is natural to question why the joint distribution of the response and outcome is needed. Two factorizations of the complete data density—selection model factorization and pattern-mixture modeling—may make this more intuitive. Each factorization requires a set of assumptions to jointly model the response indicator and outcome, thus providing useful categories for existing methods, to be discussed in more detail in Section 2.2. Recalling the complete data density $f(r, y, x)$ is equal to $f(r, y \mid x)f(x)$, factorizations discussed below will factorize $f(r, y \mid x)$, supposing the initial factorization of the complete data density as $f(r, y \mid x)f(x)$ has been completed.

Using the *selection model factorization*, the complete density density $f(r, y \mid x)f(x)$ is factorized as $f(r \mid y, x)f(y \mid x)f(x)$, which depends on the *selection model* $f(r \mid x, y)$ and the density of interest $f(y \mid x)$ [Little and Rubin, 2002, pg 313]. With this factorization, $f(y \mid x)$ is directly specified, which cannot be empirically verified using the data since $[Y \mid x]$ is not fully observed. Suppose we wish to estimate parameters θ indexing $f(y \mid x)$ using a fully parametric approach with observed data. As we will illustrate, the density of observed data depends on θ, and hence the likelihood of parameters indexing this density does define the likelihood for θ using all available data. Also recall the complete data likelihood depends on missing data, and therefore cannot be used directly. Applying the factorization to the log of the observed data likelihood in Equation (1.1), the contribution of individual i to the

observed data log likelihood is

$$\begin{aligned}
&\log f(x_i) + r_i \log f(y_i, r_i \mid x_i) + (1 - r_i) \log f(r_i \mid x_i) \\
&= \log f(x_i) + \underbrace{r_i \log \left(f(r_i \mid y_i, x_i; \boldsymbol{\beta}) f(y_i \mid x_i; \theta) \right)}_{\text{observed } Y} + \underbrace{(1 - r_i) \log f(r_i \mid x_i)}_{\text{unobserved } Y} \\
&= \log f(x_i) + r_i \log f(r_i \mid y_i, x_i; \boldsymbol{\beta}) + r_i \log f(y_i \mid x_i; \theta) \qquad (1.2) \\
&\qquad + (1 - r_i) \log f(r_i \mid x_i), \text{ where} \\
&f(r_i \mid x_i; \boldsymbol{\beta}, \theta) = \int_{y_i} f(r_i \mid y_i, x_i; \boldsymbol{\beta}) f(y_i \mid x_i; \theta) dy_i. \qquad (1.3)
\end{aligned}$$

Note that $r_i \log f(r_i \mid y_i, x_i; \boldsymbol{\beta})$ will not contribute to the log likelihood of θ if θ and $\boldsymbol{\beta}$ have an empty intersection. Hence, defining the likelihood of θ using only individuals with observed outcomes Y is equivalent to using only $f(y_i \mid x_i; \theta)$ to define the likelihood. However, the integration in Equation (1.3) shows $f(r_i \mid x_i)$ may be governed by both θ and $\boldsymbol{\beta}$, and hence $f(r_i \mid x_i)$ contributes to the observed data log likelihood of θ. Defining the likelihood of θ using only individuals with observed outcome (i.e. when $R_i = 1$) is tantamount to throwing away terms $f(r_i \mid x_i)$ from Equation (1.2), and hence is not the true likelihood of θ.

With *pattern-mixture modeling*, the joint density $f(r, y \mid x) f(x)$ is factorized to be [Little and Rubin, 2002, pg 313]

$$f(y \mid r, x) f(r \mid x) f(x).$$

The contribution of individuals to the log of the observed data likelihood in Equation (1.1) is written as

$$r_i \log f(y_i \mid R_i = 1, x_i) + \log f(r_i \mid x_i) + \log f(x_i). \qquad (1.4)$$

The three components of the pattern-mixture modeling factorization are $f(y \mid x, R = r)$ (i.e. one defined for each $R \in \{0, 1\}$), $f(r \mid x)$, and $f(x)$. Note that using these component parts, we may define $f(y \mid x) = \sum_{r \in \{0,1\}} f(y \mid r, x) f(r \mid x)$. Hence $f(y \mid x)$, and therefore $\mathbb{E}[Y \mid x]$, are indexed by (potentially unique) parameters indexing *observable* $f(r \mid x)$ and

$f(y \mid x, R = 1)$ and *unobservable* $f(y \mid x, R = 0)$; hence, we can not define the likelihood of all parameters indexing $f(y \mid x)$ using observed data only. Parameters indexing the latter unobservable density must be fixed, related to other elements of θ, or estimated by some other means. Like with the other parameterization, θ potentially includes parameters indexing $f(r \mid x)$. We also need to separately define $f(y \mid x, R = 1)$ and $f(y \mid x, R = 0)$, and therefore consider how the distribution of Y changes depending on response status.

We will use pattern mixture modeling in our novel methodology. We do define the component parts of the factorization, but focus our discussion on marginalizing the factorized density with respect to $[R \mid x]$ to recover $f(y \mid x)$. Hence, we have no need to define $f(x)$, but it is easily recoverable. We name the individual components of $f(y \mid x)$ as

$$f(y \mid x) = \underbrace{\mathbb{P}[R = 1 \mid x]}_{\text{CDRP}} \underbrace{f(y \mid x, R = 1)}_{\text{Verifiable density}} + \underbrace{\mathbb{P}[R = 0 \mid x]}_{\text{1-CDRP}} \underbrace{f(y \mid x, R = 0)}_{\text{Unverifiable density}}, \tag{1.5}$$

where CDRP stands for *covariate-dependent response probability*. We also name individual components of the regression function as

$$\mathbb{E}[Y \mid x] = \underbrace{\mathbb{P}[R = 1 \mid x]}_{\text{CDRP}} \underbrace{\mathbb{E}[Y \mid x, R = 1]}_{\text{observable regression}} + \underbrace{\mathbb{P}[R = 0 \mid x]}_{\text{1}-\text{CDRP}} \underbrace{\mathbb{E}[Y \mid x, R = 0]}_{\text{unobservable regression}} . \tag{1.6}$$

Since R and X are fully observed, the CDRP is empirically verifiable. The *observable regression* from the *recoverable density* $f(y \mid x, R = 1)$ describe behavior of fully-observed realizations of Y, and hence is empirically-verifiable. We observe no realizations of $[Y \mid x, R = 0]$; therefore, neither the *unobservable regression* nor the *unverifiable density* are empirically verifiable using observed data. Without some further assumptions, we say $f(y \mid x)$ is *not identifiable*. We use the definition of *not identifiable* in this context as used by Sadinle and Reiter [2019], who note that even with infinite data, only the density of observed data is recoverable.

Our novel approach focuses on one method of defining $f(y \mid x, R = 0)$ via the selection model, which is often used in the literature. Suppose $f(y \mid x, R = 1)$ is known and the

selection model $f(r \mid x, y)$ is specified. The latter may be transformed to be

$$\text{logit}(\mathbb{P}[R = 0 \mid x, y]) = b(x, y) \tag{1.7}$$

without loss of generality [Rotnitzky and Robins, 1997]. The unverifiable density $f(y \mid x, R = 0)$ relates to the verifiable density $f(y \mid x, R = 1)$ via [Kim and Yu, 2011]:

$$f(y \mid x, R = 0) = f(y \mid x, R = 1)\frac{\exp\{b(x, y)\}}{\mathbb{E}[\exp\{b(x, Y)\} \mid x, R = 1]}. \tag{1.8}$$

1.3 Existing Methods

In some cases, we need not worry about the specific form of the selection model, a classic example being ignorable missingness with a likelihood-based approach to inference. Estimation in these cases is discussed in Section 1.3.1. Then, we consider estimation when the missingness mechanism cannot be neglected in Section 1.3.2. In Section 1.3.2.1, we discuss identifiability of parameters θ—which is of particular concern with nonignorable missingness—and definitions of relevant terms. Finally, we summarize existing methods to estimate θ using selection model factorization and pattern-mixture model factorization in Sections 1.3.2.2 and 1.3.2.3, respectively.

1.3.1 When Missingness May Be Neglected

As previously mentioned, the missingness mechanism may be *neglected* under certain conditions when estimating parameters θ. One of the common set of assumptions is *ignorability*, defined to be (1) missingness at random (MAR) and (2) the empty intersection of parameters indexing $f(y \mid x)$ and $\boldsymbol{\beta}$ [Little and Rubin, 2002, pg 119-120]. *Ignorability* is an appropriate moniker because the contribution of an individual to the log of the observed data likelihood

(using the selection model factorization in Equation (1.2)) under these conditions is now

$$\begin{aligned}
&\log f(x_i) + r_i \log f(r_i \mid y_i, x_i; \boldsymbol{\beta}) + r_i \log f(y_i \mid x_i; \theta) + (1 - r_i) \log f(r_i \mid x_i) \\
=&\log f(x_i) + r_i \log f(r_i \mid x_i; \boldsymbol{\beta}) + r_i \log f(y_i \mid x_i; \theta) + (1 - r_i) \log f(r_i \mid x_i; \boldsymbol{\beta}) \\
=&\log f(x_i) + r_i \log f(y_i \mid x_i; \theta) + \log f(r_i \mid x_i; \boldsymbol{\beta}),
\end{aligned}$$

recalling that under ignorability, $f(r \mid x, y) = f(r \mid x)$, which cannot depend on θ. Therefore, the partial derivative of the log of the observed data likelihood with respect to θ depends only on θ and *not* on any part of the response mechanism. Hence, when using a fully parametric approach, the response mechanism need not be defined nor estimated to acquire the MLE for θ.

This justification may be simpler using the pattern mixture modeling factorization. The identity in Equation (1.8) implies $f(y \mid x, R = 1) = f(y \mid x, R = 0)$, and noting $f(y \mid x) = \sum_{r \in \{0,1\}} f(y \mid r, x) f(r \mid x)$, we then have $f(y \mid x) = f(y \mid x, R = 1)$. Therefore, the observed data likelihood in Equation (1.4) defines the likelihood for all parameters indexing the selection model. Furthermore, the MLE for θ using the observed data likelihood depends only on the verifiable density evaluated among individuals with observed outcomes Y.

Ignorability directly pertains to the ability to ignore the missingness mechanism with a likelihood-based approach to inference. However—in some cases—we may require specification of the selection model under MAR to avoid bias, e.g. when auxiliary variables excluded from the regression function must be adjusted for in the selection model to make the MAR assumption tenable. However, recall the selection model depends on fully observed data, and hence is empirically-verifiable. In these cases, parameters indexing the regression function may be efficiently estimated using augmented inverse probability weighting [Robins et al., 1995]. Also note that if even if complete-case analyses are appropriate, estimates using the observed data likelihood may be substantially less efficient than those acquired using complete data; we may wish to leverage available data from individuals with missing outcomes to reduce this discrepancy, e.g. using carefully-implemented multiple imputation. (Davidian

[2017] has a helpful summary of this procedure.)

As previously discussed in Section 1.1, missingness at random is sometimes scientifically implausible and quite restrictive. Additionally, as noted in Section 1.2, this assumption is not testable using the data. Clearly, assuming ignorability when it does not exist will lead to incorrect specification of the observed data likelihood and will in general lead to bias.

Some methods do exist that maintain the ease of ignorability, while relaxing the assumption of missingness at random. Ibrahim et al. [2001] reduce the bias normally incurred by inappropriately assuming ignorability by using a set of auxiliary variables V collected without missingness that (1) are correlated with Y conditional on X and (2) have an empty intersection with X. Surrogate endpoints, e.g. CD4 counts for symptoms of AIDS or X-ray images in place of magnetic resonance images for cancer outcomes, are suitable choices. These auxiliary variables should be considered determinants of response, such that the selection model $f(r \mid y, x, v)$ has non-zero dependence on V. The new full data density may be factorized as

$$f(y, r, v, x) = f(r \mid y, x, v) f(v \mid y, x) f(y \mid x) f(x).$$

Ibrahim et al. [2001] suppose there exists sufficient absolute correlation between Y and V conditional on X such that $f(r \mid y, x, v) \approx f(r \mid x, v)$. If the equality does hold, missingness is at random. This replacement in the likelihood results in partial derivatives with respect to θ that do not depend on $\boldsymbol{\beta}$; however, the density $f(v \mid y, x)$ must be specified and the parameters indexing it estimated in most applications. The reduction in bias from assuming $f(r \mid y, x, v) = f(r \mid x, v)$ has been empirically shown to increase monotonically with absolute correlation between V and Y. Note that unbiased estimation of θ is not feasible unless (1) $f(r \mid y, x, v)$ does not depend on Y, in which case MAR holds or (2) Y and V are perfectly (absolutely) correlated, rendering collection of outcome Y meaningless. Note the association between V and Y must also be estimable using observed outcomes Y, despite missingness not being ignorable for Y.

Definition of the density $f(y \mid x)$ may also render the specification of the selection model obsolete. For instance, Luo and Tsai [2012] defined the semiparametric model

$$f(y \mid x) = \frac{\exp(\theta^T xy)g(y)}{\int \exp(\theta^T xy)dG(y)}$$

for unspecified baseline distribution function $G(y)$ with related density $g(y)$. Note that for arbitrary values y_1 and y_2,

$$\log \frac{f(y_2 \mid x)}{f(y_1 \mid x)} = \log \frac{g(y_2)}{g(y_1)} + \theta^T x(y_2 - y_1).$$

The target of estimation is still θ, defining a regression with a likelihood ratio as the outcome, and values X and contrast in Y as the covariates. We suppose the "intercept" term $\log \frac{g(y_2)}{g(y_1)}$ is not of interest. By supposing Y is binary, we can see θ is a generalization of the log odds ratio. Chan [2013] assumes the true selection model falls into the broad class of $\mathbb{P}[R = 1|x, y] = b_1(x)b_2(y)$ for some functions $b_1(\cdot)$ and $b_2(\cdot)$, thereby allowing for nonignorable missingness. Under this assumption,

$$f(y \mid x, R = 1) = \frac{\exp(\theta^T xy)g_1(y)}{\int \exp(\theta^T xy)dG_1(y)}$$

for a *new* distribution $G_1(y)$ and related density $g_1(y)$. Chan [2013] is able to estimate θ without defining or estimating $G_1(y)$, leading to unbiased estimates of θ for any true missingness mechanism in this class using only observed outcome data. Therefore, the missingness mechanism may be neglected.

1.3.2 When Missingness Cannot Be Neglected

If conditions in Section 1.3.1 are not met, we must consider the missingness mechanism when estimating θ. First, we further explore the issue of identifiability of θ under these circumstances and how it can be ensured in Section 1.3.2.1. These considerations drive modeling choices in existing methods discussed in Sections 1.3.2.2-1.3.2.3.

1.3.2.1 Identifiability of Parameters

When missingness cannot be neglected, we must account for the joint behavior of R and Y as discussed in Section 1.2.1. The full data density is not identifiable using observed data alone; some *identifying assumptions*—unverifiable assumptions about the behavior of data—must be made to define the joint distribution. Identifying assumptions are commonly made about the selection model, ranging from semiparametric specification of the functional form of the mean model to a completely fixed function for the mean model.

A useful benchmark for assumptions made about the full data density is *nonparametric identifiability* (also known as *just identifiability*) [Daniels and Hogan, 2008]. This quality ensures a number of useful properties, to be described in more detail later. Most importantly, it makes parameters indexing the full data density (and therefore the regression function) identifiable. Furthermore, differences in inference will be due entirely to differences in assumptions about missingness, and not lack of corroboration of the implied observed data distribution with observed data.

Definitions and interpretation of defined terms through the remainder of this section are due mostly to Sadinle and Reiter [2019]. *Nonparametric identifiability* is the culmination of increasingly strict qualities of classes of full data distributions; the simplest will be defined first. A class of complete data distributions C_Ω indexed by parameter space Ω is *full-data identifiable* if there is a bijection between Ω and C_Ω, i.e. the class of complete data distributions is properly parameterized.

Let $\text{obs}(C_\Omega)$ denote the class of observed data distributions implied by C_Ω, recalling the observed data density may be derived by integrating the complete data density with respect to missing outcomes Y. For *identifiability* for class C_Ω, we further require a bijection between $\text{obs}(C_\Omega)$ and C_Ω, i.e. there is a unique way to go back and forth between the observed data distribution and complete data distribution. Hence, the observed data density is properly parameterized as well.

Assumptions used to construct the full data distribution may lead to a restricted class

obs(C_Ω). Under *nonparametric identifiability*, obs(C_Ω) must be parametrically unrestricted, and we can think of the class of full data distributions being indexed by the set of all observed data distributions. To contextualize this definition, suppose the observed data distribution (assumed known) and some identifying assumptions have been used to define the complete data distribution. If it is *just identified*, integrating the complete data density with respect to missing Y will yield the known observed data density, ensuring complete corroboration of the complete data density with observed data behavior. With assured reconstruction of the known observed data density, the identifying assumption used to define any nonparametrically identified class of complete data distributions cannot be rejected using observed data alone. Finally, as mentioned previously, the ability to construct complete data distributions that are *observationally equivalent*, i.e. have the same implied observed data distribution, ensures differences in inference for nonparametrically identified classes are due entirely to differing missing data assumptions.

Note that despite having sufficient assumptions to *just identify* the full data distribution, the functional form may not be completely specified in practice. Knowledge that the full data distribution is just identified for any arbitrary observed data density is sufficient to ensure identifiability of parameters indexing it. Furthermore, the identifying assumptions may be semiparametrically specified and still yield just-identified full data distributions; examples will be illustrated below. With functional forms of identifying assumptions fixed, θ may remain non-identifiable until some parameters indexing the identifying assumptions are fixed; if this is the case, *sensitivity analyses* should be conducted across a range of values (Council et al. [2010, pg 5], Scharfstein et al. [1999], Little and Rubin [2002, pg 314]).

Finally, it is worth noting nonparametric identifiability of $f(r, y, x)$ ensures identifiability of *all* parameters indexing the density. However, our targets of estimation are parameters θ indexing $\mathbb{E}[Y \mid x]$, with the remainder being nuisance parameters. Therefore, nonparametric identifiability of the full data distribution is sometimes more strict than necessary; see for example how Rotnitzky and Robins [1997] determine if regular estimates of θ exist, when the same does not hold for parameters $\boldsymbol{\beta}$ indexing the selection model.

Choosing identifying assumptions ensuring θ is identifiable is a daunting task. However, published methods offer guidance on classes of identifying assumptions fulfilling this requirement. We focus on the use of selection models as identifying assumptions, first within the selection model factorization, and finally to define the relationship between $f(y \mid x, R = 0)$ and $f(y \mid x, R = 1)$ using Equation (1.8) via pattern-mixture modeling.

1.3.2.2 Selection Model Factorization

Two components of the selection model factorization are not empirically verifiable: $f(y \mid x)$ and $f(r \mid x, y)$. First, consider specification of $f(y \mid x)$. In practice, if the target of estimation is $\mathbb{E}[Y \mid x]$, specifying the functional form of the regression function may be sufficient. We focus on this latter semi-parametric case, in the sense that neither the distribution of X nor the distribution of the errors $\epsilon = y - \mathbb{E}[Y \mid x]$ are specified [Rotnitzky and Robins, 1997]. Published results offer guidance on two methods for defining the selection model, the remaining component of the factorization to be defined, to ensure parameters indexing $\mathbb{E}[Y \mid x]$ are identifiable.

The first method entails specifying the complete functional form of the selection model that seems scientifically plausible, indexed by unknown parameters, checking later for estimability of parameters indexing $\mathbb{E}[Y \mid x]$. Rotnitzky and Robins [1997] use the augmented inverse probability weighted estimator achieving the semi-parametric variance bound for the specified model to check for feasibility of acquiring consistent and asymptotically normal (CAN) estimates of θ; CAN estimators exist if and only if the semi-parametric variance bound is finite. Rotnitzky and Robins [1997] offer specific sufficient and necessary conditions for lack of finiteness of this bound, and hence nonexistence of CAN estimators.

The second method entails choosing from a class of selection models known to yield just-identified complete data distributions. For example, Rotnitzky et al. [1998] note that any selection model under certain regularity conditions that has a completely-specified functional form and fixed parameters indexing the functional dependence of missingness on outcome Y ensures identifiability of θ.

If a CAN estimator is found to exist, Rotnitzky and Robins [1997] recommend using the augmented inverse probability weighted estimator reaching the semi-parametric variance bound. The estimator is specifically defined for missing outcome data and completely-observed covariate data.

1.3.2.3 Pattern-Mixture Model Factorization

Recall that under pattern-mixture modeling factorization, the only component of the factorization that is not empirically verifiable is $f(y \mid x, R = 0)$. This density may be specified a number of ways. Below, we review literature defining $f(y \mid x, R = 0)$ by relating it to empirically-verifiable $f(y \mid x, R = 1)$ using the selection model (see Equation (1.8)). Note that parameters indexing $f(y \mid x, R = 1)$ and $f(r \mid x)$ are identifiable using observed data if the densities are properly parameterized. Therefore, all parameters indexing $\mathbb{E}[Y \mid x] = \sum_{r \in \{0,1\}} \mathbb{E}[Y \mid r, x] f(r \mid x)$ are identifiable if parameters indexing the functional relationship between $f(y \mid x, R = 1)$ and $f(y \mid x, R = 0)$, which also index the selection model in this case, are fixed or identifiable.

The functional form of the selection model may be fixed to ensure some parametric form of $f(y \mid x, R = 0)$ or chosen to align with scientific understanding of the missingness mechanism. While the latter approach may also be used in selection model factorization, pattern-mixture modeling may make determining the functional form of the selection model and finding reasonable ranges of parameters indexing the selection model simpler. This factorization illuminates how parameters indexing the selection model also index the unverifiable density and *vice versa*. As such, both the parametric form of $f(y \mid x, R = 0)$ and the appropriateness of the chosen selection model are typically considered in tandem.

For example, Scharfstein et al. [2014] found that for any choice of function $r(x, y)$, construction of $f(y \mid x, R = 0)$ via

$$f(y \mid x, R = 0) = f(y \mid x, R = 1) \frac{\exp\{r(x, y)\}}{\mathbb{E}[\exp\{r(x, Y)\} \mid x, R = 1]} \tag{1.9}$$

implies that the $b(x, y)$ from Equation (1.7) defining the selection model takes the form $b(x, y) = l(x) + r(x, y)$, where $l(x)$ depends on $f(r \mid x)$, which is entirely observable, and $r(x, y)$. Note Equation (1.9) bears a striking resemblance to Equation (1.8). This decomposition shows that if $b(x, y)$ is additive in some function that depends on only X and some function that depends on both X and Y, the former does not define the functional relationship between $f(y \mid x, R = 0)$ and $f(y \mid x, R = 1)$, because it is cancelled when writing out Equation (1.8). Scharfstein et al. [2014] call for fixing the functional form and parameters indexing $r(x, y)$, and hence identifiability of all θ follows from the assumptions made by Scharfstein et al. [2014].

For another example, Kaciroti and Raghunathan [2014] suppose $f(y \mid x, R = 0)$ and $f(y \mid x, R = 1)$ are members of the same exponential family with mean models of the same functional form with different coefficients and intercepts. This approach is an application of the recommendation of Daniels and Hogan [2008] for sensitivity analyses. The implied selection model must take a specific form, depending in part on the difference in coefficients, which may be assessed for plausibility. The differences between coefficients and intercepts are not identifiable, and are therefore assigned a prior. Seeing how these same parameters index the selection model may help guide this process. The final estimate of $\mathbb{E}[Y \mid x]$ is a function of $\mathbb{E}[Y \mid x, R = 1]$, $\mathbb{E}[Y \mid x, R = 0]$, and $f(r \mid x)$; the authors recommend using Monte Carlo methods for inference.

The relationship may also be established semi-parametrically. Hogan and Laird [1997] modeled how covariates D related to missingness impact the outcome mean model, i.e. $\mathbb{E}[Y \mid x, D = d]$, in a so-called *general linear mixture model*. This model was later reparameterized by Fitzmaurice et al. [2001] with covariates D centered about their mean so that marginalization of $\mathbb{E}[Y \mid x, d]$ over D to acquire $\mathbb{E}[Y \mid x]$ remains linear in X. Estimation of parameters indexing $\mathbb{E}[Y \mid x, d]$ (and therefore $\mathbb{E}[Y \mid x]$) is possible via maximum likelihood estimation using a Fisher scoring algorithm.

A common choice of selection models defines $f(y \mid x, R = 0)$ through an *exponential tilt* of $f(y \mid x, R = 1)$ [Shardell et al., 2013]. Suppose $b(x, y) = \alpha(x) + \beta y$ for arbitrary function

$\alpha(\cdot)$ and coefficient β, i.e. $b(x, y)$ is additive in X and Y, with linear dependence on Y. Then,

$$f(y \mid x, R = 0) = f(y \mid x, R = 1)\frac{\exp\{\beta y\}}{\mathbb{E}[\exp\{\beta Y\} \mid x, R = 1]}. \tag{1.10}$$

The tilted density is often, but not always, in the same family as $f(y \mid x, R = 1)$. Now, the entirety of parameters indexing $f(y \mid x)$ are those indexing $f(r \mid x)$, $f(y \mid x, R = 1)$, and β. Therefore, all that remains is finding the value of β.

However, β is not identifiable from observed data alone. This parameter may be fixed across a range of values [Scharfstein et al., 1999, Shardell et al., 2013] or assigned a prior [Wang et al., 2010]. Birmingham et al. [2003] extend the definition of exponential tilting by allowing the coefficient to Y in $b(x, y)$ to depend on observed data, i.e. $\beta(x)$, assuming the functional form and parameters indexing $\beta(x)$ are fixed.

Even in the simpler case of β not depending on X, estimation of its value requires some data apart from X, Y, and R. Kim and Yu [2011] assume a sample of original nonrespondents may be reached to acquire the outcome measure, called a *validation sample*, to estimate β. The estimate of β may be plugged into a nonparametric kernel-based regression estimator of $\mathbb{E}[Y \mid x, R = 0]$, an extension of the estimation method of Cheng [1994] to nonignorable data. Tang et al. [2014] and Zhao et al. [2013] further develop nonparametric estimators using exponential tilts, assuming β is known, estimated from a validation sample, or estimated using an independent sample. Shao and Wang [2016] depend on an *instrument*—a covariate collected without missingness during the study that is associated with the outcome that has zero effect on the selection model—to estimate β via profiling and the two-step generalized method of moments. The use of the instrument enables definition of additional estimating equations; these moments motivate a similar kernel-based estimator of $\alpha(x)$ to Kim and Yu [2011], except the estimator is a function of β. Finally, by plugging in the estimate of $\alpha(x)$ to the remaining estimating equations, β may be estimated via the generalized method of moments.

The method of Shao and Wang [2016] is applicable to a selection model with any func-

tional dependence on Y, as long as X and Y are additive in the selection model on the logit scale. The selection model assumed to hold true in the novel methodology defined in Section 1.4.2.3 is a member of the same class. Instead of *instruments*, we use summaries of Y collected outside the current study to identify parameters indexing the functional dependence of the selection model on Y. The functional form of the dependence is dictated by the functions of the outcome Y with accessible mean estimates. This novel methodology is described in detail in Section 1.4.

1.4 Defining the Regression Function Under Additive Nonignorability

This section begins with a summary of theoretical results due to Sadinle and Reiter [2019] on a class of selection models yielding just-identified full data distributions given some auxiliary marginal information in Section 1.4.1 . This class of selection models is additive in baseline covariates and the potentially-missing outcome, hence the moniker of *additive nonignorability*. Nonparametric identifiability of the full data distribution guarantees all parameters θ indexing the regression function will be identifiable using the combination of observed data and auxiliary marginal information. We suppose the true selection model falls in this very broad class.

We then apply these results to identify $f(y \mid x)$ using the pattern-mixture modeling factorization in Equation (1.5). First, we consider empirically verifiable components of the decomposition, defining classes of covariate-dependent response probability models and verifiable densities in Sections 1.4.2.1 and 1.4.2.2 respectively that corroborate with our proposed estimation technique. However, note that identifiability is maintained for *any* observed data density. We will then use a subclass of the selection models defined by Sadinle and Reiter [2019] to define the unverifiable density, using the procedure summarized in Section 1.3.2.3. With this specific subclass, there is no need to define $f(x)$ nor to specify functional dependence of the selection model on baseline covariates.

The final result is a regression model indexed by parameters that are identifiable using observed data and auxiliary marginal information. A proposed technique to estimate θ

will be described in Section 1.5, including further consideration of how to best incorporate estimates of auxiliary marginal information, accounting for their uncertainty. For now, we rely on point estimates.

1.4.1 Nonparametric Identifiability Under Additive Nonignorability Using Auxiliary Marginal Information

Results on nonparametric identifiability by Sadinle and Reiter [2019] apply to any arbitrary observed data density. This density may be factorized into $f(x, y \mid R = 1)$, $f(x \mid R = 0)$, and $\mathbb{P}[R = r]$, but other factorizations are available. Sadinle and Reiter [2019] also assume there is access to some *auxiliary marginal information*, or marginal means of functions of the outcome. More specifically for $K \geq 1$, we know $\{\mathbb{E}[h_k(Y)] : k = 1, ..., K\}$ for some functions $h_k(\cdot)$.

The identifying assumption used to define the full data density is the functional form of the selection model, depending on unknown parameters. Sadinle and Reiter [2019] suppose the selection model falls in a class that is additive in X and Y after transformation by function $q(\cdot)$ that is differentiable and monotonically increasing on $(0, 1)$, i.e.

$$q\left(\mathbb{P}[R = 0 \mid x, y]\right) = \alpha(x) + \beta(y), \tag{1.11}$$

where $\alpha(\cdot) \in L_1\{f(x \mid R = 0)\}$, i.e. $\int |\alpha(x)| f(x \mid R = 0) dx < \infty$, and $\beta(\cdot)$ is in the linear span of $\{h_k(\cdot) : k = 1, ..., K\}$. The function $\beta(y)$ may also be explicitly written as $\sum_{k=1}^{K} \beta_k h_k(y)$. By denoting $h(y) = [h_1(y)...h_K(y)]^T$, this sum may be more easily notated by $\boldsymbol{\beta}^T h(y)$, where we redefine $\boldsymbol{\beta} = [\beta_1...\beta_K]^T$ to be the parameters indexing the linear span of $\{h_k(\cdot) : k = 1, ..., K\}$. Later, we will show we need only concern ourselves with estimation of this subset of parameters indexing the selection model, and not those indexing $\alpha(x)$, hence the change in definition of $\boldsymbol{\beta}$.

Note the selection model in Equation (1.11) has the flexibility to encompass different types of missingness. With $\alpha(x) = \alpha$ and $\beta(y) = 0$, we have missingness completely at random;

with $\beta(y) = 0$, we have missingness at random; finally, if $\beta(y) \neq 0$, then missingness is not at random.

With this class of selection models, the full data distribution is just-identified. In other words, there is a one-to-one mapping between the set of observed data distributions and auxiliary marginal information to the set of full data distributions. For more on the implications of nonparametric identifiability, see Section 1.3.2.1.

Note that at this juncture, we suppose full specification of the functional form of $\alpha(x)$ to just-identify the full data density, and therefore ensure all parameters indexing the full data distribution are identifiable. Later on, we will define a subclass of selection models defined in Equation (1.11) so that, for the purposes of identifying $f(y \mid x)$, there is no need to specify the functional form of $\alpha(x)$. In so doing, we retain identifiability of all parameters θ indexing the regression function using observed data.

As noted above, as in Sadinle and Reiter [2019], we retain the use of random X to define the observed and complete data, so that the complete data distribution models behavior of X in addition to R and Y. Considering X to be random is critical to our proposed estimation strategy in Section 1.5. However, since there is no need to estimate parameters indexing $\alpha(x)$ as a result of the subclass of selection models used, we do not need to specify the density $f(x)$. We will only depend on the empirical distribution of X to estimate $\boldsymbol{\beta}$.

With these results in hand, we now define components of the factorization of the complete data distribution, but more directly define $f(y \mid x)$ using the parameterization in Equation 1.5, with assured identifiability of parameters indexing $f(y \mid x)$. First, we provide guidance on defining the covariate-dependent response probability and verifiable density. Then, we propose a method to use results due to Sadinle and Reiter [2019] to define the unverifiable density.

1.4.2 Defining Components of the Regression Function

1.4.2.1 Covariate-Dependent Response Probability

The covariate-dependent response probability $\mathbb{P}[R = r \mid x]$ in the decomposition in Equation (1.5) depends only on fully observed data X and R. Therefore, assumptions about its parametric form are empirically verifiable using observed data.

For the sake of illustration of the estimation methodology outlined in Section 1.5, we initially suppose $\mathbb{P}[R = 1 \mid x]$ follows a logistic regression, i.e.

$$\text{logit}(\mathbb{P}[R = 1 \mid x]) = \delta^T x_\delta \tag{1.12}$$

for some coefficients δ and some subset X_δ of X. Parameters δ may then be added to parameters θ indexing $\mathbb{E}[Y \mid x]$. This class of response probabilities is flexible and widely understood. However, a sufficient condition to employ the same methodology outlined in Section 1.5 is the ability to estimate the parameters indexing the covariate-dependent response probability using estimating equations with mean zero estimating functions. The final functional form of the covariate-dependent response probability should be verified using the data and a model-checking technique of the investigator's choice.

1.4.2.2 Verifiable Density

The verifiable density $f(y \mid x, R = 1)$ is so-named because of its dependence on observed data only. Like the covariate-dependent response probability, rather than considering the infinitude of densities available, for pedagogical purposes, we focus on $[Y \mid x, R = 1]$ in the class of generalized linear models [McCullagh and Nelder, 1989], i.e.

$$f(y \mid \xi, \eta) = \exp\left[\frac{y\xi - \zeta(\xi)}{\eta} + c(y, \eta)\right],$$

for some dispersion parameter η, parameter ξ dependent on subset X_λ of X (therefore defining the density specifically for each study subject), and functions $\zeta(\cdot)$ and $c(\cdot)$. The observ-

able regression function $\mathbb{E}[Y \mid x, R = 1]$ is now defined to be $\mathbb{E}[Y \mid x, R = 1] \equiv \zeta'(\xi)$, where $\nu(\mathbb{E}[Y \mid x, R = 1]) = \lambda^T x_\lambda$ for link function $\nu(\cdot)$ and some coefficients x_λ. Now, parameters λ and η may be added to vector θ, i.e. $\theta = (\delta, \lambda, \eta, ...)$.

As in the case of the covariate-dependent response probability, verifiable densities to which the methodology outlined in Section 1.5 applies need not fall into the class of GLMs. The ability to estimate parameters indexing $f(y \mid x, R = 1)$ using estimating equations with mean zero estimating functions is sufficient. The final functional form of the verifiable density should be empirically checked using a model-checking technique of the investigator's choice; an example is introduced in Section 1.7 and empirically verified in Appendix A.2.

1.4.2.3 Unverifiable Density

We suppose there is access to the auxiliary marginal information $\{\mathbb{E}[h_k(Y)] : k = 1, ..., K\}$ in the target population (defined in Section 1.4.1) and that the selection model is in the class defined in Equation (1.11). We also assume that $q(\cdot)$ is a logit link, and therefore

$$\begin{aligned}
\text{logit}\left(\mathbb{P}[R = 0 \mid x, y]\right) &= \log \frac{f(x, y \mid R = 0) f(R = 0)}{f(x, y \mid R = 1) f(R = 1)} = \alpha(x) + \beta(y) \\
\implies f(x, y \mid R = 0) &= \frac{f(R = 1)}{f(R = 0)} f(x, y \mid R = 1) \exp[\alpha(x) + \beta(y)] \qquad (1.13) \\
\implies f(y \mid x, R = 0) &= f(y \mid x, R = 1) \frac{f(R = 1)}{f(R = 0)} \frac{f(x \mid R = 1)}{f(x \mid R = 0)} \exp[\alpha(x) + \beta(y)] \\
&= \left[\exp[\alpha(x)] \frac{f(R = 1)}{f(R = 0)} \frac{f(x \mid R = 1)}{f(x \mid R = 0)}\right] \exp[\beta(y)] f(y \mid x, R = 1). \qquad (1.14)
\end{aligned}$$

The auxiliary marginal information fixes the functional form of $\beta(y)$ (see Section 1.4.1). Therefore, the functional form of all components of the right-hand side of Equation (1.14) may be empirically verified or have been fixed except $\alpha(x)$. Ideally, the user would not need to specify that unverifiable form. To that end, we express $\alpha(x)$ via Equation (1.13), noting

that when both sides of Equation (1.13) are integrated with respect to y, we have

$$\begin{aligned} f(x \mid R=0) &= \exp[\alpha(x)]\frac{f(R=1)}{f(R=0)}\int f(x,y \mid R=1)\exp[\beta(y)]dy. \\ \implies \exp[\alpha(x)] &= \frac{(f(R=0)/f(R=1))\, f(x \mid R=0)}{\int f(x,y \mid R=1)\exp[\beta(y)]dy} \\ &= \frac{\frac{f(R=0)}{f(R=1)} f(x \mid R=0)}{\int f(y \mid x,R=1) f(x \mid R=1)\exp[\beta(y)]dy} \\ &= \frac{f(R=0 \mid x)/f(R=1 \mid x)}{\int f(y \mid x,R=1)\exp[\beta(y)]dy} \\ &= \frac{f(R=0 \mid x)/f(R=1 \mid x)}{\mathbb{E}[\exp(\beta(Y)) \mid x,R=1]}. \end{aligned} \tag{1.15}$$

Now, plugging in Equation (1.15) for $\exp[\alpha(x)]$ in Equation (1.14), $f(y \mid x, R=0)$ takes the form

$$f(y \mid x,R=0) = f(y \mid x,R=1)\frac{\exp[\beta(Y)]}{E[\exp[\beta(Y)] \mid x,R=1]}. \tag{1.16}$$

This result is a more specific form of the selection model defined in Equation (1.9) with $r(x,y)$ that depends on Y only. It may also be considered a generalization of the exponential tilt in Equation (1.10). Like exponential tilts, to define the unverifiable density all we require is specification of the verifiable density $f(y \mid x, R=1)$ (defined in Section 1.4.2.2) once the functional form of $\beta(Y)$ has been fixed. Moreover, we do not need to write an explicit formula for $\alpha(x)$. Note that, as a consequence of this definition of the unverifiable density,

$$\mathbb{E}[Y \mid x,R=0] = \frac{\mathbb{E}[Ye^{\beta(Y)} \mid x,R=1]}{\mathbb{E}[e^{\beta(Y)} \mid x,R=1]}, \tag{1.17}$$

completing the definition of the unobservable regression in Equation (1.6), and therefore the functional form of the entire regression function $\mathbb{E}[Y \mid x]$. Hence, $\boldsymbol{\beta} = \{\beta_1, ..., \beta_K\}$ defining the linear span $\beta(y) = \sum_{k=1}^{K} \beta_k h_k(y)$ are a subset of θ.

1.4.3 Fully Parametric Specification of Selection Model

As noted above, defining $f(y \mid x, R = 0)$ using the identifying assumptions in Section 1.4.2.3 obviates the need to specify the form of $\alpha(x)$. However, as a consequence of defining the covariate-dependent response probability $\mathbb{P}[R = r \mid x]$ and verifiable density $f(y \mid x, R = 1)$, the parametric form of $\alpha(x)$ is inherently fixed using Equation (1.15), which is equivalent to

$$\alpha(x) = \log\left[\frac{f(R = 0 \mid x)}{f(R = 1 \mid x)}\right] - \log \mathbb{E}[\exp\{\beta(Y)\} \mid x, R = 1],$$

defined entirely using verifiable densities and the functional form $\beta(\cdot)$. This is a more specific form of results on selection models by Scharfstein et al. [2014] in Equation (1.9) and discussed in Section 1.3.2.3).

For example, with the logistic regression for response in Section 1.4.2.1, $\delta^T x_\delta = \log\left[\frac{f(R=1|x)}{f(R=0|x)}\right]$, and we have

$$\begin{aligned}\alpha(x) &= -\delta^T x_\delta - \log \mathbb{E}[\exp\{\beta(Y)\} \mid x, R = 1] \\ &= -\delta^T x_\delta - \log \mathbb{E}[\exp\{\beta(Y)\} \mid x_\lambda, R = 1]. \end{aligned} \tag{1.18}$$

With the form of $\alpha(x)$ fixed, the parametric form of the entire selection model is also fixed. All parameters indexing $\alpha(x)$ are a subset of θ defined above. With estimates of θ in hand, we may therefore conduct inference on $\mathbb{P}[R = 0 \mid x, y]$ for any given value (x, y), which may be useful to determine how reasonable the implied selection model is. If estimates of parameters indexing the selection model yield an unreasonable missingness mechanism, then perhaps use of these identifying assumptions should be reconsidered.

1.5 Estimation Procedure

Through the process outlined in Section 1.4, the parametric form of the regression function $\mathbb{E}[Y \mid x]$ has been defined, indexed by parameters θ. We now look to define consistent and asymptotically normal (CAN) estimates of θ via the generalized method of moments

(GMM), to which the delta method [Van der Vaart, 1998, pg 25-26] may be applied to conduct inference on the regression function evaluated at a given $X = x$.

This section begins with a general description of GMM and how it relates to estimating equations, followed by a discussion of the specific use of GMM to estimate θ indexing $\mathbb{E}[Y \mid x]$ here. The section concludes with notes on consistent estimation of the asymptotic covariance matrix for estimates of θ and use of the delta method to conduct inference on values of the regression function.

1.5.1 Generalized Method of Moments

The generalized method of moments (GMM) is commonly used in econometrics to acquire CAN estimates $\hat{\theta}$ of parameters θ and to define the asymptotic covariance matrix of those estimators [Newey and McFadden, 1994, Jesus and Chandler, 2011].

If the length of θ is p, the method depends on a moment vector $m(Z, \theta)$ at least p-long depending on random variables Z (e.g. in our case, $Z = (X, Y, R)$) and parameters θ we wish to estimate. For simplicity, we suppose $m(Z, \theta)$ is exactly p-long; for results for different lengths of $m(Z, \theta)$, see a description of the methodology by Jesus and Chandler [2011]. If the true value of θ is θ_0, moment vector $m(Z, \theta)$ must be defined so that $\mathbb{E}[m(Z, \theta_0)] = 0$.

A simple way to use $m(Z, \theta)$ to estimate θ is via the *method of moments*, or setting the sample moment vector $m_{(n)} \equiv \sum_{i=1}^{n} m(z_i, \theta)$ equal to 0 and solving for θ [Van der Vaart, 1998, pg 35-37]. However, a solution may not always exist due to sampling variability in $m_{(n)}$ [Jesus and Chandler, 2011]. For this reason, we turn to the generalized method of moments, where $\hat{\theta}$ is acquired by minimizing the weighted square of the sample moment vector, i.e.

$$\begin{aligned} \hat{\theta} &= \text{argmin}_\theta \left(\left(\gamma_n m_{(n)} \right)^T W_n \left(\gamma_n m_{(n)} \right) \right) \\ &\equiv \text{argmin}_\theta Q(m_{(n)}, \theta), \end{aligned} \tag{1.19}$$

for some $p \times p$ weighting matrices W_n and γ_n [Jesus and Chandler, 2011], which may both

depend on data, but not on θ. Matrix W_n must be positive definite, converge in probability to matrix W, and is often a function evaluated at some preliminary estimate of θ [Newey and McFadden, 1994]; optimal choice for W_n is discussed in Section 1.5.2. Matrix γ_n is atypical in the GMM literature, but ensures equal rates of convergence in distribution of all elements of $\hat{\theta}$; choices of γ_n will be discussed in Section 1.5.2. For notational simplicity, the vector $\tilde{m}_{(n)} = \gamma_n m_{(n)}$ will be used throughout the remainder of this section.

Note that solutions to GMM exist more often across applications than the method of moments [Jesus and Chandler, 2011]. Even if a solution does not exist, approximate solutions have been shown to have the same asymptotic properties as exact solutions [Van der Vaart, 1998, pg 42], ameliorating issues with the method of moments.

Taking special care with choices of W_n and γ_n, solutions to the method of moments may also be the solution to the GMM, e.g. in the case of generalized linear models (GLMs). Hence, estimates of parameters indexing GLMs via the standard method of moments approach, as well as their asymptotic distribution, will be the same as those acquired using GMMs, as long as the appropriate moment vector is chosen. This useful equivalence that will be exploited in Section 1.5.2.

There are a series of results published on required conditions for CAN estimates of parameters using GMM [Van der Vaart, 1998, Newey and McFadden, 1994]; we will use results from Jesus and Chandler [2011] as a guideline, with reference to other useful results as they apply. Jesus and Chandler [2011] start by showing the GMM is a specific form of estimating equations by noting that $\hat{\theta}$ solving Equation (1.19) is also the solution to $\frac{\partial}{\partial\theta}Q(m_{(n)}, \theta) = 0$, which forms a set of estimating functions. Then, they translate general requirements for CAN estimates of θ in estimating equations to requirements for GMMs, with specifications in terms of the moment vector $m(Z, \theta)$ when possible (instead of the estimating functions due to their complexity). Sufficient conditions to establish consistency, convergence in distribution, and asymptotic normality are listed in that order.

Jesus and Chandler [2011] assume the moment vector takes a specific form—additive separability of random variables and functions depending on θ—for which sufficient conditions

for consistency of $\hat{\theta}$ are met in all but pathological situations. Since our definition of the moment vector is more general, we require more careful consideration of consistency; for this reason, we turn our focus to the system of estimating equations inherently used with GMM and apply existing results on consistency of estimates from estimating equations. We more formally define this system of estimating equations to be

$$\begin{aligned} g_n(\theta; z_n) &= \frac{\partial}{\partial\theta} Q(m_{(n)}, \theta) \\ &= \left(\frac{\partial}{\partial\theta}\tilde{m}_{(n)}\right)^T W_n \tilde{m}_{(n)} \\ &\equiv 0. \end{aligned}$$

Assuming there is some solution $\hat{\theta}$ to the estimating equations, consistency and uniqueness of $\hat{\theta}$ are assured if two conditions are met:

1. There exists a sequence (d_n) of $p \times p$ matrices that do not depend on θ such that $d_n g_n(\theta; z_n)$ converges uniformly in probability to a non-random function $g_l(\theta)$.

2. To prove uniqueness of the solution, $g_l(\theta)$ must have a unique root at $\theta = \theta_0$ to the equation $g_l(\theta) = 0$ and $g_l(\cdot)$ is bounded away from zero except in the neighborhood of θ_0.

Uniform convergence in probability is achieved if the class of functions is *Glivenko-Cantelli* [Van der Vaart, 1998, 265-291]; however, this condition is often more restrictive than necessary to prove consistency of solutions to estimating equations and may be difficult to prove in some cases [Van der Vaart, 1998, pg 46-47]. If the moments are continuous and parameter space Θ assumed to be compact, then conditions for *Glivenko-Cantelli* are quite simple to prove [Van der Vaart, 1998, 272]; there appears to be precedent for making the compactness assumption without known restrictions on the parameter space [Newey and McFadden, 1994, 2131-2133]. Without the assumption of compactness, Theorem 2.7 of Newey and McFadden

[1994, pg 2133] and discussion in the remainder of the section written by Newey and McFadden [1994] in which that theorem appears might prove to be useful. We note that proof of consistency will depend on assumptions about finiteness of moments of X, for which we have not defined a distribution.

In place of the potentially complex dependence on $g_l(\theta)$ above to prove uniqueness of the solution to Equation (1.19), one might potentially use Lemma 2.3 by Newey and McFadden [1994, pg 2126]. However, note that for consistency and asymptotic normality, uniqueness need not hold. If there are multiple solutions (i.e. local minima) in practice, values of the objective function $Q(m_{(n)}, \theta)$ may be compared at each solution to find the global minimum [Jesus and Chandler, 2011]. Newey and McFadden [1994, pg 2126-2128] list sufficient conditions for local identification—unique solutions within a neighborhood of θ_0. Also note this same passage acknowledges that the uniqueness of global solutions is often simply assumed.

In addition to conditions for consistency, sufficient conditions to prove convergence in distribution with asymptotic variance of the form specified below include the following restrictions on moment vector $m(Z, \theta)$:

1. $m(Z, \theta)$ is twice continuously differentiable with respect to θ.

2. Elements of $m(Z, \theta)$ are linearly independent.

Further conditions depend on additional matrices $\Omega_n^{(r)} = \frac{\partial^2 \tilde{m}^r_{(n)}}{\partial\theta\partial\theta'}$, where $\tilde{m}^r_{(n)}$ is the r^{th} element of $\tilde{m}_{(n)}$. Suppose the sequence of matrices (γ_n) are chosen such that $\tilde{m}_{(n)}$ evaluated at $\theta = \theta_0$ has a limiting covariance matrix that tends to a limiting positive definite matrix S. Also suppose there exists a sequence (τ_n) of invertible $p \times p$ matrices that do not depend on θ and are such that

1. $\lim_{n\to\infty} \tau_n = 0$

2. For all $r \in \{1, ..., p\}$ the matrices $\left(\frac{\partial}{\partial\theta}\tilde{m}_{(n)}\right)\tau_n$ and $\Omega_n^{(r)}(\theta)\tau_n$ converge uniformly in probability to matrices $T(\theta)$—which must be invertible—and $\Omega^{(r)}(\theta)$, respectively, all elements of which are continuous functions of θ.

If all these requirements are fulfilled, then the covariance matrix of $\tau_n^{-1}\hat{\theta}_n$ converges to

$$\Sigma = [M(\theta_0)]^{-1}\tilde{A}[M'(\theta_0)]^{-1}, \tag{1.20}$$

where

$$\tilde{A} = T'(\theta_0)WSW'T(\theta_0) \text{ and}$$
$$M(\theta_0) = T'(\theta_0)WT(\theta_0).$$

Uniform convergence in probability of the functions composing $\left(\frac{\partial}{\partial\theta}\tilde{m}_{(n)}\right)\tau_n$ and $\Omega_n^{(r)}(\theta)\tau_n$ is simple to prove if the parameter space for θ is compact and $m(\cdot)$ is composed of functions absolutely continuous with respect to θ [Van der Vaart, 1998, 272].

If choosing $\gamma_n = \tau_n = \frac{1}{\sqrt{n}}I_{p\times p}$ and $m(\cdot)$ is continuously differentiable with respect to θ in a neighborhood of θ_0, Theorem 3.4 from Newey and McFadden [1994, 2148] might be used instead to prove convergence in distribution. In particular, the parameter space Θ need only be an open set (which may be $\mathbb{R}^p$) so that θ cannot take values on the boundary of Θ. Therefore, we do not require Θ to be a compact space.

To prove the asymptotic distribution is normal in the more general case, in addition to conditions for convergence in distribution, proof that $\mathbb{E}\left[\frac{\partial^2 Q_n}{\partial^2\theta}\mid_{\theta=\theta_0}\right]$ exists is sufficient to prove asymptotic normality [Jesus and Chandler, 2011]. The form of $\frac{\partial^2 Q_n}{\partial^2\theta}$ is

$$\frac{\partial^2 Q_n}{\partial^2\theta} = \Lambda_n + \left(\frac{\partial}{\partial\theta}\tilde{m}_{(n)}\right)^T W_n \left(\frac{\partial}{\partial\theta}\tilde{m}_{(n)}\right), \tag{1.21}$$

where Λ_n is a $p \times p$ matrix with $(i,j)^{\text{th}}$ element

$$\Lambda_{(i,j)} = \sum_{t=1}^{p}\sum_{s=1}^{p} W_{(t,s)}\Omega_{(i,j)}^t\tilde{m}_s, \tag{1.22}$$

where $W_{(t,s)}$ is the $(t,s)^{\text{th}}$ element of W, $\Omega_{(i,j)}^t$ is the $(i,j)^{\text{th}}$ element of $\Omega_{(n)}^t$, and $\tilde{m}_s$ is the

s^{th} element of $m_{(n)}$. Elements of $\frac{\partial^2 Q_n}{\partial^2 \theta}$ will often depend on random variable X, for which (as previously mentioned) we have not defined a distribution.

If choosing $\gamma_n = \tau_n = \frac{1}{\sqrt{n}} I_{p\times p}$ and $m(\cdot)$ is continuously differentiable with respect to θ in a neighborhood of θ_0, in place of the theory defined by Jesus and Chandler [2011], Theorem 3.4 by Newey and McFadden [1994, pg 2148] specifically proves convergence to a normal distribution under certain conditions. Moreover, there is no need to define the somewhat complicated $\mathbb{E}\left[\frac{\partial^2 Q_n}{\partial^2 \theta}|_{\theta=\theta_0}\right]$. No matter the theorem used to prove asymptotic normality, without specification of the distribution of X, assumptions will be required about finiteness of moments of X.

1.5.2 Applying Generalized Method of Moments to Additive Nonignorability

GMM may be used to estimate $\theta = (\lambda, \eta, \boldsymbol{\beta}, \delta)$, which indexes the regression function $\mathbb{E}[Y \mid x]$, as outlined below. As a quick recap, δ indexes the covariate-dependent response probability, λ and η are the coefficients in a linear function and dispersion parameter, respectively, for the verifiable density, and $\boldsymbol{\beta} = \{\beta_1, ..., \beta_K\}$ indexes the functional dependence of the selection model on the potentially-missing outcome. By construction, the unverifiable density may depend on λ, η, and/or $\boldsymbol{\beta}$.

We begin by defining the moment vector $m(X, Y, R, \theta)$, which we restrict to have the same length as θ. The moment vector is constructed to yield estimates of some components of θ identical to popular existing methods, enabling use of existing software to acquire some estimates. This selection will also enable step-wise estimation of parameters, limiting simultaneous estimation and easing computation.

The moment vector is considered in four separate components; each component is defined by the subscript, e.g. the first component is $m_1(X, Y, R, \theta)$. The ordering of these components is critical. The parameters governing individual components may be estimated separately, at times depending on estimates acquired in previous steps. A detailed description of the estimation methodology is left to later in this section. For clarity, we initially suppose assumptions about distributions of observed data from Sections 1.4.2.1-1.4.2.2 hold, but

consider extensions to other assumptions throughout.

Recalling $\nu(\mathbb{E}[Y \mid x, R = 1]) = \lambda^T x_\lambda$ for link function $\nu(\cdot)$ with dispersion parameter η for the distribution of $[Y \mid x, R = 1]$, the first component of the moment vector is of the same length as λ and is defined as

$$\begin{aligned} m_1(X, Y, R, \theta) &= R\frac{\partial}{\partial \lambda} \log f(Y \mid X, R = 1; \lambda, \eta) \\ &= R\left(\frac{\partial \mathbb{E}[Y \mid X, R = 1; \lambda]}{\partial \lambda}\right)^T \frac{Y - \mathbb{E}[Y \mid X, R = 1; \lambda]}{\text{Var}(Y \mid X, R = 1; \lambda, \eta)}, \end{aligned}$$

which is the same score function used in a typical GLM procedure [Wakefield, 2013, pg 260], where the likelihood for λ and η is defined using the verifiable density. Use of the canonical link for $\nu(\cdot)$ [Wakefield, 2013, 256-257] simplifies this formulation slightly, but is not necessary.

The second component (of length one) is the score for η using the same likelihood function used for the first component, which may generally be denoted by

$$m_2(X, Y, R, \theta) = R\frac{\partial}{\partial \eta} \log f(Y \mid X, R = 1; \lambda, \eta).$$

As previously mentioned, estimation of η is often unnecessary. For some choices of exponential family for $[Y \mid x, R = 1]$, the value of η is known. Otherwise, the necessity of estimating η depends entirely on if/how the third component of the moment vector and/or the asymptotic covariance matrix Σ depends on η.

We note that if the verifiable density is selected outside exponential families that may be used for a GLM, and if there is a method available to estimate parameters indexing such a density using mean zero estimating functions, then the first two components of the moment vector described above may be replaced with these estimating functions without substantial changes to the estimation methodology.

The third component of the moment vector is defined to depend on $\boldsymbol{\beta} = \{\beta_1, ..., \beta_K\}$. Note that, as a consequence of the construction of the unverifiable density, for any $k \in \{1, ..., K\}$,

$$\mathbb{E}[h_k(Y) \mid x, R=0] = \frac{\mathbb{E}[h_k(Y)e^{\beta(Y)} \mid x, R=1]}{\mathbb{E}[e^{\beta(Y)}|x, R=1]}, \tag{1.23}$$

with both expectations on the right-hand-side taken with respect to the verifiable density. Therefore, $\mathbb{E}[h(Y) \mid x, R=0]$ (recalling notation defined in Section 1.4.1) may depend on all elements of $\boldsymbol{\beta}$. Dependence of $\mathbb{E}[h(Y) \mid x, R=0]$ on all elements of $\boldsymbol{\beta}$ should be verified before continuing. Note that by the law of iterated expectation,

$$\mathbb{E}_Y[h_k(Y) \mid R=0] = \mathbb{E}_X[\mathbb{E}_Y[h_k(Y) \mid X, R=0]]. \tag{1.24}$$

We will use this identity to define the third component of the moment vector. Finally, note that using Equation (1.23), we may evaluate $\mathbb{E}[h_k(Y) \mid x, R=0]$ for any individual (whether or not they were lost to follow-up) after defining the verifiable density.

To define the third component of the moment vector, we use a modification of the typical form of the method of moments by setting the empirical average of $\mathbb{E}[h_k(Y) \mid x, R=0]$ among individuals with missing outcome data (i.e. the outer empirical average in Equation (1.24)) for each $k \in \{1, ..., K\}$ equal to the fixed marginal mean $\mathbb{E}[h_k(Y)|R=0] \equiv \mu_{0k}$. Methods of acquiring these latter values using auxiliary marginal information and observed data are described in Section 1.5.2.1. Each of the K elements of the third component of the moment vector may then be defined as

$$m_{3k}(X, Y, R, \theta) = (1-R)\left(\mathbb{E}[h_k(Y) \mid X, R=0] - \mu_{0k}\right) \text{ for } k = 1, ..., K.$$

The fourth and final component of the moment vector is defined to depend on δ, which governs the covariate-dependent response probability $\text{logit}(\mathbb{P}[R=1 \mid x]) = \delta^T X_\delta$. The score of the likelihood of δ defined using the covariate-dependent response probability may be used across all individuals. This score is identical to the score used for a GLM to estimate the values of coefficients defining the mean model of a Bernoulli-distributed random variable

with logit link:

$$\begin{aligned} m_4(X, Y, R, \theta) &= \frac{\partial}{\partial \delta} \log f(R \mid X; \delta) \\ &= X_\delta^T \left(R - \frac{\exp(\delta^T X_\delta)}{1 + \exp(\delta^T X_\delta)} \right). \end{aligned}$$

If assumptions about the covariate-dependent response probability $\mathbb{P}[R = 1 \mid x]$ are not the same as laid out in Section 1.4.2.1, and another parametric form is chosen for the response probability indexed by parameters estimable using mean zero estimating functions, these estimating functions may replace the fourth component of the moment vector with little change to the estimation methodology.

In addition to the moment vector $m(X, Y, R, \theta)$, use of GMM requires defining matrices W_n, τ_n, and γ_n. To achieve optimal efficiency, published results prove the optimal choice for weight matrix W_n is a consistent estimate of covariance matrix S [Jesus and Chandler, 2011]. This estimate can be an iterative process, using previous estimates of θ to update W_n to then acquire another estimate of θ. A default initial choice for W_n is the identity matrix. If the solution to $Q(m_{(n)}, \theta) = 0$ may be acquired via the method of moments, i.e. solving $m_{(n)} = 0$ (as in the example in Section 1.6), choice of W_n does not impact the value of the estimate nor the efficiency. For simple estimation of Σ, we arbitrarily select W_n to be the identity matrix.

Choosing $\tau_n = \frac{1}{\sqrt{n}} I_{p \times p}$ will yield an asymptotic sampling distribution of the form $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \Sigma)$ for asymptotic covariance Σ. Both this distribution and the distribution derived from the delta method [Van der Vaart, 1998, pg 26] for estimates of the regression function will take forms useful for inference. To enable use of the considerable literature regarding uniform convergence in probability of sample means and/or Theorem 3.4 by Newey and McFadden [1994, 2148], we also recommend choosing $\gamma_n = \frac{1}{\sqrt{n}} I_{p \times p}$.

With these choices of weighting matrices, we consider solutions to the GMM, which may be found in a step-wise fashion with $m(X, Y, Z, \theta)$ as defined above. First, consider the first component of the moment vector $m_1(X, Y, R, \theta)$ only. If $[Y \mid x, R = 1]$ is in the class of

GLMs, the solution to $\sum_{i=1}^{n} m_1(x_i, y_i, r_i, \lambda) = 0$ (which will also be the solution acquired from minimizing the quantity $Q(m_{(n)}, \theta)$ in Equation (1.19)) may be acquired using the standard Fisher scoring for coefficients indexing mean models for GLMs. The values of $\hat{\lambda}$ may then be plugged into $\sum_{i=1}^{n} m_2(x_i, y_i, r_i, \hat{\lambda}, \eta) = 0$ to find the maximum likelihood estimate for η; like $\hat{\lambda}$, these values of $\hat{\eta}$, by default, will be the solution to $\min_\theta Q(m_{(n)}, \theta)$. Estimates of δ defining the fourth component of $m(X, Y, R, \theta)$ may similarly be calculated using the standard Fisher scoring for coefficients indexing mean models for GLMs.

Acquiring estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ might present a bigger challenge. After plugging in $\hat{\lambda}$ and $\hat{\eta}$ to the third component of the moment vector $m_3(X, Y, R, \theta)$, with choices of weighting matrices above, estimates $\hat{\boldsymbol{\beta}}$ are the solutions to Equation (1.19), and hence the minimizing values of $\left(\sum_{i=1}^{n} m_3(x_i, y_i, r_i, \hat{\lambda}, \hat{\eta}, \boldsymbol{\beta})\right)^T \left(\sum_{i=1}^{n} m_3(x_i, y_i, r_i, \hat{\lambda}, \hat{\eta}, \boldsymbol{\beta})\right)$. If a method of moments estimator exists, i.e. $\sum_{i=1}^{n} m_3(x_i, y_i, r_i, \hat{\lambda}, \hat{\eta}, \boldsymbol{\beta}) = 0$ for some $\hat{\boldsymbol{\beta}}$, the estimator is simple to calculate. Otherwise, finding $\hat{\beta}$ might require an iterative search.

With assumptions made in Sections 1.4.2.1-1.4.2.2 about observed data distributions, consistency of estimates of parameters indexing the verifiable density and covariate-dependent response probability is readily apparent [Wakefield, 2013, pg 261]. Consistency of $\hat{\boldsymbol{\beta}}$ may be proven separately or together with the other elements of $\hat{\theta}$; see Section 1.5.1 for tips on proving sufficient conditions for both consistency and asymptotic normality of $\hat{\theta}$.

1.5.2.1 Specifying Marginal Means

As previously mentioned, we must fix the unobserved marginal means $\mathbb{E}[h_k(Y) \mid R = 0] = \mu_{0k}$, $k \in \{1, ..., K\}$ to define the third component of the moment vector. However, we suppose there is access to $\mathbb{E}[h_k(Y)] = \mu_{0k}$, $k \in \{1, ..., K\}$. The former is a function of the latter using the following identity:

$$\mathbb{E}[h_k(Y) \mid R = 0] = \frac{\mathbb{E}[h_k(Y)] - \mathbb{P}[R = 1]\mathbb{E}[h_k(Y) \mid R = 1]}{\mathbb{P}[R = 0]}. \tag{1.25}$$

Throughout this document, we suppose $\mathbb{E}[h_k(Y)] \equiv \mu_k$, $k \in \{1, ..., K\}$ are fixed. They may be estimated—using external surveys, administrative data sets, or expert opinion—or take a variety of values in a sensitivity analysis. By plugging in this fixed value to Equation (1.25) as well as estimates of $\mathbb{P}[R = r]$ and $\mathbb{E}[h_k(Y)|R = 1]$ from the given data set, we may estimate $\mathbb{E}[h_k(Y) \mid R = 0]$.

Default estimates of $\pi_1 \equiv \mathbb{P}[R = 1]$ and $\mu_{1k} \equiv \mathbb{E}[h_k(Y)|R = 1]$ are sample averages, i.e.

$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} r_i, \tag{1.26}$$

$$\hat{\mu}_{1k} = \frac{\sum_{i=1}^{n} r_i h_k(y_i)}{\sum_{i=1}^{n} r_i}. \tag{1.27}$$

However, other estimates may be used.

Use of the identity in Equation (1.25) supposes $\mathbb{P}[R = 1]$ and $\mathbb{E}[h_k(Y)|R = 1]$ are fixed quantities. However, these estimates have variability of their own. Estimates $\hat{\boldsymbol{\beta}}$ are clearly a function of these estimates, in addition to $\hat{\lambda}$ and $\hat{\eta}$. Therefore, to calculate the true variance of $\hat{\boldsymbol{\beta}}$, we must account for variability of estimates of $\mathbb{P}[R = r]$ and $\mathbb{E}[h_k(Y)|R = 1]$, in addition to considering the variance of $\hat{\lambda}$ and $\hat{\eta}$.

A natural approach is to estimate $\mathbb{P}[R = 1]$ and $\mathbb{E}[h_k(Y)|R = 1]$ via the same GMM used to estimate θ by adding an additional $K + 1$ elements to the moment vector. For example, components taking the form

$$\begin{aligned} m_5(X, Y, R, \theta, \pi_1, \mu_{1k}) &= R - \pi_1, \\ m_{6k}(X, Y, R, \theta, \pi_1, \mu_{1k}) &= R\left(h_k(Y) - \mu_{1k}\right) \text{ for } k \in \{1, ..., K\} \end{aligned}$$

will yield the same default empirical averages defined in Equations (1.26) and (1.27). The value μ_{0k} in the third component of the moment vector if then replaced with the right-hand side of Equation 1.25 (i.e. a function of μ_k, π_1, and μ_{1k}). Note all estimates of λ, η, β, and δ will remain unchanged.

Recall we have previously fixed parametric forms of $f(y \mid x, R=1)$ and $f(r \mid x)$, both of which are empirically verifiable. Therefore, using estimates $\hat{\lambda}$ and $\hat{\delta}$, μ_{1k} and π_1 may generally be consistently estimated by taking the empirical average of $\mathbb{E}[h_k(Y) \mid x, R=1; \hat{\lambda}]$ among observed individuals and $\mathbb{P}[R=1 \mid x; \hat{\delta}]$ among all individuals, i.e.

$$\hat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}[R_i = 1 \mid x_i; \hat{\delta}], \tag{1.28}$$

$$\hat{\mu}_{1k} = \frac{\sum_{i=1}^{n} r_i \mathbb{E}[h_k(Y_i) \mid x_i, R_i = 1; \hat{\lambda}]}{\sum_{i=1}^{n} r_i}. \tag{1.29}$$

Consistency of these estimates may generally be proven using Lemma 4.3 due to Newey and McFadden [1994, pg 2156].

In these estimates, data R and $h_k(Y)$ in Equations (1.26)-(1.27) are replaced with the parametric forms of their expected values conditional on X. The estimates in Equations (1.28)-(1.29) are unbiased because, by the law of iterated expectation,

$$\begin{aligned}
\mathbb{E}_X[\mathbb{E}_R[R \mid X]] &= \mathbb{E}_X[\mathbb{P}[R=1 \mid X]] \\
&= \mathbb{E}_R[R] \\
&= \mathbb{P}[R=1], \text{ and} \\
\mathbb{E}_X[\mathbb{E}_Y[h_k(Y) \mid X, R=1]] &= \mathbb{E}_Y[h_k(Y) \mid R=1].
\end{aligned}$$

Similar logic was used to define the third component of the moment vector. With correctly-specified models, these estimates will be asymptotically equivalent to Equations (1.26) and (1.27). Equations (1.28) and (1.29) may be acquired by replacing $m_5(\cdot)$ and $m_6(\cdot)$ above with

$$\begin{aligned}
m_5(X, Y, R, \theta, \pi_1, \mu_{1k}) &= \mathbb{P}[R=1 \mid X; \delta] - \pi_1, \\
m_{6k}(X, Y, R, \theta, \pi_1, \mu_{1k}) &= R\left(\mathbb{E}[h_k(Y) \mid X, R=1; \lambda] - \mu_{1k}\right) \text{ for k } \in \{1, ..., K\}.
\end{aligned}$$

We have not found a universal result on relative efficiency of estimates of π_1 and μ_{1k} using these approaches. In particular, efficiency of $\hat{\mu}_{1k}$ depends on efficiency of estimates of λ. With correctly-specified covariate-dependent response probability models and verifiable densities, we suspect the second set of proposed additions to the moment vector will yield more efficient estimation of π_1 and μ_{1k} due to their dependence on the functional forms of $\mathbb{E}[h_k(Y) \mid x, R = 1; \lambda]$ and $\mathbb{P}[R = 1 \mid x; \delta]$. In general, we recommend calculating the covariance matrix using both approaches and comparing efficiency. Since estimates of the asymptotic covariance matrix are subject to error themselves, we recommend comparing the actual asymptotic covariance when possible. Note that the typical advantage of using $m_5(\cdot)$ and $m_6(\cdot)$ yielding estimates in Equations (1.26)-(1.27) is that they are robust to misspecification of $\mathbb{E}[h_k(Y) \mid x, R = 1; \lambda]$ and $\mathbb{P}[R = 1 \mid x; \delta]$. This quality is not important here, since our methodology depends quite heavily on correctly specified models. We suppose the model for covariate-dependent response probability is correct and that the entire verifiable density is correct to ensure the unobservable regression model used in $m_3(\cdot)$ is correctly specified.

We may also wish to account for variability in estimates of $\mathbb{E}[h_k(Y)]$, $k \in \{1, ..., K\}$. Fixing this value may be appropriate in some circumstances, e.g. if the standard error of the estimate is so small, it is effectively known, or if the intent is to use a number of values for sensitivity analyses. In the latter case, we may also directly fix a number of values of μ_{0k}; an example of this approach is illustrated in Section 1.7.

Otherwise, variability in the estimation of $\mathbb{E}[h_k(Y)]$ may be accounted for by adding another component to the moment vector, i.e. adding $\mu_k = \mathbb{E}[h_k(Y)]$, $k \in \{1, ..., K\}$ to the list of parameters being estimated using the same GMM procedure. Estimates of $\mathbb{E}[h_k(Y)]$ from auxiliary data are often acquired via method of moments—often more specifically Horvitz-Thompson weighted estimators [Little and Rubin, 2002, pg 44-47]—and may therefore be easily incorporated into the existing GMM. The addition to the moment vector to estimate μ_{1k} in Equation (1.27) is an example. We often expect access to the approximate sampling distribution of μ_k from auxiliary information when individual-level data are not available

(e.g. published results from surveys without publicly accessible data); we leave accounting for variability due to estimation of μ_k in these cases to future research.

1.5.2.2 Estimating the Asymptotic Covariance

We separately estimate each component of the covariance matrix Σ in Equation (1.20). With fixed W, we must estimate $T(\theta_0)$ and S. We achieve a consistent estimate $\hat{\Sigma}$ of Σ by evaluating Equation (1.20) using consistent estimates $\hat{T}(\theta_0)$ of $T(\theta_0)$ and $\hat{S}$ of S. We start by proposing estimators for each component, followed by a discussion of conditions under which these estimates are consistent. Throughout, we suppose $\gamma_n = \tau_n = \frac{1}{\sqrt{n}} I_{p\times p}$ and $W_n = W$ is a known fixed matrix.

Since $\mathbb{E}[m(Z,\theta_0)] = 0$, the covariance matrix of $m(Z,\theta_0)$ is $S = \mathbb{E}[m(Z,\theta_0)m(Z,\theta_0)^T]$, with the expectation taken with respect to the joint distribution of X, Y, and R. We may estimate each component of S using an empirical average evaluated at estimates $\hat{\theta}_n$ of θ, so that the $(j,k)^{\text{th}}$ component of $\hat{S}$ is

$$\hat{S}_{jk} = \frac{1}{n}\sum_{i=1}^{n} m_j(x_i, y_i, r_i, \hat{\theta}_n) m_k(x_i, y_i, r_i, \hat{\theta}_n). \tag{1.30}$$

With the values of γ_n and τ_n we have fixed, $T(\theta_0) = \mathbb{E}_{(X,Y,R)}\left[\frac{\partial}{\partial\theta} m(Z,\theta)|_{\theta=\theta_0}\right]$. Matrix $T(\theta_0)$ is $p \times p$, but it is not symmetric; however, dependence of Σ on $T(\theta_0)$ means the orientation we use to define the matrix does not matter. We arbitrarily suppose row j corresponds to the j^{th} element of θ, and column k the k^{th} element of $m(X,Y,R,\theta)$. We continue to use empirical averages evaluated at $\hat{\theta}_n$ to estimate $T(\theta_0)$, so that the $(j,k)^{\text{th}}$ element of $\hat{T}(\theta_0)$ is

$$\hat{T}(\theta_0)_{jk} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta_j} m_k(x_i, y_i, r_i, \theta)|_{\theta=\hat{\theta}_n}.$$

It is simplest to divide $\hat{T}(\theta_0)$ into individual matrices according to the four groups of the parameter vector θ (λ, η, β, and δ) and the four components of the moment vector defined in Section 1.5.2. Many components of $m(X,Y,R,\theta)$ depend only on a subset of θ, and therefore

some elements of $\hat{T}(\theta_0)$ will be 0. Denote $\hat{T}(\theta_0)$ by

$$\hat{T}(\theta_0) = \begin{bmatrix} \hat{T}_{11} & \hat{T}_{12} & 0 & 0 \\ \hat{T}_{21} & \hat{T}_{22} & 0 & 0 \\ \hat{T}_{31} & \hat{T}_{32} & \hat{T}_{33} & 0 \\ 0 & 0 & 0 & \hat{T}_{44}. \end{bmatrix}$$

With the assumptions made about the verifiable density in Section 1.4.2.2 and covariate-dependent response probability in Section 1.4.2.1, the nonzero matrices composing $\hat{T}(\theta_0)$ may be more explicitly written as

$$\hat{T}_{11} = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\partial}{\partial\lambda}\left[\left(\frac{\partial \mathbb{E}[Y \mid x_i, R_i = 1; \lambda]}{\partial\lambda}\right)^T \frac{y_i - \mathbb{E}[Y \mid x_i, R_i = 1; \lambda]}{\mathrm{Var}(Y \mid x_i, R_i = 1; \lambda, \eta)}\right]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{21} = \frac{1}{n}\sum_{i=1}^{n} r_i \left(\frac{\partial \mathbb{E}[Y \mid x_i, R_i = 1; \lambda]}{\partial\lambda}\right)^T (y_i - \mathbb{E}[Y \mid x_i, R_i = 1; \lambda]) \frac{\partial}{\partial\eta}\left[\mathrm{Var}^{-1}(Y \mid x_i, R_i = 1; \lambda, \eta)\right]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{12} = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\partial^2}{\partial\eta\partial\lambda}[\log f(y_i \mid x_i, R_i = 1; \lambda, \eta)]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{22} = \frac{1}{n}\sum_{i=1}^{n} r_i \frac{\partial^2}{\partial\eta^2}[\log f(y_i \mid x_i, R_i = 1; \lambda, \eta)]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{13} = \frac{1}{n}\sum_{i=1}^{n} (1 - r_i) \frac{\partial}{\partial\lambda}\mathbb{E}[h(Y) \mid x_i, R_i = 0; \lambda, \eta, \boldsymbol{\beta}]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{23} = \frac{1}{n}\sum_{i=1}^{n} (1 - r_i) \frac{\partial}{\partial\eta}\mathbb{E}[h(Y) \mid x_i, R_i = 0; \lambda, \eta, \boldsymbol{\beta}]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{33} = \frac{1}{n}\sum_{i=1}^{n} (1 - r_i) \frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}[h(Y) \mid x_i, R_i = 0; \lambda, \eta, \boldsymbol{\beta}]\Bigg|_{\theta=\hat{\theta}_n},$$

$$\hat{T}_{44} = \frac{1}{n}\sum_{i=1}^{n} -x_i x_i^T \frac{\exp(\hat{\delta}^T x_{\delta i})}{(1 + \exp(\hat{\delta}^T x_{\delta i}))^2}.$$

If η is not estimated, and hence $m_2(\cdot)$ is not a component of the moment vector used to estimate θ, related elements may be removed from $\hat{T}(\theta_0)$ and $\hat{S}$. To assess when $\hat{T}(\theta_0)$ and

$\hat{S}$ consistently estimate $T(\theta_0)$ and S respectively, note that each element of $T(\theta_0)$ and S is an expected value. Suppose $\mathbb{E}[\rho(X, Y, R, \theta_0)]$ is any element of either matrix with expectation taken with respect to the joint distribution of (X, Y, R) for some function $\rho(\cdot)$. Therefore, the estimate of the relevant component of either matrix may be written as $\frac{1}{n}\sum_{i=1}^{n}\rho(x_i, y_i, r_i, \hat{\theta}_n)$. In general, continuing to take expected values with respect to the joint distribution of (X, Y, R), for estimate $\hat{\theta}_n$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\rho(x_i, y_i, r_i, \hat{\theta}_n) &= \frac{1}{n}\sum_{i=1}^{n}\rho(x_i, y_i, r_i, \hat{\theta}_n) - \mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)] + \mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)] \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\rho(x_i, y_i, r_i, \hat{\theta}_n) - \mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)]\right| + \mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)] \\
&\leq \underbrace{\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\rho(x_i, y_i, r_i, \theta) - \mathbb{E}[\rho(X, Y, R, \theta)]\right|}_{\text{First part}} + \underbrace{\mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)]}_{\text{Second part}}.
\end{aligned} \tag{1.31}$$

If conditions are met to use the theory due to Jesus and Chandler [2011] to prove asymptotic normality, then we must have the first part of Equation (1.31) converge in probability to 0 for all elements of $T(\theta_0)$. Similarly, if functions $\rho(X, Y, R, \theta_0)$ whose expected values define the components of S are Glivenko-Cantelli [Van der Vaart, 1998, pg 269]), then convergence in probability to 0 of the first part of Equation (1.31) holds true. Also note that if the theory due to Jesus and Chandler [2011] is used to prove asymptotic normality, then all elements $\mathbb{E}[\rho(X, Y, R, \theta)]$ composing $T(\theta)$ are continuous with respect to θ, and hence $\mathbb{E}[\rho(X, Y, R, \hat{\theta}_n)]$ is a consistent of $\mathbb{E}[\rho(X, Y, R, \theta_0)]$ due to the continuous mapping theorem, recalling $\hat{\theta}_n$ was proven to be consistent for θ_0. The same should hold for all elements of S in all but pathological cases.

If requirements are met of Theorem 3.4 by Newey and McFadden [1994, pg 2148]—previously used to prove asymptotic normality in some cases—consistency of the proposed estimator $\hat{\Sigma}$ holds if two additional conditions are met: (1) continuity of $m(X, Y, R, \theta)$ at θ_0

and (2) for a neighborhood $\mathcal{N}$ of θ_0, $\mathbb{E}[\sup_{\theta\in\mathcal{N}} \|m(X,Y,R\theta)^2\|] < \infty$. (See Theorem 4.5 by Newey and McFadden [1994, pg 2160].)

1.5.3 Estimating Values of the Regression Function

To estimate values of $\mathbb{E}[Y \mid x;\theta_0] \equiv \psi(x,\theta_0)$ at the true value θ_0 of θ, we suppose the functional form of the regression $\psi(x,\theta)$ is continuously differentiable with respect to θ. Under this assumption, given $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0,\Sigma)$, we have

$$\sqrt{n}(\psi(x,\hat{\theta}_n) - \psi(x,\theta_0)) \to_d N\left(0, \left[\frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{\theta=\theta_0}\right]^T \Sigma \left[\frac{\partial}{\partial\theta}\psi(x,\theta)\bigg|_{\theta=\theta_0}\right]\right)$$

via the delta method [Van der Vaart, 1998, pg 25-26], for which the asymptotic covariance may be consistently estimated using $\hat{\Sigma}$ (see Section 1.5.2.2) and $\frac{\partial}{\partial\theta}\psi(x,\theta)|_{\theta=\hat{\theta}_n}$ according to the continuous mapping theorem.

Using assumptions laid out in Sections 1.4.2.1-1.4.2.2 about the covariate-dependent response probability $\mathbb{P}[R = 1 \mid x]$ and verifiable density $f(y \mid x, R = 1)$ respectively, the regression function $\mathbb{E}[Y \mid x;\theta]$ takes the form

$$\begin{aligned}\mathbb{E}[Y \mid x;\theta] &= \mathbb{P}[R=1 \mid x;\delta]\,\mathbb{E}[Y \mid x, R=1;\lambda] + \mathbb{P}[R=0 \mid x;\delta]\,\mathbb{E}[Y \mid x, R=0;\lambda,\eta,\boldsymbol{\beta}] \\ &= \frac{\exp(\delta^T x_\delta)}{1+\exp(\delta^T x_\delta)}\mathbb{E}[Y \mid x, R=1;\lambda] \\ &\quad + \frac{1}{1+\exp(\delta^T x_\delta)}\frac{\mathbb{E}[Y\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}{\mathbb{E}[\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]},\end{aligned}$$

using the definition of $\mathbb{E}[Y \mid x, R = 0]$ from Equation (1.23). The gradient $\frac{\partial}{\partial\theta}\psi(x,\theta)$ is simplest to define by four groups of parameter vector θ (according to λ, η, $\boldsymbol{\beta}$, and δ). Each

of these four components of $\frac{\partial}{\partial\theta}\psi(x,\theta)$ in this case may be written as

$$\begin{aligned}
\frac{\partial}{\partial\lambda}\psi(x,\theta) &= \frac{\exp(\delta^T x_\delta)}{1+\exp(\delta^T x_\delta)}\frac{\partial}{\partial\lambda}(\mathbb{E}[Y \mid x, R=1;\lambda]) \\
&\quad + \frac{1}{1+\exp(\delta^T x_\delta)}\frac{\partial}{\partial\lambda}\left(\frac{\mathbb{E}[Y\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}{\mathbb{E}[\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}\right), \\
\frac{\partial}{\partial\eta}\psi(x,\theta) &= \frac{1}{1+\exp(\delta^T x_\delta)}\frac{\partial}{\partial\eta}\left(\frac{\mathbb{E}[Y\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}{\mathbb{E}[\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}\right), \\
\frac{\partial}{\partial\boldsymbol{\beta}}\psi(x,\theta) &= \frac{1}{1+\exp(\delta^T x_\delta)}\frac{\partial}{\partial\boldsymbol{\beta}}\left(\frac{\mathbb{E}[Y\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}{\mathbb{E}[\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}\right), \\
\frac{\partial}{\partial\delta}\psi(x,\theta) &= \frac{x_\delta\exp(\delta^T x_\delta)}{[1+\exp(\delta^T x_\delta)]^2}\left(\mathbb{E}[Y \mid x, R=1;\lambda] - \frac{\mathbb{E}[Y\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}{\mathbb{E}[\exp\{\beta(Y)\} \mid x, R=1;\lambda,\eta,\boldsymbol{\beta}]}\right).
\end{aligned}$$

Partial derivatives with respect to η are removed if η is not estimated.

1.6 Simulated Data Example: Poisson-Distributed Outcomes

We conducted a simulation study to assess performance of this method with Poisson-distributed outcomes and a single continuous baseline covariate. We were particularly interested in sensitivity of coverage of parameter values and regression values to the sample size and the percentage of individuals lost to follow-up. The procedure to generate data is described in Section 1.6.1. Then, taking the perspective of an investigator, we apply the method outlined in Section 1.4 to define the regression function and Section 1.5 to estimate parameter and regression values. Results on coverage are summarized in Section 1.6.4.

1.6.1 Data Generation

Across all generated data sets, we supposed a single baseline covariate X_1 was collected; values of this covariate were generated using a $N(1,1)$ distribution. Observed outcomes Y were generated according to the distribution $[Y \mid x, R=1] \sim \text{Poisson}(\exp\{\lambda^T x_\lambda\})$ where $x_\lambda = (1\ x_1)^T$. We supposed the covariate-dependent response probability follows a logistic regression, i.e. response indicators were realizations of $[R \mid x] \sim \text{Bernoulli}\left(\frac{\exp\{\delta x_\delta\}}{1+\exp\{\delta x_\delta\}}\right)$ where $x_\delta = x_1$, and therefore the regression does not include an intercept term. Data were generated

using δ equal to 2, 1, and 0.2 to get $\mathbb{P}[R = 0]$ equal to 22%, 30%, and 45% respectively.

While not necessary for data generation, for each value of δ, we calculated the *true* values of unobserved marginal means of the outcome. In principle, when employing the novel methodology, functions $h_k(Y)$, $k \in \{1, ..., K\}$ defining the selection model dictate the marginal means we must fix; see Section 1.4.1 for further details. However, to use the GMM outlined in Section 1.5.2, we *directly* depend on estimates of
$\mu_{0k} = \mathbb{E}[h_k(Y) \mid R = 0]$, $k \in \{1, ..., K\}$. Hence, we calculated the true values of these values (as opposed to $\mu_k = \mathbb{E}[h_k(Y)]$) to be used when estimating θ later. In practice, μ_k, $k \in \{1, ..., K\}$ are estimated using auxiliary data, and μ_{0k}, $k \in \{1, ..., K\}$ estimated using a method outlined in Section 1.5.2.1.

To calculate μ_{0k}, $k \in \{1, ..., K\}$, we suppose the selection model falls in the subclass defined in Section 1.4.2.3, further supposing $\beta(y) = \beta y$ (with some abuse of notation) for single coefficient β, i.e. $\boldsymbol{\beta} = \beta$ and is of length one. Therefore,

$$f(y \mid x, R = 0) = f(y \mid x, R = 1)\frac{\exp\{\beta y\}}{\mathbb{E}[\exp\{\beta y\} \mid x, R = 0]}.$$

We kept $\beta = 0.75$ fixed across all generated data sets, and data are therefore missing not at random. Therefore, we must estimate/calculate $\mathbb{E}[Y \mid R = 0]$ to use the outlined GMM procedure in Section 1.5.2 to estimate θ. Note that the support of $[Y \mid R = 0]$ is all nonnegative integers, and therefore

$$\mathbb{E}[Y \mid R = 0] = \sum_{y=0}^{\infty} y\mathbb{P}[Y = y \mid R = 0]. \tag{1.32}$$

Since $[X \mid R = 0]$ has continuous support,

$$\begin{aligned} f(y \mid R = 0) &= \int_x \mathbb{P}[Y = y \mid x, R = 0] f(x \mid R = 0) dx \\ &= \int_x \mathbb{P}[Y = y \mid x, R = 0] \frac{f(x)\mathbb{P}[R = 0 \mid x]}{\mathbb{P}[R = 0]} dx \\ &= \int_x \mathbb{P}[Y = y \mid x, R = 0] \frac{f(x)\mathbb{P}[R = 0 \mid x]}{\int_x f(x)\mathbb{P}[R = 0 \mid x] dx} dx \\ &= \frac{\int_x \mathbb{P}[Y = y \mid x, R = 0] f(x)\mathbb{P}[R = 0 \mid x] dx}{\int_x f(x)\mathbb{P}[R = 0 \mid x] dx}. \end{aligned}$$

Plugging this in to Equation (1.32), we have

$$\begin{aligned} \mathbb{E}[Y \mid R = 0] &= \sum_{y=0}^{\infty} y \frac{\int_x \mathbb{P}[Y = y \mid x, R = 0] f(x)\mathbb{P}[R = 0 \mid x] dx}{\int_x f(x)\mathbb{P}[R = 0 \mid x] dx} \\ &= \frac{\sum_{y=0}^{\infty} y \int_x \mathbb{P}[Y = y \mid x, R = 0] f(x)\mathbb{P}[R = 0 \mid x] dx}{\int_x f(x)\mathbb{P}[R = 0 \mid x] dx}. \end{aligned}$$

While the final form of $\mathbb{E}[Y \mid R = 0]$ depends on known densities, the analytic form is intractable here. To calculate $\mathbb{E}[Y \mid R = 0]$, we estimated all integrals using numerical integration and all sums by adding summands for increasing values of y until convergence was reached (which for us, was the addition of a value to the summand below a quantity recognizable by the R computing software). For each value of δ, we generated a total of 10,000 data sets.

1.6.2 Defining the Regression Function

Taking the perspective of an investigator, we specified functional forms of the distributions of $[Y \mid x, R = 1]$ and $[R \mid x]$; we specified the same distributions used to generate the data. In practice, these assumptions would be empirically verified by the investigator. We also correctly supposed the selection model fell in the subclass defined in Section 1.4.2.3 with linear functional dependence on Y, i.e. $\beta(y) = \beta y$.

Note that by exponential tilting of the verifiable density, $[Y \mid x, R = 0]$ also has a Poisson

distribution indexed by the same log linear function as $[Y \mid x, R = 1]$ with the addition of intercept β. In other words, $[Y \mid x, R = 1] \sim \text{Poisson}(\exp\{\lambda^T x_\lambda + \beta\})$ because

$$\begin{aligned} f(y \mid x, R = 0) &= f(y \mid x, R = 1) \frac{\exp\{\beta y\}}{\mathbb{E}[\exp\{\beta Y\} \mid x, R = 1]} \\ &= \frac{\exp\{-\exp(\lambda^T x_\lambda)\} \exp\{\lambda^T x_\lambda y\}}{y!} \frac{\exp\{\beta y\}}{\exp\{\exp(\lambda^T x_\lambda)(\exp(\beta) - 1)\}} \\ &= \frac{\exp\{y(\lambda^T x_\lambda + \beta)\} \exp\{-\exp(\lambda^T x_\lambda + \beta)\}}{y!}. \end{aligned}$$

Hence, the final complete regression function is

$$\mathbb{E}[Y \mid x] = \left[\frac{\exp(\delta x_1)}{1 + \exp(\delta x_1)} \exp(\lambda^T x_\lambda) \right] + \left[\frac{1}{1 + \exp(\delta x_1)} \exp(\lambda^T x_\lambda + \beta) \right].$$

1.6.3 Estimation Procedure

We fixed $\mu_0 = \mathbb{E}[Y \mid R = 0]$ at the value calculated in Section 1.6.1; see Section 1.5.2.1 for more information on estimating $\mathbb{E}[Y \mid R = 0]$ in practice. Using aforementioned assumptions about the verifiable density, covariate-dependent response probability, and selection model, we may use the following moment vector in a GMM to estimate θ, where each element listed below is of length 1:

$$\begin{aligned} m_{1A}(X, Y, R, \theta) &= R(Y - \exp(\lambda^T X_\lambda)), \\ m_{1B}(X, Y, R, \theta) &= R(Y - \exp(\lambda^T X_\lambda)) X_1, \\ m_2(X, Y, R, \theta) &: NA, \\ m_3(X, Y, R, \theta) &= (1 - R)[\exp(\lambda^T X_\lambda + \beta) - \mu_0], \\ m_4(X, Y, R, \theta) &= X_1 \left(R - \frac{\exp(\delta X_1)}{1 + \exp(\delta X_1)} \right). \end{aligned}$$

What we call the "first component" of the moment vector in Section 1.5.2 is split into the two separate elements, m_{1A} and m_{1B}, for each of the two elements of $\lambda = (\lambda_0, \lambda_1)$, respectively.

Since the dispersion parameter for the verifiable density $\eta = 1$ is known, there is no need to estimate that value, and so the second component of the moment vector remains undefined.

Weighting matrices to use GMM were fixed as recommended in Section 1.5.2, i.e. $W_n = I_{p\times p}$ and $\gamma_n = \tau_n = \frac{1}{\sqrt{n}} I_{p\times p}$. Vector θ is of length 4, and hence $p = 4$. Because $\hat{\lambda}$ and $\hat{\delta}$ are MLEs of parameters governing GLMs with canonical links (i.e. default estimators of parameters governing GLMs with canonical links), solutions to $\hat{\lambda}$ and $\hat{\delta}$ are known to be consistent [Wakefield, 2013, pg 261]; uniqueness of these solutions (supposing full-rank x_λ and x_δ) are proven by Wedderburn [1976] for these choices of link functions. Therefore, we only require proof of uniqueness and consistency of $\hat{\beta}$. First, note that $\hat{\beta}$ may be analytically derived using $\hat{\lambda}$ via

$$\hat{\beta} = \log(\mu_0) - \log\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(1-r_i)\exp(\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i})}{\frac{1}{n}\sum_{i=1}^{n}(1-r_i)}\right). \tag{1.33}$$

Therefore, $\hat{\beta}$ is unique. To prove consistency of $\hat{\beta}$, note that by Lemma 4.3 due to Newey and McFadden [1994, pg 2156-2157], for true values (λ_0, λ_1) governing the data generating mechanism,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(1-r_i)\exp(\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i}) &\to_p \mathbb{E}[(1-R)\exp(\lambda_0 + \lambda_1 X_1)] \\
&= \sum_{r\in\{0,1\}} \mathbb{P}[R=r]\,\mathbb{E}[(1-R)\exp(\lambda_0 + \lambda_1 X_1) \mid R=r] \\
&= \mathbb{P}[R=0]\,\mathbb{E}[\exp(\lambda_0 + \lambda_1 X_1) \mid R=0].
\end{aligned}$$

Plugging this quantity into Equation (1.33), we have

$$\begin{aligned}
\hat{\beta} &= \log(\mu_0) - \log\left(\frac{\mathbb{P}[R=0]\,\mathbb{E}[\exp(\lambda_0+\lambda_1 X_1) \mid R=0] + o_p(1)}{\frac{1}{n}\sum_{i=1}^{n}(1-r_i)}\right) \\
&= \log \mathbb{E}[Y \mid R=0] - \log\left(\frac{\mathbb{P}[R=0]\,\mathbb{E}[\exp(\lambda_0+\lambda_1 X_1) \mid R=0] + o_p(1)}{\mathbb{P}[R=0] + o_p(1)}\right) \\
&= \log \mathbb{E}[Y \mid R=0] - \log(\mathbb{E}[\exp(\lambda_0+\lambda_1 X_1) \mid R=0]) + o_p(1) \\
&= \log \mathbb{E}[Y \mid R=0] + \log(\exp\{\beta\}) - \log(\exp\{\beta\}) - \log(\mathbb{E}[\exp(\lambda_0+\lambda_1 X_1) \mid R=0]) + o_p(1) \\
&= \log \mathbb{E}[Y \mid R=0] + \log(\exp\{\beta\}) - \log(\mathbb{E}[\exp(\lambda_0+\lambda_1 X_1+\beta) \mid R=0]) + o_p(1) \\
&= \log \mathbb{E}[Y \mid R=0] + \log(\exp\{\beta\}) - \log(\mathbb{E}_X[\mathbb{E}_Y[Y \mid X, R=0]]) + o_p(1) \\
&= \log \mathbb{E}[Y \mid R=0] + \log(\exp\{\beta\}) - \log \mathbb{E}[Y \mid R=0] + o_p(1) \\
&= \beta + o_p(1),
\end{aligned}$$

using the law of iterated expectation for the penultimate line (the same logic used to define $m_3(\cdot)$ in Section 1.5.2), thus proving consistency of $\hat{\beta}$. To prove asymptotic normality of the estimates, we turn to Theorem 3.4 by Newey and McFadden [1994, pg 2148], with special attention brought to the note at the bottom of the page: if estimates are consistent and conditions (i)-(v) of the theorem are satisfied, then $\hat{\theta}$ are CAN. These conditions are easily met if assumptions are made about the finiteness of expected values of functions of X and assumptions are further made about nonsingularity of $T'(\theta)T(\theta)$. We assume the latter condition is met.

Using Theorem 3.4 by Newey and McFadden [1994, pg 2148] (or more directly, Lemma 4.3 by Newey and McFadden [1994, pg 2156-2157]), estimates of Σ using sample gradients and moments, plugging in $\hat{\theta}$ for θ, will be consistent (see Section 1.5.2.2). Recall Σ is a function of W, S, and $T(\theta)$. The first matrix W is fixed as an identity matrix, and the second may be estimated using sample moments evaluated at $\hat{\theta}$ (as in Equation (1.30)). The general form of $\hat{T}(\theta_0)$ defined in Section 1.5.2.2 applies here. Since $\eta = 1$ is known, the second row and column are removed from $\hat{T}(\theta_0)$. We retain the same subscripts for all other

matrices composing $\hat{T}(\theta_0)$. The nonzero matrices in this example are

$$\hat{T}_{11} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} -r_i \exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i}\} & \frac{1}{n}\sum_{i=1}^{n} -r_i x_{1i} \exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i}\} \\ \frac{1}{n}\sum_{i=1}^{n} -r_i x_{1i} \exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i}\} & \frac{1}{n}\sum_{i=1}^{n} -r_i x_{1i}^2 \exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i}\} \end{bmatrix},$$

$$\hat{T}_{31} = \frac{1}{n}\sum_{i=1}^{n}(1 - r_i)\exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i} + \hat{\beta}\},$$

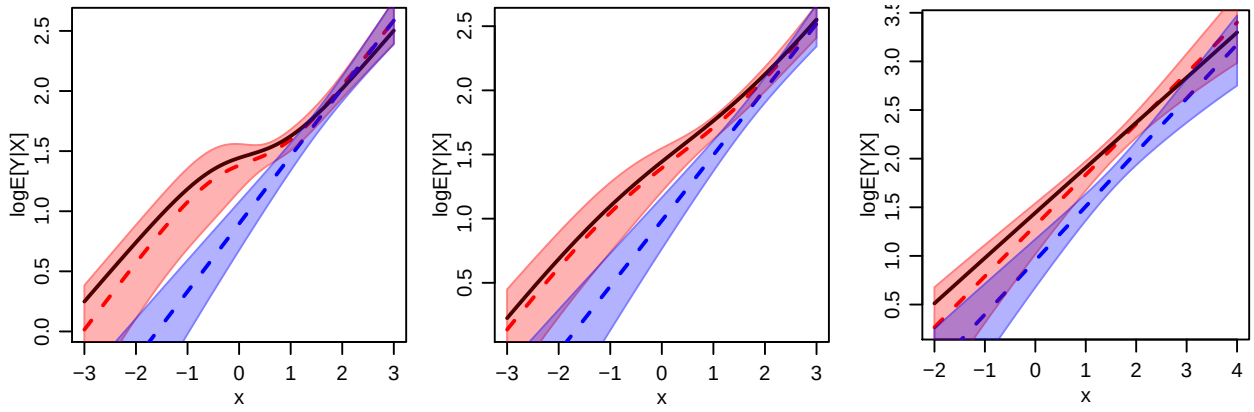
$$\hat{T}_{33} = \frac{1}{n}\sum_{i=1}^{n}(1 - r_i)x_{1i}\exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i} + \hat{\beta}\},$$

$$\hat{T}_{44} = \frac{1}{n}\sum_{i=1}^{n}(1 - r_i)\exp\{\hat{\lambda}_0 + \hat{\lambda}_1 x_{1i} + \hat{\beta}\}.$$

1.6.4 Results

First, to exemplify the impact of erroneously assuming ignorable missingness, Figure 1.1 shows estimates of the regression function for a range of X_1 values and their related 95% pointwise confidence bands for a single generated data set vs. the true regression function (in black). Estimates assuming ignorable missingness (in blue) were acquired using the likelihood approach. Recall that under ignorability, the overall regression function is the same as the observable regression function; we supposed the verifiable density was correctly specified. Then, $\hat{\theta}$ were default estimates of parameters governing a GLM (i.e. using the first component of the moment vector defining the GMM procedure). Estimates allowing for nonignorability acquired using our novel approach are displayed in red. Sample sizes were kept small at $n = 100$ to make the confidence bands visible over a wide range of X_1. Across all three probabilities of missingness, the true regression line is contained in the pointwise confidence band that allowed for nonignorable missingness, but is largely excluded if ignorable missingness is assumed. Moreover, allowing for nonignorable missingness better captures the curvature of the true regression line.

Note that, since choice of W_n does not impact values of estimates, optimal efficiency is achieved for this moment vector. Figure 1.2 shows observed coverage of $\hat{\theta}$ using all 10,000 generated data sets; nominal coverage was 95% using $\hat{\Sigma}$ to calculate the confidence intervals.



(a) $\mathbb{P}(R = 0) = 0.22$ (b) $\mathbb{P}(R = 0) = 0.30$ (c) $\mathbb{P}(R = 0) = 0.45$

Figure 1.1: True regression line in black, with estimates of the regression line and related 95% pointwise confidence bands allowing for nonignorable missingness using our method (in red) and assuming ignorable missingness (in blue); $n = 100$ throughout.

Lower coverage with smaller n appears to be quickly remedied with increasing sample sizes, even with high proportions of individuals missing outcome data. While it is not surprising coverage for parameters indexing the recoverable density is worse as the proportion of missing individuals increases, it is interesting that coverage of $\hat{\beta}$ is relatively constant across $\mathbb{P}[R = 0]$ with fixed sample size.

To show the impact of this coverage on regression values (i.e. expected behavior of the pointwise confidence bands in Figure 1.1), observed vs. nominal (of 95%) coverage across the same 10,000 generated data sets is plotted in Figure 1.3 of $\mathbb{E}[Y \mid x_1]$ for $X_1 \in \{0, 1, 2\}$. These values were selected from high-density areas of X_1, representing values above and below the average. Overall, observed coverage is nearly nominal, even with smaller sample sizes. Like $\hat{\theta}$, low coverage is quickly recovered with increasing n.

1.7 Application to Binary Dental Utilization Outcome

To illustrate this method in practice, we apply our method to an exploratory analysis [Blasi et al., 2018] of data from the Oral Health for Life (OH4L) study [McClure et al., 2018].

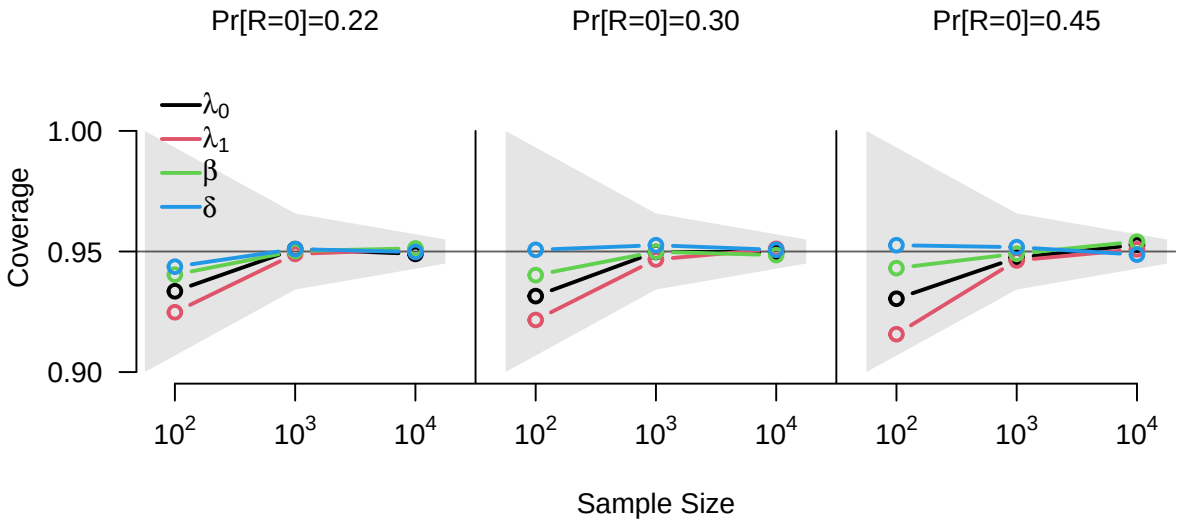


Figure 1.2: Observed coverage (with nominal coverage of 95%) for $\hat{\theta}$ for orders of magnitude increases in sample size; grey area shows *relative* width of 95% confidence intervals. A total of 10^4 data sets were generated to estimate coverage; the covariance matrix for $\hat{\theta}$ was consistently estimated (see Section 1.5.2.2 for more details) to calculate confidence intervals.

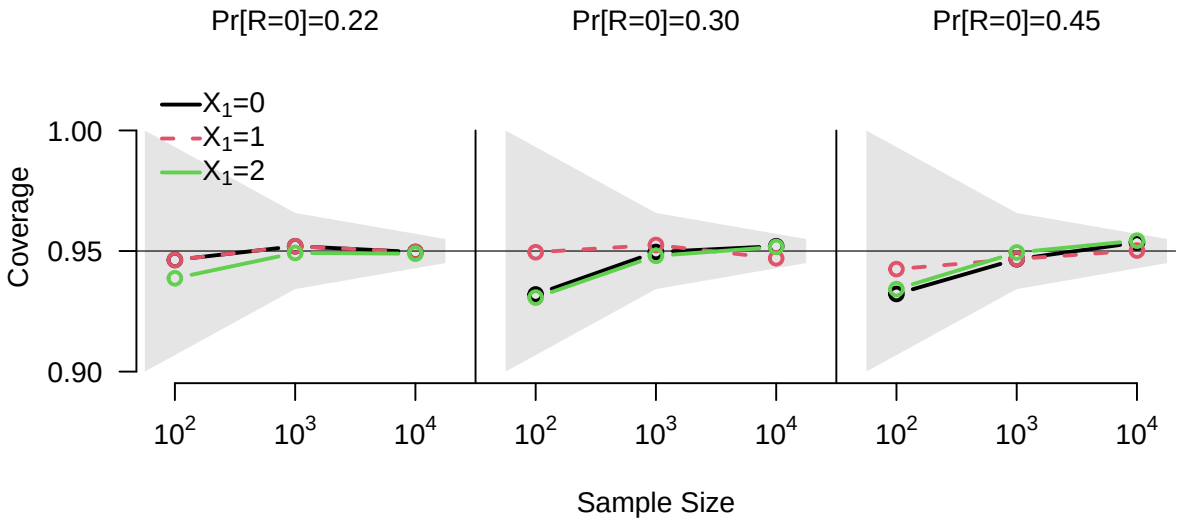


Figure 1.3: Observed coverage (with nominal coverage of 95%) for $\mathbb{E}[Y|X_1 = x_1]$ using pointwise confidence bands for orders of magnitude increases in sample size; grey area shows *relative* width of 95% confidence intervals. A total of 10^4 data sets were generated to estimate coverage; the covariance matrix for $\hat{\theta}$ was consistently estimated (see Section 1.5.2.2 for more details) to calculate confidence intervals.

OH4L's study population was composed of smokers motivated to quit, who were disproportionately low-income. Low-income smokers face disproportionate rates of oral health disease [Office of the Surgeon General and National Institute of Dental and Craniofacial Research (US), 2000], and tend to be delinquent in seeking timely dental care [Bloom et al., 2012, Drilea et al., 2005]. Little research is available on methods to encourage professional dental healthcare utilization in this target population [Blasi et al., 2018]. Exploratory analyses by Blasi et al. [2018] aimed to fill this knowledge gap, depending on single imputation for missing outcome data.

The rest of this section proceeds as follows. First, we will summarize data from OH4L, the methods used by Blasi et al. [2018] to find determinants of seeking professional dental care, and results of this exploratory analysis. Next, we will use our novel approach to answer the same scientific question using the same data. The methods used to specify the verifiable density, covariate-dependent response probability, and auxiliary marginal information will be laid out. Finally, the estimated impacts (and related inference) of covariates on seeking professional dental care using our approach are compared to estimates and inference from Blasi et al. [2018]. We also assess the sensitivity of the novel results to values of auxiliary marginal information, akin to a tipping point analysis.

1.7.1 Oral Health for Life Data

The original study comprised 718 individuals calling into smoking quitlines—free services provided by states offering coaching and connection to smoking-cessation resources [CDC, 2020]. In addition to answering extensive questions about demographics, health, dental insurance/cost, and personal feelings about dentistry at baseline, individuals were asked 2 and 6 months after baseline if they had sought professional dental care over the past 6 months. Individuals were included in the study only if they had *not* sought professional dental care over the past 6 months at baseline, and were therefore overdue for a visit to a dental professional (i.e. if all individuals followed recommended dental care guidelines, 100% of study participants would have seen the dentist by the end of follow-up). Due to loss to

follow up, sufficient information to determine if someone sought professional dental care over the past 6 months at the end of the study was collected from 531 (74%) of the individuals included in the study at baseline [McClure et al., 2018]. Among individuals with collected outcome data, 25% sought professional dental care over the course of follow-up.

1.7.2 Original Exploratory Analysis

Blasi et al. [2018] wished to use these data to find determinants of seeking professional dental care among low-income smokers, aiding the development of future targeted behavioral interventions. In particular, Blasi et al. [2018] wanted to find determinants of dental care utilization beyond financial factors (e.g. dental insurance and cost of dental care).

Supposing loss to follow-up indicated lack of personal motivation, Blasi et al. [2018] imputed missing dental utilization data as "none", i.e. the original analysis assumed 718-531=187 individuals did not seek professional dental care over the course of follow-up.

The search for determinants of professional dental care utilization entailed two stages. First, for each covariate of interest, relative risks (RRs) of the outcome—an indicator of whether they sought professional dental care during the 6-month follow-up—with changes in the covariate were estimated using separate Poisson regression models (see Table 3 by Blasi et al. [2018] for a complete list). If covariates were deemed significantly associated with the outcome at the 0.10 level in these models, they were included in a single multivariate Poisson regression model with the same outcome, the source of the final estimates of relative risk and inference. All models adjusted for financial covariates and potential confounders. (See Table 1.1 for a complete list.) Blasi et al. [2018] used generalized estimating equations to estimate coefficients and conduct inference with robust standard errors. In all regression models, the aforementioned single imputation was used for missing outcomes.

The column labeled Original in Table 1.1 contains p-values from omnibus Wald tests for significance of at least one association between any covariate in the listed category (all categorical) and risk of the outcome of seeking professional dental care during follow-up. For a complete list of point estimates of relative risks, see Table 4 of Blasi et al. [2018] or the

depictions in Figures 1.4-1.6 in this chapter (labeled with log(RR)). The lack of significance of financial covariates was somewhat surprising. In the final regression model, both being married (RR 0.67) and living with a disability (RR 1.63) had point estimates of effect in the direction opposite of expected, and were both deemed significant at the 0.05 level used for the final inference. Sensitivity of these results to the assumed missingness mechanism is of particular interest.

	Categories of Covariates	X_δ	X_λ	**Original**	**New**
Potential Confounders	Age	✓	✓	0.96	0.032
	Gender		✓	0.83	0.765
	State		✓	0.45	0.208
Financial	Cost barrier		✓	0.88	0.865
	Insurance barrier		✓	0.61	0.995
	Dental insurance		✓	0.33	0.104
Additional	Marital status	✓	✓	0.02	0.010
	Education	✓	✓	0.013	0.023
	Living with a disability	✓	✓	0.004	0.004
	Time since last dental cleaning	✓	✓	< 0.001	0.011
	Motivation to see dentist in next 6 months		✓	0.029	0.001
	Self-efficacy to see dentist in next 6 months		✓	0.019	$<$0.001
	Perceived gum disease			0.014	n/a
	Number daily cigarettes	✓		n/a	0.020

Table 1.1: For each row, p-values from the omnibus Wald test for the null hypothesis [H_0 : all coefficients for covariates in the category equal to 0] using the analytic approach of Blasi et al. [2018] and our novel approach. The first two columns mark covariates included in the novel approach's covariate-dependent response probability model (X_δ) and those included in the observable regression model (X_λ).

1.7.3 Applying the New Method

To use our new method with these data, we must specify (1) auxiliary marginal information, (2) the verifiable density, and (3) the covariate-dependent response probability model.

Each will be considered in turn. Throughout, outcome Y refers to the indicator of seeking professional dental care over the past six months.

1.7.3.1 Auxiliary Marginal Information

Wording for the question on professional dental healthcare utilization in OH4L came from the "sample adult" questionnaire from the National Health Interview Survey (NHIS), an annual survey jointly conducted by the Centers for Disease Control and U.S. Census Bureau [Botman et al., 2000]. Available data releases for NHIS include weights for each participating individual, allowing for generalization of estimates to the U.S. resident civilian non-institutionalized population. The NHIS data release from the calendar time of OH4L serves as a natural source of auxiliary marginal information. We will fix only one marginal mean $\mathbb{E}[Y]$, or the probability of seeking professional dental care over the past 6 months. This choice is particularly appropriate for a binary outcome, noting any choice of functional dependence of the selection model on Y in Equation 1.11 would simplify to be linear in Y.

Respondents from the 2017 sample adult NHIS survey were first selected and then reweighted to make estimates generalizable to the OH4L target population. We focused on two criteria to make estimates generalizable to the OH4L target population: (1) the desire to quit smoking and (2) income status. It is critical to focus on variables with different distributions among respondents to the NHIS and OH4L that *also* impact the propensity for the outcome. The first variable is an inclusion criteria for OH4L and potentially impacts the outcome; the necessity of accounting for the latter variable is investigated below. We assumed the target populations were otherwise similar in distribution.

To address the first criteria, we selected respondents to the 2017 NHIS only if they attempted to quit smoking over the past year, resulting in data from a total of 1897 people. This number may suggest there is little to gain by using NHIS data to estimate $\mathbb{E}[Y]$ over OH4L. However, recall that *all* data is collected from individuals at the same time from NHIS. Therefore, we anticipate the motivation to respond to NHIS is similar to OH4L at baseline (as opposed to the OH4L follow-up call). Furthermore, individual responses to NHIS

come with carefully considered weights, which when thoughtfully utilized, allow for estimates generalizable to the US civilian population who attempted to quit smoking in the past year; the methodology is written in more detail below.

To address the second criteria, the selected NHIS respondents were stratified by income. Annual *family* income was collected from NHIS, while annual *household* income was collected from OH4L. However, since 98% of those surveyed in NHIS lived in single-family homes, household income was assumed to be the same as family income. Income data were available on 95.20% of respondents who attempted to quit smoking in the past year. Since the categorical responses from OH4L and NHIS did not align, we divided NHIS respondents into two groups according to income categories that were similar between OH4L and NHIS: less than \$35,000/year (similar to the \$40,000/year distinction from OH4L) and more than \$35,000/year.

For individuals who attempted to quit smoking in the past year, separate estimates of the probability of seeking professional dental care in the past six months were acquired by binary annual income category. These estimates used modified weights allocated to each individual response; the *new* weights were equal to the original weight provided by NHIS divided by the sum of NHIS weights in that income category. Accounting for income proved quite important: 42% of individuals earning over \$35,000 sought professional dental care, as opposed to only 23% of those earning less. With 87% of participants in OH4L making less than \$40,000/year, while an estimated 38.87% of smokers who attempted to quit smoking in the past year in the US earned less than \$35,000/year, re-weighting estimates from NHIS according to income is crucial. Hence, the final estimate of the marginal mean was acquired by the weighed sum of the probability of seeking dental care among individuals making more and less than \$35,000 annually: $(0.87 \times 0.23) + (0.13 \times 0.42) = 0.25$.

Two inclusion criteria of note for OH4L could not be accounted for: state of residence (Louisiana, Nebraska, or Oregon) and being delinquent in seeking professional dental care at baseline (i.e. 2-6 months prior to being questioned). While the former is expected to have little impact, the latter might be important. NHIS might include more individuals with

regularly-scheduled professional dental visits. For this reason, sensitivity analyses will be conducted across a range of values of the probability of seeking professional dental care.

With an estimated 25% of those not lost to follow-up seeking professional dental care (i.e. we estimate $\mathbb{E}[Y \mid R = 1]$ to be 0.25), and assuming 25% of individuals in the OH4L target population sought professional dental care, the estimate of $\mu_0 = \mathbb{E}[Y \mid R = 0]$ is now also 25%. Hence, this value of the marginal mean suggests propensity to seek professional dental care is *marginally* identical between observed and unobserved individuals. This value differs substantially from the value of $\mu_0 = 0$ implied by the imputation of Blasi et al. [2018].

Assuming this estimated equivalence holds, it does *not* imply data are MAR. Thus, to avoid bias, we must consider the possibility data are MNAR. This topic is discussed in detail using pertinent examples in Appendix A.1.

1.7.3.2 Verifiable Density and Covariate-Dependent Response Probability

The Poisson regression used by Blasi et al. [2018] was not available for us to use since we require a fully parametric model for the binary outcome. We opted instead to specify the verifiable distribution and the distribution of response depending on X (dictated by the covariate-dependent response probability) to be Bernoulli with logit link mean functions, i.e.,

$$\begin{aligned} \text{logit}\left(\mathbb{P}[R = 1 \mid x]\right) &= \delta^T x_\delta, \\ \text{logit}\left(\mathbb{P}[Y = 1 \mid x, R = 1]\right) &= \lambda^T x_\lambda, \end{aligned}$$

for covariates X_δ and X_λ collected at baseline. Both vectors of covariates include 1 as the first value (for an intercept) and may include common covariates. Covariates X_λ were chosen to be identical to those in the analysis by Blasi et al. [2018] to ensure inclusion of covariates assessed by Blasi et al. [2018] in our final model to compare the results. This model

also happens to have nearly the highest AIC among all combinations of possible covariates adjusting for confounders and financial covariates. The collection of covariates, X_δ, was that yielding the highest AIC in logistic regression. For complete lists of covariates included in each model, see Table 1.1.

Upon re-examination of the data, self-reported belief in current affliction with gum disease and income were removed from consideration for either model due to non-negligible nonresponse at baseline. We leave adjustment for these covariates to future research, with special consideration regarding reasons for nonresponse to these questions at baseline (and any potentially necessary MNAR adjustment).

The logit link function was chosen for convenience for this illustration, and as mentioned, choices of covariates defining the verifiable density were chosen to match covariates included in the original regression. However, validity of the novel method does depend on correct specification of the verifiable density and covariate-dependent response probability model. In practice, we recommend empirical verification of both the link function and choices of covariates in both of these models. We illustrate how models may be verified with binary outcomes using the verifiable density in Appendix A.2.

1.7.3.3 Unverifiable Density and Regression Function

We use the method outlined in Section 1.4.2.3 to define the unverifiable density. As discussed in Section 1.7.3.1, we fix a single marginal mean of the outcome. Therefore, dependence on Y in Equation 1.11 is linear indexed by a single coefficient β. The unverifiable density is therefore an exponential tilt of the verifiable density dictated by the value of β. Extending the example used in Appendix A.1 to multivariate X, the unverifiable distribution is also Bernoulli with a logit link linear mean model, taking the same functional form as the verifiable density with the additional intercept β. Adding these densities to Equation 1.6, the final regression function takes the form

$$\mathbb{E}[Y \mid x] = \left[\underbrace{\frac{\exp(\delta^T x_\delta)}{1+\exp(\delta^T x_\delta)}}_{\text{CDRP}} \underbrace{\frac{\exp(\lambda^T x_\lambda)}{1+\exp(\lambda^T x_\lambda)}}_{\text{Observable regression}} \right] + \left[\underbrace{\frac{1}{1+\exp(\delta^T x_\delta)}}_{\text{1-CDRP}} \underbrace{\frac{\exp(\lambda^T x_\lambda + \beta)}{1+\exp(\lambda^T x_\lambda + \beta)}}_{\text{Unobservable regression}} \right]. \tag{1.34}$$

Unfortunately, no manipulation of $\mathbb{E}[Y \mid x]$ that is linear is analytically tractable and interpretable; therefore, this regression function offers no natural estimate of association using a single parameter. We have chosen to calculate relative risks to enable comparisons to the original analysis; however, the association between covariates and the outcome is not estimable using a single relative risk. Section 1.7.3.5 below includes recommended illustration of estimated relative risks using the regression function in Equation (1.34), and how to compare them to the original analysis.

1.7.3.4 Generalized Method of Moments

The moment vector to define the GMM may be written as

$$\begin{aligned}
m_1(X,Y,R,\theta) &= R\,X_\lambda \left(Y - \frac{\exp(\lambda^T X_\lambda)}{1+\exp(\lambda^T X_\lambda)} \right), \\
m_2(X,Y,R,\theta) &: \text{n/a}, \\
m_3(X,Y,R,\theta) &= (1-R) \left[\frac{\exp(\lambda^T X_\lambda + \beta)}{1+\exp(\lambda^T X_\lambda + \beta)} - \mu_0 \right], \\
m_4(X,Y,R,\theta) &= X_\delta \left[R - \frac{\exp(\delta^T X_\delta)}{1+\exp(\delta^T X_\delta)} \right],
\end{aligned} \tag{1.35}$$

with the first and fourth components of the same lengths as λ and δ, respectively. The dispersion parameter η for the distribution of $[Y \mid x, R=1]$ is known to be one and does not require estimation; hence $m_2(X,Y,R,\theta)$ remains undefined. Component $m_3(X,Y,R,\theta)$ is of the same length as coefficients $\boldsymbol{\beta}$ indexing $h(Y)$ in Equation 1.11, and hence is of length 1 here. Weighting matrices were fixed as recommended in Section 1.5.2, i.e. $W_n = I_{p\times p}$ and

$\gamma_n = \tau_n = \frac{1}{\sqrt{n}} I_{p\times p}$.

Since λ and δ govern GLMs with canonical links, and the GMM procedure recovers MLEs of these estimators (i.e. default estimates of parameters indexing GLMs), consistency of $\hat{\lambda}$ and $\hat{\delta}$ are readily apparent [Wakefield, 2013, pg 261]. Wedderburn [1976] proves uniqueness of these solutions with the logit link provided x_δ and x_λ are full-rank. To prove uniqueness of $\hat{\beta}$, note that

$$\frac{\partial}{\partial\beta} \sum_{i=1}^{n_{\text{unobserved}}} \frac{\exp(\hat{\lambda}^T x_{\lambda,i} + \beta)}{1 + \exp(\hat{\lambda}^T x_{\lambda,i} + \beta)} = \sum_{i=1}^{n_{\text{unobserved}}} \frac{\exp(\hat{\lambda}^T x_{\lambda,i} + \beta)}{[1 + \exp(\hat{\lambda}^T x_{\lambda,i} + \beta)]^2} > 0, \text{ with}$$
$$\lim_{\beta\to-\infty} \sum_{i=1}^{n_{\text{unobserved}}} \frac{\exp(\hat{\lambda}^T x_{\lambda,i} + \beta)}{1 + \exp(\hat{\lambda}^T x_{\lambda,i} + \beta)} = 0, \text{ and } \lim_{\beta\to\infty} \sum_{i=1}^{n_{\text{unobserved}}} \frac{\exp(\hat{\lambda}^T x_{\lambda,i} + \beta)}{1 + \exp(\hat{\lambda}^T x_{\lambda,i} + \beta)} = 1.$$

Therefore, for any μ_0 bounded away from 0 and 1, there exists a single estimate $\hat{\beta}$ for a given $\hat{\lambda}$ such that $\sum_{i=1}^{n} m_3(x_i, y_i, r_i, \theta) = 0$, and is therefore a unique solution to the GMM. Given uniqueness, Theorem 2.7 by Newey and McFadden [1994, pg 2133] may be applied to prove consistency of $\hat{\theta}$ without requiring compactness of the parameter space. To prove asymptotic normality, Theorem 3.4 by Newey and McFadden [1994, pg 2148] may again be applied; any requirements regarding finiteness of moments of X and Y are naturally fulfilled by virtue of them being categorical. See Section 1.5.2.2 for guidance on consistent estimation of the covariance matrix Σ; Section 1.6.3 contains a more detailed example that is useful for reference.

1.7.3.5 Results

We begin by comparing inference via significance tests using the novel and original inferential techniques. However, the primary focus in both our analysis and the original is on point estimates of relative risks and related confidence intervals, discussed in further detail later in the section. Over-interpretation of significance tests in exploratory analyses with many tests conducted simultaneously should generally be avoided. We further caution against reliance on significance tests using the novel approach informed by the comparison of results

of significance tests and estimated relative risks at the end of the section.

To determine significance of covariates in the final regression model, we conducted omnibus Wald tests for each covariate; the null hypotheses was defined as equivalence to 0 of *all* coefficients for a given covariate. If d elements of θ are being tested simultaneously, the null hypothesis may be written as $H_0 : B\theta = 0$ for $d \times p$ matrix B; each row of B takes the value 1 corresponding to each component of θ being tested and is equal to 0 otherwise. Using estimate of the covariance matrix $\hat{\Sigma}$ and estimate of parameter vector $\hat{\theta}$, the test statistic is

$$(B\hat{\theta})^T[B(\hat{\Sigma}/n)B^T]^{-1}(B\hat{\theta}),$$

which asymptotically has a χ^2_d distribution. See column "New" in Table 1.1 for p-values from these tests and how they compare to p-values from tests of significance of covariates conducted by Blasi et al. [2018]. Note that inference for the novel methodology does not depend on the assumed value of μ_0 in this case, since coefficients to any covariate in the final regression model do not depend on β; coefficients are only components of λ and δ, which are estimated using complete data. Hence, a "tipping point" analysis—generally defined as assessing sensitivity of estimation/inference to the assumed missingness mechanism, which is directly impacted by fixed values of auxiliary marginal information—impacts estimation only, not inference.

Significance of covariates—defined as the p-value being below the type I error rate threshold of 0.05—remained largely unchanged from the original analysis, with a few notable exceptions:

- At least one age-related coefficient was deemed to significantly depart from zero.
- Significance of the coefficients related to time since an individual's last dental cleaning was retained, but not as strongly.
- The reverse holds true for motivation and self-efficacy to see a dentist in the next 6 months.

- Number of daily cigarettes was divided into three categories and was only included in the covariate-dependent response probability model; at least one coefficient was deemed to be significantly different from zero. This covariate failed to reach the necessary p-value threshold to be included in the original analysis.

Unlike $\hat{\lambda}$ and $\hat{\delta}$, both estimates and inference for $\hat{\beta}$ change with fixed values of μ_0. Table 1.2 contains the estimated relative odds $\exp(\hat{\beta})$ of being *missing* from follow-up for those who sought professional dental care compared to those who did not, all else being equal at baseline. The first column is the probability unobserved individuals sought professional dental care; the second column is a reversal of the conditioning, being a function of the probability individuals who sought professional dental care were unobserved. Note that the novel methodology prohibits fixing $\mu_0 = 0$—the assumption made by Blasi et al. [2018] in their single imputation procedure—because a solution to $\hat{\beta}$ cannot be found. However, with $\mu_0 = 0.005$, fewer than one individual of the 187 individuals lost to follow-up will be expected to have sought professional dental care, approximating the assumptions made by Blasi et al. [2018]. The assumption effectively supposes all individuals lost to follow-up did not seek professional dental care; therefore, the estimated probability of being lost to follow-up among individuals who did not seek professional dental care is the proportion of individuals who either were lost to follow-up or had observed negative outcomes who were lost to follow-up, which is a large proportion. As a further consequence of imputing the outcome among non-respondents, we also effectively suppose all individuals with positive outcomes were observed, so that finally the estimated odds ratio of 0.01 under this assumption suggests individuals who sought professional dental care were much less likely to be lost to follow-up compared to those who did not. The fourth row of Table 1.2 supposes the probability unobserved individuals sought professional dental care is equal to the value acquired from NHIS data (see Section 1.7.3.1); this inference suggests β may not differ from 0 and there is insufficient evidence (with provided assumptions) to claim data are missing not at random. In effect, under these assumptions, the observable regression function approximates the overall

regression function $\mathbb{E}[Y \mid x]$ quite well, and data are nearly ignorable.

μ_0	OR (95% CI)
0.005	0.01 (0.009,0.016)
0.08	0.22 (0.17,0.29)
0.16	0.53 (0.41,0.69)
0.25	1.01 (0.78,1.31)
0.50	3.71 (2.84,5.03)
0.99	626.17 (411.91,951.87)

Table 1.2: For a given fixed $\mu_0 = \mathbb{P}[Y = 1 \mid R = 0]$, the estimated relative odds (Wald-type 95% confidence interval) of being lost to follow-up for individuals who sought professional dental care compared to those who did not, all else being equal at baseline.

The final regression model using the novel methodology does not provide a single point estimate of the relative risk (RR) of the outcome for a change in a single covariate, keeping all others constant. However, for any combination of covariates $X = x$, $\mathbb{P}[Y = 1 \mid x]$ may be estimated using Equation 1.34 evaluated using $\hat{\theta}$. Therefore, the RR for a change in a single covariate for a given combination of *all remaining* covariates may be estimated, e.g. for a binary covariate X_l and a total of L covariates, we may estimate

$$\frac{\mathbb{P}[Y = 1 \mid X_1 = x_1, X_2 = x_2, ..., X_l = 1, ..., X_L = x_L]}{\mathbb{P}[Y = 1 \mid X_1 = x_1, X_2 = x_2, ..., X_l = 0, ..., X_L = x_L]}. \tag{1.36}$$

For tertiary covariates, we may have $X_l = 2$ in the numerator. For each covariate X_l, we estimated the RR in Equation 1.36 for every possible combination of covariates $\{X_1, ..., X_{l-1}, X_{l+1}, ..., X_L\}$—a total of 13,824 combinations for each binary covariate and 9216 for each tertiary covariate. This process was repeated for each value of μ_0 in Table 1.2. Violin plots of the estimated log(RR)s are displayed in Figures 1.4-1.6, using default kernel density smoothers of the *violins* function [Schruth, 2013]; despite the use of a kernel density smoother, we note the estimated RRs are *not* realizations of a sampling distribution. Horizontal lines labeled in the plotting margins with estimates of the log(RR) from Blasi et al. [2018] are superimposed on the figures, in addition to confidence intervals displayed

using a grey band. While we do not display confidence intervals using our approach, envisioning confidence intervals with approximately the same width as the original inference centered at a given point estimate approximates overlap in confidence intervals using both approaches. First, we examine the sensitivity of estimated RRs to values of μ_0 using the novel methodology as part of a tipping point analysis; then, we compare estimates from the novel methodology to results from Blasi et al. [2018].

Across values of $\mu_0 \leq 0.50$, estimated RRs remain quite similar for most covariates, suggesting insensitivity to values of μ_0 below 0.50. The estimated value of $\mu_0 = 0.25$ from NHIS data falls in the middle of this wide range, providing ample support for RRs estimated using this fixed marginal mean. Among values of μ_0 used, the only notable shift in estimated RRs occurs for most covariates when $\mu_0 = 0.99$ (or when unobserved individuals are assumed extremely likely to seek professional dental care). Even for $\mu_0 = 0.99$, estimated RRs are typically on the same side of RR=1 as estimated RRs when $\mu_0 \leq 0.50$. Exceptions to this trend are covariates related to age, education, and disability status—with downward trends in estimated RRs as μ_0 gets bigger—as well as the number of daily cigarettes—with upward trends in estimated RRs as μ_0 gets bigger. This sensitivity is particularly important when the estimated value of the relative risk compared to 1 changes. For instance, with $\mu_0 = 0.005$ (approximating assumptions made by Blasi et al. [2018]), individuals aged 45-64 were estimated to be more likely to seek professional dental care compared to those below the age of 45; however, as unobserved individuals become more likely to seek professional dental care, we estimate the opposite trend holds true. Reversal of the direction of relative risk occurs for values of μ_0 as small as 0.25, the value of the marginal mean estimated from NHIS.

Estimates of RRs from Blasi et al. [2018] typically fall in the range of RRs from our methodology across any μ_0, but sometimes take values estimated for a small portion of fixed covariates $\{X_1, ..., X_{l-1}, X_{l+1}, ..., X_L\}$. For better interpretation of the comparison, we define "similarity" in estimation to be proximity of estimated RRs from Blasi et al. [2018] to estimates from our methodology with the highest frequency among the 13,824 or

9216 estimates of relative risk. We note the scale of the vertical axis is particular to each covariate, and therefore estimates may appear more dissimilar than they are. In general, point estimates of relative risks from our approach fall well within confidence intervals from the original analysis, i.e. values of relative risk we estimate are deemed compatible with the data using the original analysis. Knowing confidence intervals using our approach will be centered at these point estimates—implying overlap in confidence intervals with the original approach—interpretation of results does not differ substantially from the original for most covariates across any assumed μ_0.

We compare estimates in more granular detail with $\hat{\mu}_0 = 0.25$ from NHIS data. For this value of μ_0, estimates are quite similar, with the exception of the comparison of Oregon to Louisiana residents (our estimates were higher) and the comparison of individuals over age 65 to those under age 45 (our estimates were lower) in Figure 1.4 and the comparison of individuals with dental insurance compared to those without (our estimates were higher) in Figure 1.5. However, note all estimates of relative risk for any of these covariates lies well within the original confidence interval. This suggests some extremity in relative risk estimation was lost by Blasi et al. [2018]. In our analysis, this extremity was deemed significant and nearly significant for age and dental insurance respectively, both of which were far from significant in the analysis by Blasi et al. [2018]. Therefore, using the 0.05 threshold for p-values, this novel approach may provide slightly more evidence for two additional determinants of professional dental care compared to the original analysis. However, with substantial overlap in confidence intervals using these approaches, final interpretation does not markedly differ between the two approaches with $\mu_0 = 0.25$.

RRs estimated using the novel approach for marital status are less than one across any value of μ_0, retaining the surprising direction of point estimates in the original analyses conducted by Blasi et al. [2018]; granted, confidence intervals may have upper bounds exceeding 1 using the novel approach. On the other hand, RRs for disability status from the new approach may be less than 1, reversing the estimated positive association between disability status and the outcome from Blasi et al. [2018]; however, μ_0 must be larger than 0.50 for the

estimated RR to be less than 1 in the novel approach. Recall, if $\mu_0 = 0.50$, then unobserved individuals are expected to be twice as likely as observed individuals on average to seek professional dental care. Therefore, it appears these surprising results are not sensitive to the missingness mechanism within a reasonable range of assumptions, and we fail to provide evidence that overturns them.

When comparing results of the novel methodology to the original analysis conducted by Blasi et al. [2018], we might expect the greatest similarity in estimates when $\mu_0 = 0.005$, approximating the implied value of $\mu_0 = 0$ from the original single imputation procedure conducted by Blasi et al. [2018]. This is not always the case; stark examples are RRs for residency in Nebraska vs. Louisiana in Figure 1.4 and indications of dental insurance being a barrier to seeking dental care in Figure 1.5. In fact, for some covariates, our estimates of RRs are strikingly similar to results due to Blasi et al. [2018] when unobserved individuals are assumed to generally seek professional dental care (i.e. $\mu_0 = 0.99$); see for example indicators of dental insurance in Figure 1.5.

Therefore, we see results are sensitive to the missingness mechanism *and* the choice of the final regression model, and comparisons between our results and those due to Blasi et al. [2018] should be interpreted carefully. An ideal procedure, particularly for *tipping point analyses*, would be to define a primary model that extends to nonignorable missingness mechanisms. For example, using our approach, the primary model might assume $\boldsymbol{\beta} = 0$ is known and fixed (for ignorable missingness) or assume a fixed value for μ_0 (e.g. $\mu_0 = 0.005$ to approximate the single-imputation assumption of Blasi et al. [2018] or $\mu_0 = 0.25$ from NHIS data). Then, we may assess how much $\boldsymbol{\beta}$ and/or μ_0 must change relative to the primary model to "tip" estimation by a meaningful amount (or, in this case, flip the sign of association). Hence, all changes in inference are due to changes in $\boldsymbol{\beta}$ or μ_0 only, allowing for easier interpretation.

The use of the term *association* is intentionally avoided in the description of the novel results. In generalized linear models, point estimates of coefficients are estimates of a change in a function of the outcome with changes in covariate values, such as relative risks, risk

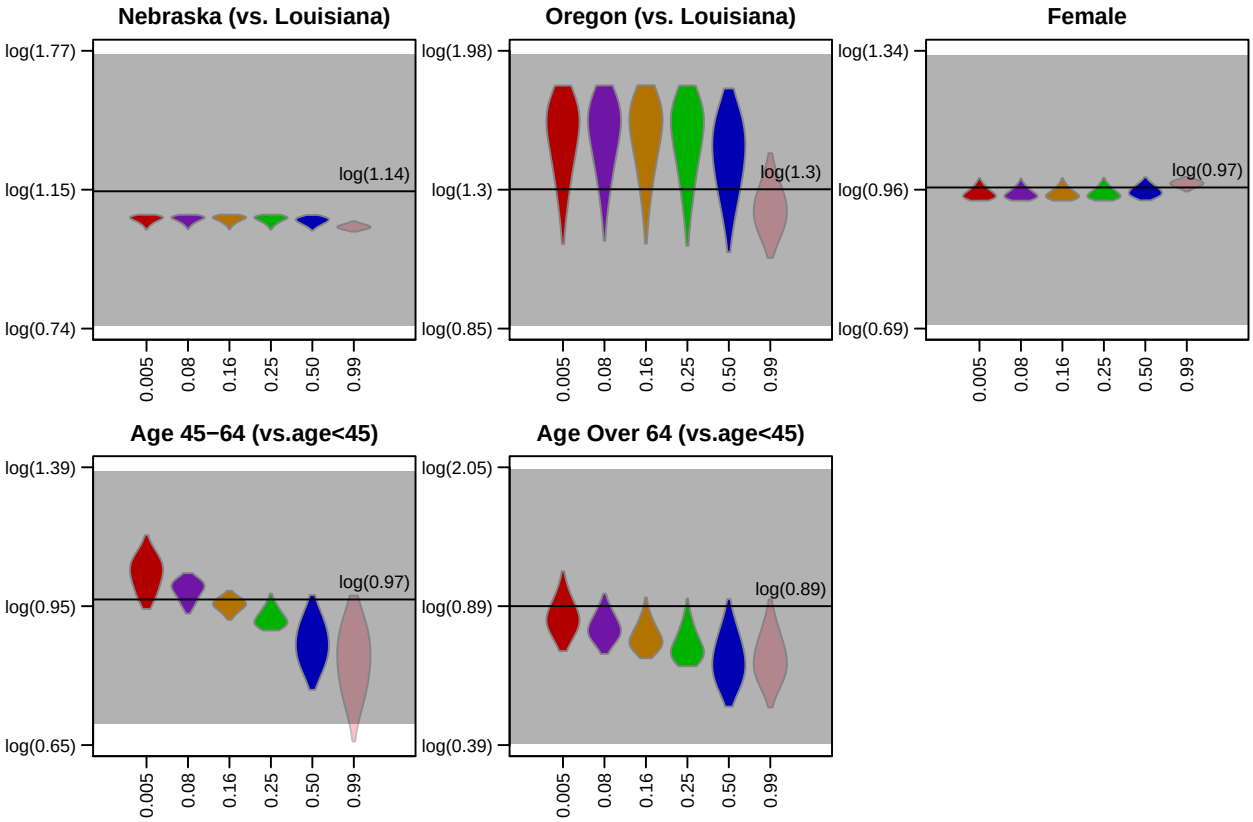


Figure 1.4: For values of μ_0 on the horizontal axis, smoothed frequencies of estimates of relative risks for the entitled confounder (compared to the absence of the covariate in the title, unless otherwise noted) for each possible combination of fixed remaining covariates in the final regression model; horizontal lines are drawn for point estimates of relative risks due to Blasi et al. [2018] and labeled in the plotting area; grey bands display 95% confidence intervals due to Blasi et al. [2018]

.

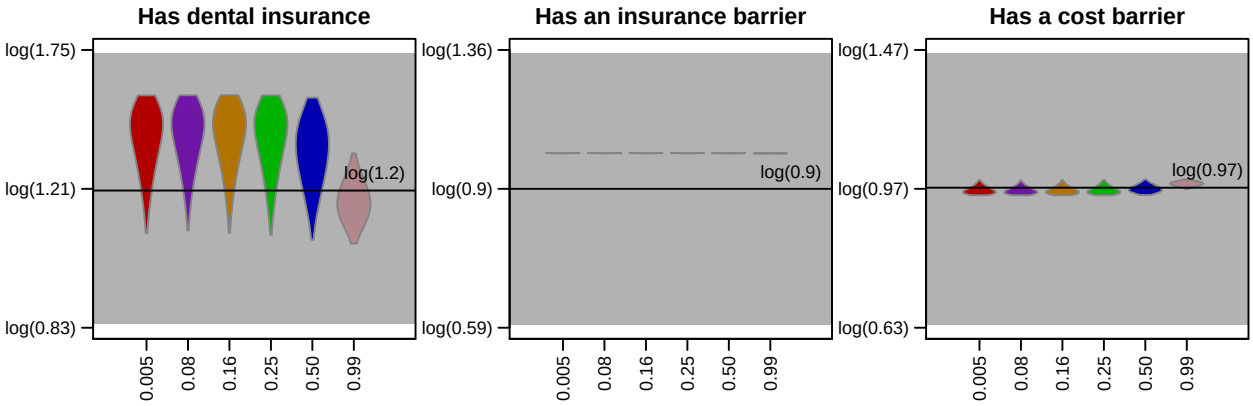


Figure 1.5: For values of μ_0 on the horizontal axis, smoothed frequencies of estimates of relative risks for the entitled financial covariate (compared to the absence of the covariate in the title, unless otherwise noted) for each possible combination of fixed remaining covariates in the final regression model; horizontal lines are drawn for point estimates of relative risks due to Blasi et al. [2018] and labeled in the plotting area; grey bands display 95% confidence intervals due to Blasi et al. [2018]. Point estimates of relative risks using the novel approach were approximately 1 across all assumed μ_0.

differences, or odds ratios. In our case, a coefficient may be significantly different from zero, but lead to negligible expected changes in a function of the outcome of interest, for reasons not attributable to scaling.

"Number of daily cigarettes" in this analysis is an example of this phenomenon. When $\mu_0 = 0.25$, all point estimates of relative risk for daily cigarettes between 11-20 and over 20 compared to 10 or fewer are nearly equal to 1 (see Figure 1.6). However, point estimates of coefficients for these covariates (only in the response probability model) are -0.81 and -0.97, respectively, and with a significant overall test. With $\hat{\beta}$ near 0 when $\mu_0 = 0.25$, the response probability has little impact on the overall regression model (since the observable and unobservable regression are nearly identical). Since cigarette smoking only appears in the covariate-dependent response probability model, it is effectively removed from the final regression function, despite its significant association with probability of response. With different choices of μ_0, inference for "number of daily cigarettes" remains identical, but with

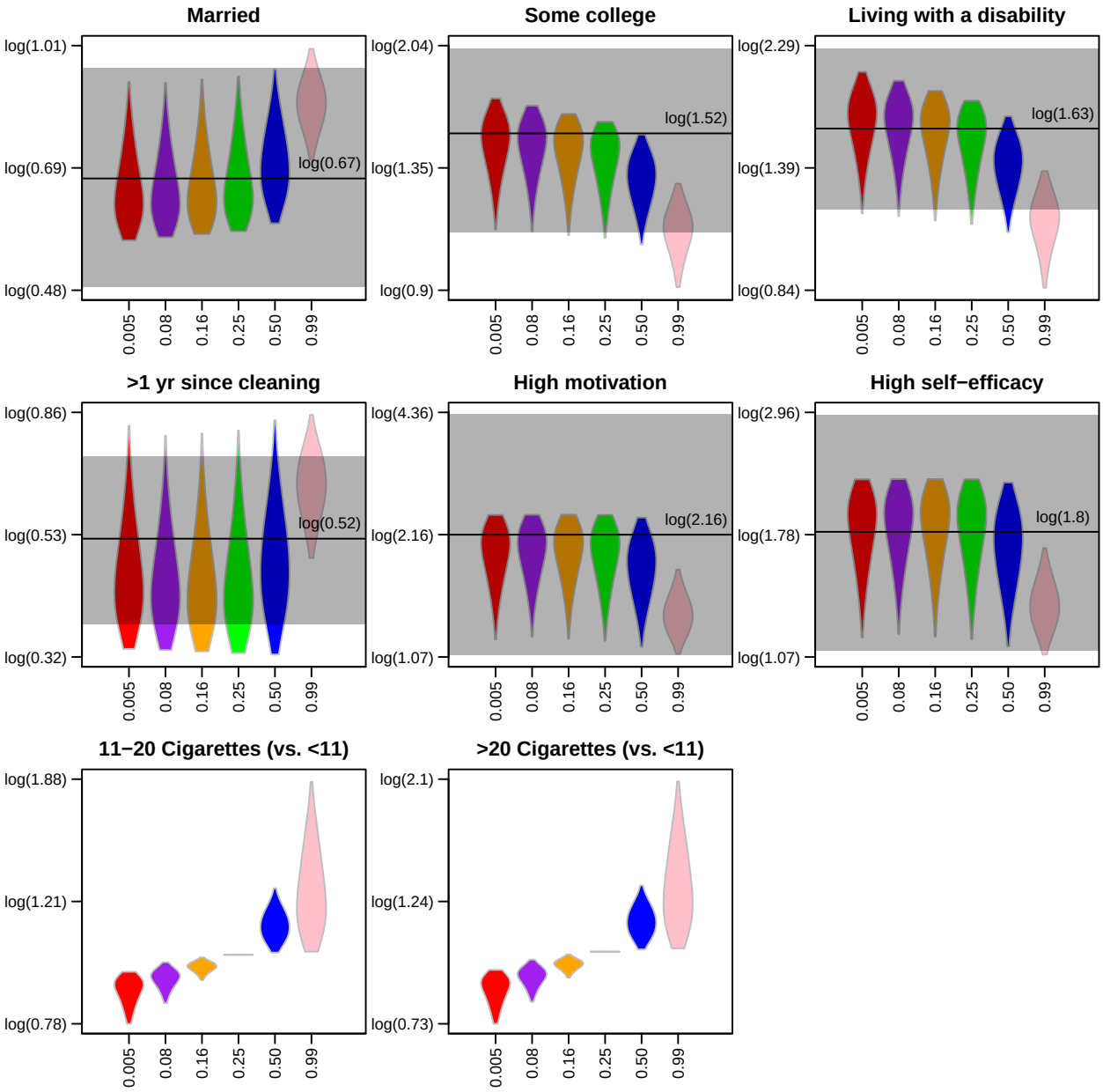


Figure 1.6: For values of μ_0 on the horizontal axis, smoothed frequencies of estimates of relative risks for the entitled covariate (compared to the absence of the covariate in the title, unless otherwise noted) for each possible combination of fixed remaining covariates in the final regression model; horizontal lines are drawn for point estimates of relative risks due to Blasi et al. [2018] and labeled in the plotting area; grey bands display 95% confidence intervals due to Blasi et al. [2018]. Note daily cigarette use was not included in the regression model fit by Blasi et al. [2018].

point estimates of relative risk changing quite drastically. For further interpretation, we recommend calculating confidence intervals for interpretable quantities using the novel approach, e.g. relative risks, rather than the parameter itself.

While determination of significance for such a coefficient is often used for model building, like coefficients in a generalized linear model, it must be interpreted alongside point estimates of changes in the outcome and related confidence intervals. Without directly interpretable coefficients in the final regression model, we may use a method such as the one illustrated in Figures 1.4-1.6 to compare point estimates calculated using the novel approach to point estimates and confidence intervals from the original analysis. Future research will investigate recommendations for comparisons of *confidence intervals* and *point estimates* from both the novel approach and the original analysis.

1.8 Final Remarks and Future Research

When estimating $\mathbb{E}[Y \mid x]$, if outcome Y is subject to nonignorable missingness, joint behavior of the response indicator and outcome must be accounted for to avoid bias. The method introduced herein uses the selection model to relate the dependence of unobserved outcome data on baseline covariates to the dependence of observed outcome data on baseline covariates. The method is applicable to a flexible class of selection models, for which (1) functional dependence on baseline covariates need not be specified and (2) functional dependence on the outcome is limited only by availability of estimates of marginal means of functions of the outcome in the target population. Thanks to Sadinle and Reiter [2019], with the aforementioned availability of auxiliary marginal information, identifiability of parameters indexing the final regression function is assured. The generalized method of moments may be used to acquire consistent and asymptotically normal estimates of parameters indexing the final regression model. Simulation studies conducted thus far show approximately nominal coverage is maintained even with small sample sizes and nearly half of participants lost to follow-up.

In practice, if an investigator intends to use this methodology to conduct a sensitivity analysis, we suggest considering how the entire moment vector in Section 1.5.2 will be

defined in the sensitivity analysis. The primary analysis may be conducted by fixing parameters indexing the selection model. For instance, fixing $\boldsymbol{\beta}$ at 0 reflects the assumption of ignorable missingness. Subsequent sensitivity analyses may then be conducted allowing $\boldsymbol{\beta}$ to be estimated using data and auxiliary marginal information. Differences in inference in the sensitivity analysis will then be due purely to differences in estimated auxiliary marginal information—implying extents to which missingness is not ignorable—rather than other modeling choices, such as how observed outcomes are modeled. Our method introduces some (albeit few) restrictions on the latter.

We have summarized an estimation method that depends only on correct specification of observable and unobservable regression functions to guarantee consistent and asymptotically normal estimates of parameters indexing the final regression function. In fact, complete specification of the verifiable density is only necessary to guarantee correct specification of the unobservable regression function. Finding a way to specify the latter without full specification of the former is an interesting aim for future research. Compiling a list of combinations of covariate-dependent response probability models, verifiable densities, and functions $h(Y)$ yielding interpretable coefficients in the final regression function is also of interest; otherwise, illustrations exhibited in this chapter may be used.

We have also made substantial progress on an entirely nonparametric estimator depending on kernel density estimates and Nadaraya-Watson estimators [Hastie et al., 2009, §6], for which there is no need to specify the verifiable density nor the covariate-dependent response probability. We leave finalizing this estimator and assessing its asymptotic properties to future research.

Thus far, we have focused on the existence of consistent and asymptotically normal estimators, not optimal efficiency. Assuming the verifiable density falls in the class of GLMs with canonical link, estimates of parameters indexing the density are recovered at the maximum likelihood estimator, and hence achieve optimal efficiency for any given fixed set of covariates. Similar logic dictates optimal efficiency of estimation of parameters indexing the covariate-dependent response probability. However, this question requires further research

given our assumption baseline covariates are *random*. Furthermore, a lingering question is how to define the third component of the moment vector in Section 1.5.2 to achieve optimal efficiency when estimating parameters $\boldsymbol{\beta}$ indexing the selection model. Without another candidate to replace the third component of the moment vector, we do not consider this question entirely pertinent, but interesting to address in future research. Rotnitzky and Robins [1997] define the *semiparametric* variance bound, to which our variance may be compared; however, we anticipate estimation is more efficient with fully-parametric specification of all models.

Chapter 2

DECISION-THEORETIC APPROACH TO *POST HOC* SIGNIFICANCE TEST ASSESSMENT

Abstract

Statistical tests—comparing a p-value to a prespecified threshold—are ubiquitous. While their use is controversial, in some areas, the binary decisions they produce are appealingly simple and reflect how knowledge will be acquired (e.g. economic policy changes, drug regulation). However, assessing a test's reliability *post hoc* is notoriously difficult; existing methods are prone to paradoxical results, misinterpretation, dependence on further arbitrary thresholds, and/or have failed to gain popularity. As an alternative, we consider a decision-theoretic approach to testing that produces natural analogs of one- and two-sided significance tests. In particular, we apply "loss estimation" (as well as risk estimation) to these decisions, deriving novel measures of test reliability. This approach provides a straightforward way to describe the merit of a significance test. The feasibility, accuracy, precision, and usefulness of these methods are explored.

2.1 Introduction

Testing is a ubiquitous part of statistical analysis, allowing inference on whether or not a parameter θ lies in some null space, determined by a null hypothesis. The implementation of frequentist tests is deceptively simple; one considers the p-value – the probability that some test statistic calculated from the available data is, under the null hypothesis, at least as extreme as was observed. If this statistic is extreme, meaning that the p-value is below some pre-specified threshold, the null is rejected. If not, the null is either not rejected (in a significance test) or accepted (in a hypothesis test). The binary nature of testing, yielding

a simple "yes" or "no"/"I don't know" response, is very tempting. Not only is it appealingly simple — as only two outcomes need be considered — but for decisions such as making an economic policy change, allowing a new medication to be sold, or changing the way patients are treated, it can be viewed as a directly-relevant distillation of the data.

Unfortunately, p-values are also highly prone to mis-interpretation and misuse, as summarized in recent ASA discussions and policy declarations [Wasserstein et al., 2016, Greenland et al., 2016]. It is particularly concerning that most of these faults result from over-interpretation of p-values — which were never created for definitive conclusions in the first place. They were instead to be considered as only part of a broader scientific context. — in which the p-values assess whether the question being addressed is "worthy of a second look" [Nuzzo, 2014], with other evidence and analyses brought to bear before firmer conclusions are drawn.

This weaker-but-correct interpretation of p-values is not obvious from standard teaching material nor common practice in the literature. In the latter, statistical significance is treated as a simple arbiter of truth, without accompanying measures of the test's reliability or evidential worth. Yet some such measures do exist, and may usefully supplement p-values and tests. We note that reporting confidence intervals may appear sufficient to make up for p-values' shortcomings—and are often cited as a solution—but they are prone to similar problems, such as dichotomizing of a continuous measure and widespread misinterpretation [Greenland et al., 2016].

In this paper we consider measures to accompany significance/hypothesis tests, summarizing the literature but also introducing some new approaches. The primary aim of this work is to help users better interpret p-values when the use of a p-value is appropriate to conduct a test. Our focus is measures that quantify how much faith to place in a test's binary result, rather than estimated magnitude of the parameter being tested; however, the former is sometimes quantified using the latter. While the alternative testing criteria uses a decision-theoretic analog to p-values, and hence might yield different inference to standard frequentist tests, we describe when tests using the new and standard frequentist criteria will

be approximately equivalent. Therefore, the new testing reliability criteria is applicable to frequentist tests.

Most previously-developed measures to accompany tests based on p-values are written assuming interpretation of the original results as a significance test, where non-significance leads only to failure to reject the null; however, extensions to hypothesis testing, where non-significance leads to a conclusion that the null holds, are possible if a hypothesis test is appropriate, i.e. the null hypotheses is at least plausibly true. While assessing reliability of results from hypothesis tests is still warranted, inconclusive results from significance tests with non-significant results (due either to insufficient evidence of the true alternative or the null being true) create particular need for significance tests to be augmented by measures of test reliability. Together with the widespread use of significance testing (or something close to it) in applied work, we are motivated to focus primarily on significance tests in this chapter.

This text is structured as follows. In Section 2.2, we describe existing methods of quantifying reliability of significance tests: *post hoc* power calculations (Section 2.2.1), severity (Section 2.2.2), fiducial distributions and confidence intervals (Section 2.2.3), errors of types S and M (Section 2.2.4), post-experimental rejection odds (Section 2.2.5), and analysis of credibility (Section 2.2.6). We then turn to decision-theoretic measures, building on recent results that motivate close analogs of one- and two-sided frequentist significance tests as Bayes rules for particular loss functions [Rice et al., 2020]. By estimating the loss and risk that might accompany this form of rule, we further quantify the information encoded in a test, and how it changes under different circumstances.

2.2 Existing methods

While a few methods do exist for leveraging data to help interpret results of frequentist significance tests beyond their reported p-values, few have gained any popularity and the motivation of almost all has been criticized. In our review of these methods below, we note that while the interpretation of non-significant results from significance tests strongly

motivate *post hoc* test assessment to uncover the reason for these results, the same assessment may be warranted for hypothesis tests, e.g. even if power is maximized for each alternative with a fixed Type I error rate, the power of the test at a reasonable alternative may have still been quite low for a given test. For brevity, we assume the original test about parameter θ was conducted using a significance test.

2.2.1 Post Hoc Power Calculations

Faced with the results of a significance test, there is never certainty about the underlying truth. A significant result may correctly indicate presence of a true effect (perhaps in some specified direction) or may be due to a false positive—where the null holds but the data are significant just by chance. Conversely, non-significant results may indicate that the null holds, or could be a false negative, again due just to chance. Distinguishing between the competing explanations for either significant or non-significant results is alluring: strong support for one or other of them greatly simplifies interpretation of the data. A particularly compelling way to distinguish between competing explanations is attempting to assess the power of the test—the probability that a significant result will be found for a given true value of the parameter being tested. So-called "*post hoc* power calculations" are problematic with both non-significant results, as discussed by Hoenig and Heisey [2001], and significant results, as discussed by both Hoenig and Heisey [2001] and Mayo [2018].

Hoenig and Heisey [2001] first focus on the problem of "observed power", defined as the value of the study's power curve at its observed point estimate $\hat{\theta}$ of parameter θ that was also used to calculate the test statistic (and p-value). In particular, prior to Hoenig and Heisey [2001], observed power calculations were frequently used to claim that non-significant results indicated that a signal was present but the study at hand was simply underpowered to find it. Therefore, with a point estimate in the alternative state space from a non-significant result, higher observed power provides evidence the null would have been rejected if false, therefore providing evidence the null is in fact true. The authors contradict this logic, proving that, for classic Normal location problems using Z-tests, observed power is a one-to-one function

of the p-value (regardless of sample size), and hence provides absolutely no information beyond the p-value with which to distinguish between competing explanations for the test result. Notably, this phenomenon occurs for both significant and non-significant results, so observed power does not contribute further evidence for either type of result.

Clearly, this result was not intuitive to the many authors who used observed power prior to Hoenig and Heisey [2001]. The lack of further information is obvious, however, if one considers simpler data, e.g. as a single Bernoulli variable. The single "bit" of information is expended when significance/non-significance is indicated, and nothing more can be inferred. However, for continuously-valued p-values, the result is less intuitive. While the one-to-one relationship between p-values and observed power is claimed universal by Hoenig and Heisey [2001]—a sentiment reiterated by numerous authors [O'Keefe, 2007, Lenth, 2001, Plate et al., 2019]—it does not appear to hold, exactly, when the measure of precision defining the test statistic depends on the null hypothesis, e.g. a score test for disease prevalence exceeding a given value. For further exploration of this result for some commonly-used statistical tests, see Appendix B.3.

Despite the results of Hoenig and Heisey [2001], some individuals continue to advocate for calculation of observed power, particularly with non-significant results. Some claim that the estimate of power is still useful for interpretation, framing the p-value on a different but helpful scale [Bababekov and Chang, 2019]. This interpretation may be deemed particularly important in fields in which non-significant results lead to adoption of new actions (e.g. a new surgical technique deemed equivalent to previous options) and/or conducting studies with enough events and/or sample size to fulfill a power threshold at an effect size of scientific interest is nearly impossible. However, Gelman [2019b] points out that observed power may be misleading; supposing the effect size takes a value in the calculated confidence interval, estimated power may vary wildly. Therefore, investigators calculating observed power may be unduly confident in their estimates of power and hence may misinterpret results.

Hoenig and Heisey [2001] provide further insight into how, with non-significant results, the use of observed power calculations can be actively misleading. Their focus is the "Power

Approach Paradox" (PAP), which is simplest to consider for one-sided significance tests, supposing $H_0 : \theta \leq \theta_0$ and $H_A : \theta > \theta_0$ without loss of generality. Panel (a) of Figure 2.1 displays the power curve for a normal location problem with $n = 3$, $\sigma^2 = 1$, and $\alpha = 0.05$. Suppose point estimate $\hat{\theta}$ varies across separate tests, keeping all other conditions (e.g. sample size, data variability and hence standard error) constant, so that the power curve in panel (a) of Figure 2.1 still applies. The power curve evaluated at any $\hat{\theta}$ is the "observed power". Consider the observed power calculations for $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ respectively, all of which yield non-significant results. Two contradictory trends occur simultaneously; first, observed power gets higher, suggesting increased support for the null; second, the p-value decreases, suggesting increased support for the alternative. (This issue persists with two-sided tests, with $\hat{\theta}$ getting larger in magnitude while staying of the same sign relative to null value θ_0.) The conclusion is that, with a non-significant result, interpreting higher observed power as greater support for the null hypothesis is fallacious.

This paradox also undermines other forms of *post hoc* power calculation. One of these is the "detectable effect size", the minimum value of $|\theta - \theta_0|$ at which a power threshold is surpassed. For example, continuing to use the one-sided Normal location problem example, to surpass power τ, θ must be greater than D_n for given sample size n in panel (b) of Figure 2.1; a two-sided analog requires θ to have at least some discrepancy from θ_0 in either direction. Under this *post hoc* argument, if an experiment fails to reject the null hypothesis, the detectable effect size may be interpreted as an upper bound on the value of θ, since if the true θ were at the detectable effect size or more extreme, it is very probable the null would have been rejected—and it was not.

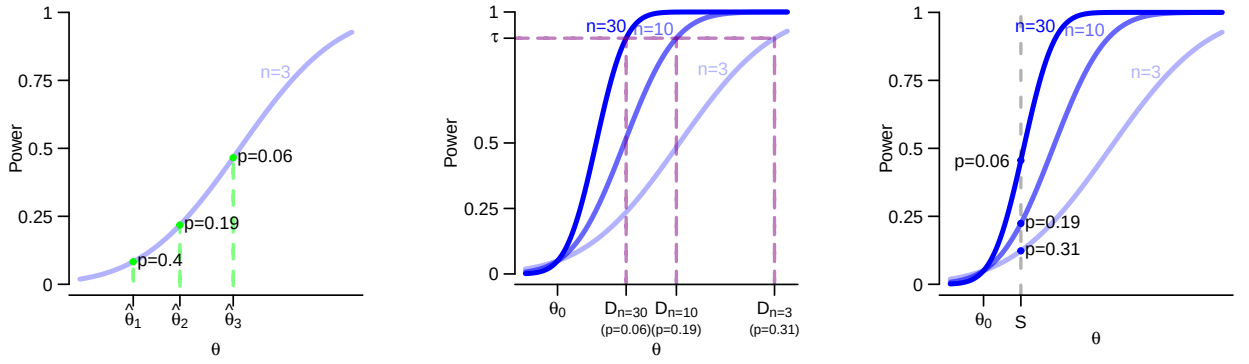
To see the fallacy, suppose a series of tests are conducted with decreasing standard error, accomplished by increasing the sample size maintaining the same data variance, while the point estimate θ remains the same, as illustrated in panel (b) of Figure 2.1. In this case, we suppose $\hat{\theta} = 0.28$ and data variability $\sigma^2 = 1$ in all experiments, while the sample size increases (from n=3 to n=10 to n=30). The resulting p-value from each test is displayed below the detectable effect size D_n for that sample size. As the sample size grows, the detectable

effect size gets closer to θ_0, providing more evidence in favor of the null. Simultaneously, the p-value decreases, providing stronger evidence against the null. The PAP is again evident, and the *post hoc* interpretation of the detectable effect size is fallacious.

The final example of *post hoc* power calculations discussed is power at an effect size of scientific interest, e.g. S in panel (c) of Figure 2.1 supposing the same one-sided Normal location problem as above. If the power calculated at the effect size of scientific interest is high, the probability of rejecting the null if the effect were as extreme is high; therefore, failure to reject the null provides evidence the true effect is not as extreme as the effect size of scientific interest. [Hoenig and Heisey, 2001] claim the PAP prevails in this instance as well. Consider a series of experiments with decreasing data variability (as may be accomplished by increasing the sample size and maintaining the same data variance), all else being identical. For example, consider the power curves for the series of tests in panel (c) of Figure 2.1, for which $\hat{\theta} = 0.28$ and data variability $\sigma^2 = 1$ across all experiments, while the sample size increases (from $n = 3$ to 10 to 30); p-values from those tests are printed by power curves for that sample size, all of which are non-significant. As the sample size increases, the calculated power at the effect size of scientific interest grows, providing increasing evidence for truth of the null. Simultaneously, the decrease in p-value provides more evidence against the null, leading the authors to conclude the PAP affects this *post hoc* power calculation as well.

Short of developing other methods, there seems no resolution of the problems the PAP raises with *post hoc* power calculations. Several proposed improvements are available but, as Hoenig and Heisey [2001] note, none avoid the PAP problem. However, the PAP does depend on the logic that if it is improbable θ exceeds some value, the only alternative is that θ takes a value in the null space (as opposed to values in the alternative state space near the null). This logic is reasonable with point nulls, but may be problematic with the ubiquitous one-sided testing example (see reasoning above).

As pointed out by Mayo [2018, pg 144 & 240], fallacies with common interpretations of post-hoc power calculations are not limited to non-significant results, but extend at least to just-significant results. Logic similar to that described above dictates interpretations of just-



(a) Power evaluated at estimates $\hat{\theta}$, a.k.a. observed power.

(b) Detectable effect sizes D_n with power threshold τ and p-values if $\hat{\theta} = 0.28$.

(c) Power at effect size of scientific interest S and p-values if $\hat{\theta} = 0.28$.

Figure 2.1: Illustrations of the PAP with observed power, the detectable effect size, and the effect size of scientific interest with one-sided normal location problem testing $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$. Data variance is assumed known at $\sigma^2 = 1$ and $\alpha = 0.05$. Values D_n denote the minimal detectable effect sizes with power threshold $\tau = 0.90$ for sample size n, and S denotes the effect size of scientific interest.

significant results that may be stated two ways for a one-sided test. Specifically, suppose that two tests are conducted yielding the same just-significant p-value, one with larger sample size (i.e. higher power) than the other. Results of the test on the smaller data set are indicative of (1) larger discrepancies from the null [Mayo, 2018, pg 240] and (2) provide more evidence in favor of θ exceeding any fixed discrepancy from the null in the alternative state space [Mayo and Morey, 2017]. Fear of "over-powered" tests is a consequence of these results, i.e. tests of sufficient size to successfully detect exceptionally small departures from the null of no scientific value.

Bayesian posterior distributions with flat priors (which are occasionally equivalent to fiducial arguments) may be used to prove the same conclusions as Mayo on relative evidence of discrepancies from the null from tests of different sizes (see Section 2.2.3 for more information on fiducial distributions), although extensions to more informative priors are unclear. Figure 2.2 displays, for a one-sided Normal location testing problem with $H_0 : \theta \leq 0$ and

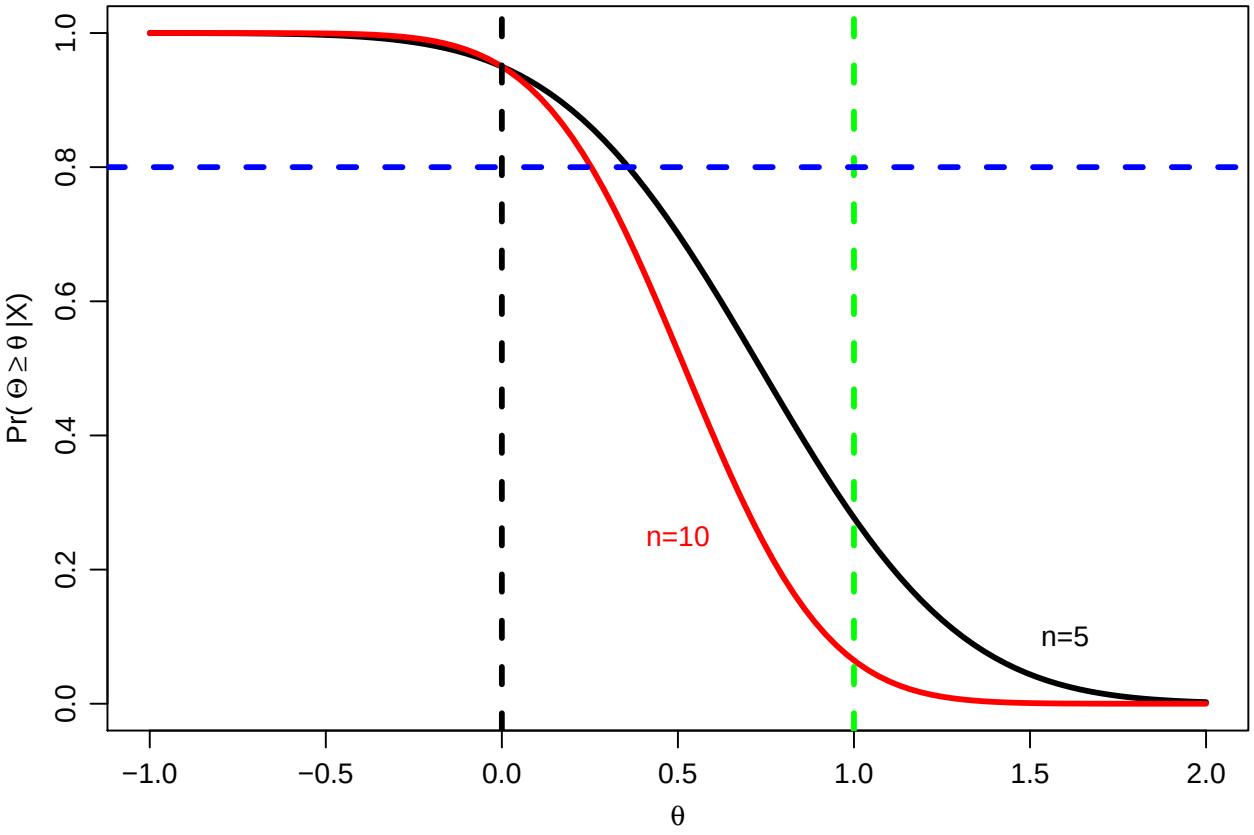


Figure 2.2: Fiducial probability θ exceeds value on the horizontal axis given data with a just-significant result with $\alpha = 0.05$ for a one-sided normal location problem testing $H_0 : \theta \leq 0$ with $n = 10$ and $n = 5$.

just-significant results, the fiducial probabilities that Θ exceeds the value on the horizontal axis. It shows results from samples of sizes 5 and 10, the latter having higher power to reject the null for each $\theta > 0$. For any fixed discrepancy from the null (e.g. $\theta = 1$, denoted by the vertical green dashed line), the probability the discrepancy exceeds that value is higher in the lower-powered study. Additionally, for any probability threshold required to claim $Pr(\Theta \geq \theta)$ (e.g. 0.80, denoted by the horizontal blue dashed line), the threshold is met for a smaller value of θ with the larger study, i.e. in the smaller study, we may claim smaller discrepancies from the null are unlikely (by looking at the difference between 1 and the curve).

Mayo claims interpretation of results is often the opposite of those given in (1) [Mayo, 2018, pg 240] and (2) [Mayo and Morey, 2017] respectively, i.e. many claim the larger study provides more evidence of a larger discrepancy from the null and more evidence θ exceeds any given discrepancy than the smaller study [Mayo, 2018, pg 240]. This fallacy is coined "making mountains out of molehills", and could be considered another example of the PAP. However, the true prevalence of this fallacy in current literature is difficult to gauge, and is perhaps lower than claimed by Mayo (as echoed by Owen [2019]).

Clearly, the p-value itself is insufficient to interpret either non-significant or just-significant results. While problematic with non-significant results, *post hoc* power calculations may however aid in interpretation of just-significant results, though it provides little or no extra information beyond the p-value.

2.2.2 Severity

With the same focus on magnitude of effect supported by the data in order to interpret the original binary test result, Mayo extends the logic used for *post hoc* power calculations with just-significant results to any p-value via a tool called *severity*. The motivation of *severity* [Mayo and Spanos, 2006] is to assess tests' abilities to find fault in false hypotheses and to provide "a way to use a test's error probabilities to assess the well-testedness of various statistical claims by given data" [Mayo and Morey, 2017], where "error probabilities" measure how often a frequentist test errs [Spanos, 2013].

Severity, a function of a hypothetical true value of parameter θ—which we denote by θ' in this section—requires two steps for evaluation after conducting the original significance test: (1) identifying the generic form of the hypothesis (defined using θ') the data "agrees" with—e.g. for a one-sided test, do data accord with $\theta < \theta'$ or $\theta \geq \theta'$ for an arbitrary θ'—and (2) for a specific θ', finding the probability with which, if the aforementioned hypothesis were false, future data would accord less well with that hypothesis than present data. Hence, the detectable "fault" in a false hypothesis is not defined by extremity of a test statistic relative to a critical value, but rather extremity relative to the already observed test statistic. If

the probability in (2) is high, the test (of the hypothesis defined in (1), not the original inference) passes with "high severity" [Mayo and Spanos, 2006]. We will describe each step of the evaluation in detail below, first for generic tests, and then more specifically for Z-tests. The section is complete with general findings on severity calculations, both from our own investigation and existing reviews.

The statement is step (1) due to Mayo and Spanos [2006, pg 329] may be misleading, suggesting the direction of the hypothesis is defined by accordance of the data to a specific θ'. However, a *generic* form of the hypothesis is defined. In actuality, finding a generic form of hypothesis the data accord best with for *any* arbitrary θ' is not possible. In fact, the final output of severity is calculation of accordance of data with the generic hypothesis defined for a range of θ'. In practice, it is more accurate to state the first step of the severity evaluation entails defining a generic form of hypothesis of particular interest given the result of the original inference. We have surmised this intent using examples, e.g. the one provided by Mayo [2018, pg 143].

The first step is best-illustrated using the often-cited case of a one-sided test $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$. If the null hypothesis cannot be rejected, the data are said to "agree" better with $\theta < \theta_0 + \gamma$ for some $\gamma > 0$ (as opposed to $\theta \geq \theta_0 + \gamma$). Defining $\theta' = \theta_0 + \gamma$, the generic form of the hypothesis in step (1) is now $H_0 : \theta < \theta'$. If the same null hypothesis is rejected, the data are said to "agree" better with $\theta \geq \theta_0 + \gamma$ for some $\gamma > 0$, thereby flipping the direction of the generic hypothesis in step (1). These generic hypothesis forms are motivated further below.

The probability defined in the second step is the severity of the test of the hypothesis defined in (1). Worse accordance of the data with the hypothesis defined in (1) is determined using the original test statistic $d(\cdot)$ and may be generally defined for test procedure T and realization x of data X as

$$\text{SEV}(T, X, \theta') = \min_{\theta \text{ w/ step (1) false}} \{\mathbb{P}[\, d(X) \text{ more extreme than } d(x)\,; \text{hypothesis in (1) is false}\,]\},$$

where the minimum is taken with respect to values of θ in the alternative state space of the hypothesis defined in step (1). Specification of the minimum calculation is newly made explicit here (but is generally called for by Mayo [2018]). The direction(s) of extremity is dictated by the hypothesis defined in (1), recalling we aim to find the probability of *worse* accordance of the data with a given hypothesis. In the notation provided by Mayo [2018, pg 143], the third argument of the severity function is a claim being tested about θ relative to θ'; notation has been altered here to emphasize calculation and to make the definition more general.

For general interpretation, note that with non-significant test results, if severity at a given θ' is high, the ability of the test to detect that difference was high, and therefore failure to reject the null is further indication of its absence [Mayo, 2018, pg 351]; this logic is not dissimilar to that used with detectable effect sizes (see Section 2.2.1). To further relate the two concepts, we note that severity will always exceed power calculated at a given alternative with non-significant results [Owen, 2019].

We now write out all the steps of the severity calculation for the one-sided Z-test (written as T^+) commonly used to illustrate severity. We continue to suppose the original test is an assessment of $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$. Suppose the original test result is non-significant. Recall the data are said to "agree" better with $\theta < \theta_0 + \gamma \equiv \theta'$ for some $\gamma > 0$. Hence, we will assess the severity of tests of hypotheses of the form $H_0 : \theta < \theta'$. For the moment, consider an arbitrary fixed θ'; in practice, severity is calculated across a range of θ'. Worse fit of the data with the hypothesis $H_0 : \theta < \theta'$ would occur if $d(X) > d(x)$, or the Z- score exceeded the observed one. In this case,

$$Pr(d(X) > d(x); \theta < \theta' \text{ false}) = Pr(d(X) > d(x); \theta \geq \theta') \geq Pr(d(X) > d(x); \theta = \theta'),$$

with the final probability being the value of severity at $\theta = \theta'$ with a non-significant initial inference. Using similar logic, significant initial inference yields severity equal to $Pr(d(X) < d(x); \theta = \theta')$, or one minus severity with non-significant initial inference, which quantifies the

severity of tests of the hypothesis $H_0 : \theta > \theta'$. For illustration, see Figure 2.3 for a variety of severity curves under significant/non-significant results and sample sizes for a one-sided Normal location test of $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$ evaluated using a Z-test.

In general with Z-tests, a ubiquitous example in the literature, severity is simply the p-value of the significance test of the hypothesis defined in step (1) at a given θ'. Furthermore, if the original one-sided Z-test yields a just-significant result, the resulting severity curve is $1 - \pi(\theta')$, where $\pi(\cdot)$ is the power function for the test. (Consequently, the blue and green curves in Figure 2.3 represent the tests' Type II error rates, i.e. one minus the power curves.) By extension, for any p-value, if we set the type I error rate threshold α to the p-value, the severity curve is the Type II error rate of the original test.

Despite the relationships between severity and power functions, Mayo [2018, p. 343] claims comparing power analyses and severity is akin to "compar[ing] apples and frogs" since severity calculations apply to both non-significant *and* significant test results and the claims to be tested are determined by the data (not the original hypotheses). The similarities in calculation cannot be denied with non-significant results; the primary difference between the two methods in this case is interpretation. While severity calculations assess the ability to find fault in a series of false hypotheses, in turn providing support for θ being within a certain interval of the null, *post hoc* power calculations then suppose small width of this interval provides more support for the null itself. In further contrast to power calculations, there does not appear to be consensus on default thresholds for severity, nor concrete guidelines on how to establish a threshold for a particular analysis.

As previously mentioned, severity translates to support for θ within a certain (one-sided) interval, and hence is primarily recommended for use defining bounds for θ; whether this bound is an upper or lower bound depends on the original inference, driving the generic form of the hypothesis in step (1) [Mayo, 2018, pg 351-352]. While a one-sided test is described below, findings may be extended to significant two-sided tests: the same principles of one-sided severity curves may be applied to the same side of θ_0 as the point estimate used to calculate the original test statistic [Spanos, 2013]. Non-significant two-sided testing analogs

are less clear.

In the case of a one-sided test, if the null hypothesis $H_0 : \theta \leq \theta_0$ cannot be rejected, we may suppose θ_0 is not an appropriate upper bound, motivating the search for a better one. Defining the generic form of the hypothesis to be $H_0 : \theta \leq \theta'$ as above, we wish to find the minimum value of $\theta' = \theta_0 + \gamma$, $\gamma > 0$ that passes a pre-specified severity threshold ζ, i.e. if the minimum value is θ'_{min}, then the test $H_0 : \theta \leq \theta'$ passes with high severity for any $\theta' \geq \theta'_{\text{min}}$. Therefore, θ'_{min} is a suitable upper bound for θ. For illustration, see the increase in upper bounds on θ dictated by the same severity threshold ζ with the same study design and different p-values with the red and purple curves in Figure 2.3. With results just shy of significance and severity threshold ζ approximating power thresholds, this bound will be very similar to the detectable effect size discussed by Hoenig and Heisey [2001]. Note $\theta'_{\text{min},n=10,p=0.94}$ is not positive; Mayo's interpretation of this value remains unclear Mayo [2018, pg 347]. To attain $\theta'_{\text{min}} > \theta_0$, the p-value from the original Z-test must be less than ζ; interest in severity with a p-value this large may arguably be minimal with such high compatibility of the data with the null hypothesis.

Another important note is that the upper bound θ_U of a standard one-sided confidence interval from a Z-test (i.e. $\theta \in (-\infty, \theta_U]$, as defined by Casella and Berger [2002, pg 425]) may be recovered by setting $\zeta = 1 - \alpha$, or phrased another way, $\alpha = 1 - \zeta$. Therefore, any differences in severity and one-sided confidence bound calculations amount to setting severity bounds ζ so that their difference from 1 differs from the Type I error rate used to conduct the original inference. One major practical difference between these two tools is calculation of severity for values of θ' beyond θ'_{min}, as opposed to dichotomous inclusion/exclusion criteria of confidence intervals Mayo and Spanos [2006].

Analogous interpretations with significant original results from one-sided tests may be formed, noting θ_0 is no longer a suitable lower bound for θ. Now, we may say $H_0 : \theta > \theta'$ passes with high severity for all $\theta' \leq \theta'_{\text{max}}$, where θ'_{max} is the maximum value of θ' with severity surpassing ζ. Therefore, θ'_{max} is a suitable lower bound for θ. For illustration, note how the lower bound increases for θ as the sample size decreases with just-significant results

in Figure 2.3 (also see discussion of "making mountains out of molehills" in Section 2.2.1). A lower bound θ_L for a classic one-sided confidence interval after conducting a Z-test (i.e. $\theta \in [\theta_L, \infty)$) may similarly be recovered by setting $\zeta = 1 - \alpha$, so that for any just-significant Z-test, $\theta'_{\max}$ will always be θ_0.

Mayo [2018, p. 351] also states that severity may be used to provide some indication of reasonable upper bounds on θ in the case of significant original inference and lower bounds in the case of non-significant original inference—so-called "warnings", in-keeping with the focus on one-sided bounds deemed appropriate for one-sided tests [Mayo, 2018, p. 357]. However, the method used to find these "warning" bounds is unclear; presumably, $1 - \text{SEV}(T, X, \theta')$ is calculated and values of θ' exceeding pre-specified thresholds found.

Despite advertised usefulness (as in the beginning of this section), severity does not appear particularly useful for addressing the original binary test result, a critique leveled by Robert [2014]. In addition, Robert [2014] discusses the need to set subjective and non-intuitive choices of severity threshold to find a bound for θ, a.k.a. "double calibration", therefore making severity "ineffective as an operational tool". Furthermore, discussion of severity to date has been limited to one-sided Z-tests (or turning a significant two-sided Z-test into a one-sided problem) with no nuisance parameters, leaving applicability to other scenarios in question [Robert, 2019]. Finally, due to the aforementioned relationship between one-sided confidence intervals and severity, Owen [2019] says there remains a need to motivate defining a bound for θ using a severity threshold that yields a bound that differs from a confidence interval, e.g. a value θ that should be considered plausible may pass with high severity but be excluded from a confidence interval, or vice versa; otherwise, the utility of severity remains unclear.

2.2.3 Generalized Fiducial Distributions and Confidence Intervals

With controversy surrounding the use of p-values, the practice of reporting confidence intervals in conjunction with or in place of p-values has long been promoted, and is a major component of the ASA's recent recommendations [Wasserstein et al., 2016]. Unlike p-values,

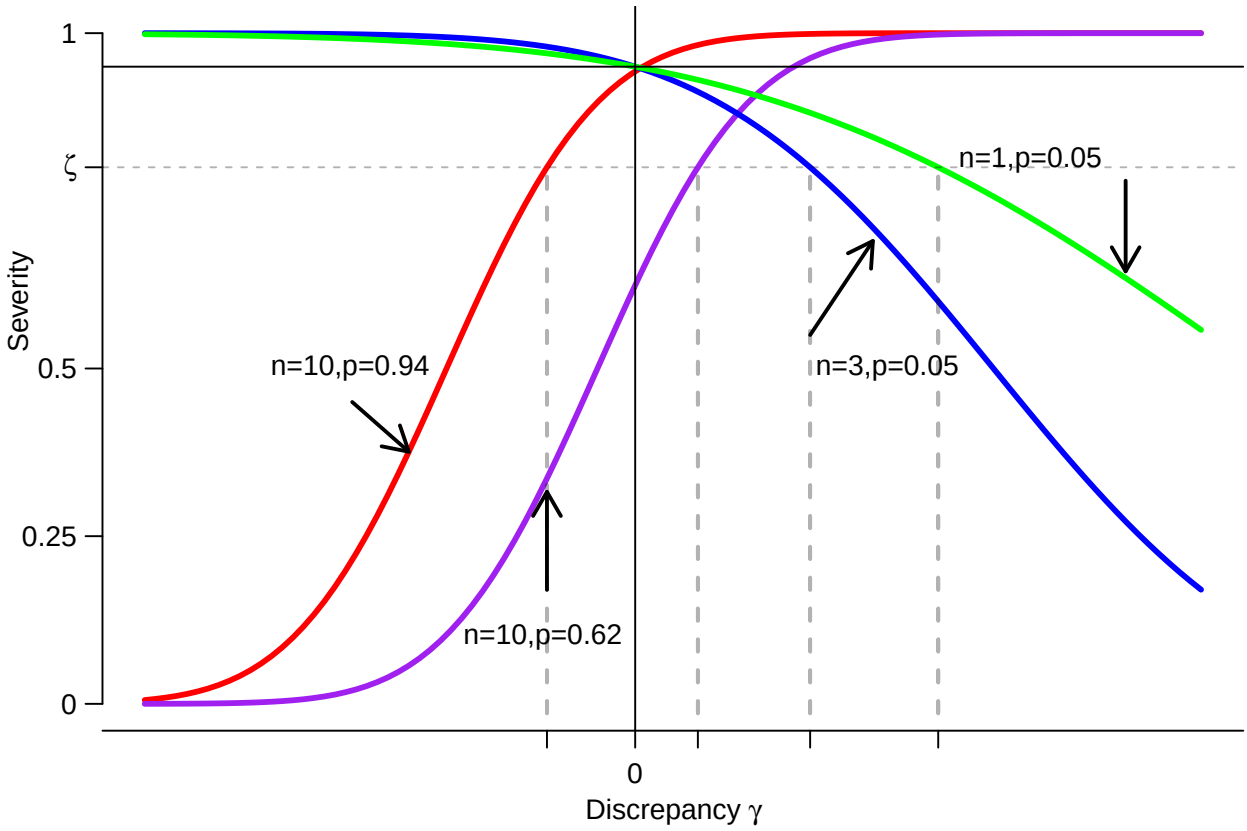


Figure 2.3: Severity curves by discrepancy γ of θ from $\theta_0 = 0$ for one-sided normal location test with data variance $\sigma^2 = 1$ and $\alpha = 0.05$ across labeled sample sizes n and p-values from the original inference. Threshold ζ is used to find tests that pass with high severity.

confidence intervals separate two components of the data: distance of the estimate $\hat{\theta}$ away from the null value, and precision of the estimate. In so doing, confidence intervals may quantify appropriate faith to place in the results of a significance test. In particular, confidence intervals provide clues regarding whether non-significant results are due to veracity of the null or lack of precision; with significant results, confidence intervals indicate how scientifically meaningful results are and may reveal probable false positives.

Unfortunately, like p-values, confidence intervals are prone to misinterpretations and have faced similar criticism. Furthermore, in practice, a strict focus on just the confidence bounds often leads to the same conclusions as p-values (via inclusion of the null value in the confidence interval), yielding a focus on binary decisions on hypotheses rather than a more holistic understanding of the relative evidence for different values of θ [Greenland et al., 2016].

Looking at confidence intervals as examples of generalized fiducial distributions, or probability statements about θ, may alleviate these issues, rendering confidence intervals more useful in practice and less prone to misinterpretation. Derivation and interpretation of such probability statements are summarized below.

The concept of fiducial distributions, or probability distributions of parameters without specification of a prior, was introduced by R.A. Fisher in the 1930s, reportedly beginning with an argument with Karl Pearson [Fisher, 1930, McGrayne, 2011]. Due to unclear distinction from Bayesian posteriors with uniform priors and lack of generalization to all data distributions (e.g. discrete data and/or multiple unknown parameters), the concept of fiducial distributions fell out of favor.

Lack of distinction from Bayesian inference is largely due to circumstantial equivalence of distributions (and therefore inference) using fiducial distributions and posterior distributions with a uniform prior for θ proven by Lindley [1958] that has been inappropriately generalized to every data distribution [McGrayne, 2011]. Hannig [2009] discredits this claim by showing that a generalized fiducial distribution (GFD) can rarely be written as a posterior distribution where any prior, uniform or otherwise, has been chosen for θ. Therefore, this process is

distinct from an update to a flat prior for θ.

Derivation of fiducial distributions starts with writing the data-generating mechanism in the form of a "structural equation" $X = G(U, \theta)$, where variable U has a known distribution and θ is an unknown parameter value. For example, if $X \sim N(\theta, 1)$, then we may write

$$\begin{aligned} X &= G(U, \theta) = \theta + U, \\ \text{where, } U &\sim N(0, 1). \end{aligned}$$

Where, for any observed x and arbitrary u, a unique inverse $Q_x(u)$ exists such that $Q_x(u) = \theta$, then we define the fiducial distribution for θ as the distribution of $Q_x(U^*)$ where U^* is an independent copy of U. (In the example where $X \sim N(\theta, 1)$, $Q_x(u) = x - u$, making $N(x, 1)$ the fiducial distribution for θ). In this way the fiducial approach obeys a "switching principle", switching the random X and fixed θ for a fiducial distribution that describes θ based on a now-fixed observation x.

However, a unique inverse $Q_x(u) = \theta$ may not exist—either because no solution exists or because multiple solutions exist. These limitations have recently been addressed using *Generalized Fiducial Distributions* (GFDs), due to Hannig [2009]. Where no solution exists, one may define the GFD to be the distribution of $Q_x(U^*)$ limited to the values of U^* for which a solution does exist (called U_x), requiring renormalization of the distribution of U^* across the limited domain (i.e. proportionate inflating to integrate to 1) to then find the distribution of $Q_x(U^*)$. However, the set U_x often has probability zero, making conditioning on such a set problematic. To combat these problems, Hannig et al. [2016] instead employ approximate inverses to "fatten up" the space U_x upon which the distribution is conditioned, defining the GFD as the distribution of

$$\lim_{\epsilon \to 0} \left(\underset{\theta^*}{\operatorname{argmin}} ||x - G(U^*, \theta^*)|| \, ; \, \min_{\theta^*} ||x - G(U^*, \theta^*)|| \leq \epsilon \right) \tag{2.1}$$

where the choice of norm is left to the discretion of the user.

Assuming approximate inverses are used to define the GFD, the issue of multiple solutions may persist; if so, random selection of solutions may be used to define the distribution, with suggested methods provided by Hannig et al. [2016].

For continuous data (generally meeting necessary criteria) with the same dimensionality as the unknown parameter space, the generalized fiducial density $r(\theta|x)$ may be written as in Theorem 1 due to Hannig et al. [2016]. Density $r(\theta|x)$ depends on the data density $f(x;\theta)$ and determinant of the gradient of the structural equation $G(U,\theta)$. If the dimension of the data does not match the dimension of the unknown parameter space, the function of the gradient of the structural equation used to define $r(\theta|x)$ depends on the norm used to define Equation 2.1 (see additional notes in Hannig et al. [2016]); since any norm may be used, $r(\theta|x)$ may not be unique.

GFDs have many strengths, including existence for any multivariate data distribution (whether continuous or discrete) with a data-generating mechanism depending on multiple parameters [Hannig, 2009, Hannig et al., 2016]. Unlike the original fiducial distribution, GFDs are always proper [Cox, 2006, Hannig, 2009]. Furthermore, while one-to-one transformations of the structural equations will often lead to different GFDs, they are invariant under smooth reparameterizations of the data distribution [Hannig et al., 2016, Hannig, 2013].

Deriving the GFD is a suitable initial step to calculating the confidence interval. Recall in this case, we have access to probability statements about the parameter itself; hence, intervals are analogous to *credible intervals* (as opposed to *confidence intervals*), and are called *confidence sets* here. For both continuous and discrete data, under certain conditions, appropriately-chosen confidence sets using GFDs provide asymptotically correct coverage [Hannig et al., 2016]. In many simple univariate cases, confidence sets derived in this way are in fact typical frequentist confidence intervals [Hannig et al., 2016].

Calculating bounds of confidence intervals as confidence sets poses two notable advantages to standard reporting of confidence intervals. First, the GFD permits the interpretation that the unknown parameter is more likely to be near the center of the confidence interval than near its end points — and second, if the true location is outside the confidence interval, it

is not likely to be far outside [Cox, 2006]. Developing confidence intervals as realizations of a process whereby the true θ is covered by the random interval in some percentage of replicate studies provides no basis for support of some values more than others. Additionally, as previously mentioned, visualizing the entire distribution of θ avoids the same binary decisions made using confidence intervals as typically made with p-values without further interpretation of strength of evidence. For all their strengths, GFDs are rarely available in closed form [Hannig, 2009], a potential impediment to their more widespread understanding. Furthermore, different quantiles for θ may be reported using the same data due to lack of uniqueness of the GFD in many circumstances; determining optimal choice of GFD with finite sample size is unclear.

2.2.4 Type S and Type M Errors

Errors of Types I and II (inappropriate rejection of null and inappropriate acceptance of a true signal, respectively) are the foundation of classical frequentist tests: Type I error rates are controlled at some preset level α while Type II error rates are minimized, by seeking optimal power for alternatives of interest. But these are not the only errors that can occur. Recently, Gelman and Carlin [2014] recommended considering two further errors in analysis of univariate parameters, primarily with two-sided tests. They define Type S errors as those where we obtain a significant result and reject the null, but where estimate $\hat{\theta}$ has the wrong sign. (This is a form of Mosteller [1948]'s "error of the third kind", correctly rejecting the null hypothesis for the wrong reason.) Gelman and Carlin [2014] also define Type M errors as the ratio of point estimate $\hat{\theta}$ to the true θ when a significant result is obtained.

Like Type I and II errors, in standard settings neither the existence/extent of Type S nor Type M errors can be determined from data. Instead, the probability of Type S errors (a.k.a. Type S error rate) and expected value of Type M errors (a.k.a. exaggeration ratio) should be calculated (with respect to repeated experiments) at plausible θ, ideally before a study is conducted to aid in study planning. However, *post hoc* calculation of these quantities may usefully quell excitement about significant findings, particularly surprising ones.

The calculation of these quantities is no more complex than standard power calculations, and Gelman and Carlin [2014] give explicit code for tests of Normal location. For low power settings, the size of both quantities may be alarming. For example, in the classical Normal location problem, with power of 10%, the Type S error rate is merely 4.5%, and rises quickly to 10% with a drop of power to only 7.5%. (The Type S error rate is always less that 0.5 and increases monotonically with decreasing power.) The exaggeration ratio is concerning at much higher values of power, reaching a value of 1.38 with 51.6% power, climbing to near expected doubling of point estimates from significant results (exaggeration ratio of 1.97) with power as high as 25.5%. It too is a monotonic function of power.

However, rather than focusing on a single value of θ, the authors suggest reporting Type S error rates and exaggeration ratios for a range of values of θ, i.e. not only at the point estimate $\hat{\theta}$. They recommend selecting values of θ deemed "plausible"—using external sources only, not the data at hand—for such calculations. This selection criteria is subtly different from choosing values of θ of interest, as typically used for power calculations, since these are prone to over-estimation by over-enthusiastic investigators. However, the inability to define plausible values of θ in some settings, e.g. innovative therapies, is an acknowledged weakness [Gelman, 2019a].

While clearly a useful tool for study planning, the informality of the system for selecting values of θ at which to evaluate Type S error rates and exaggeration ratios is a limitation. There is also no guidance on how, in any given situation, one might determine if the Tupe S error rate or exaggeration ratio is acceptable. Finally, the methodology is somewhat limited to two-sided tests: for one-sided tests, Type S errors are trivially either zero or one, and hence not useful. For two-sided tests, whether the exaggeration ratio is to be calculated only for significant results in the same direction as $\hat{\theta}$ or whether to average over all significant results remains a question to be addressed.

2.2.5 Post-experimental rejection odds

Unlike previously-discussed methods, *pre-experimental rejections odds* and *post-experimental rejection odds* [Bayarri et al., 2016] use prior information on θ to help interpret frequentist tests or provide new testing mechanisms themselves. However, for the two-sided tests for which they were developed, they explicitly require a prior with a point mass of support at the null value. This prior probability for the null is denoted τ_0, and the complementary support for the alternative is $\tau_1 = 1 - \tau_0$. The prior odds of the alternative are hence defined as

$$O_p = \frac{\tau_1}{\tau_0}$$

and the pre-experimental probability of rejection is

$$\alpha\tau_0 + (1 - \bar{\beta})\tau_1,$$

where α is the preset Type I error rate, and $1 - \bar{\beta} = \int(1 - \beta(\theta))\lambda(\theta)d\theta$ is the average power $\pi(\cdot) = 1 - \beta(\cdot)$ of correctly rejecting H_0 with respect to the prior for θ under the alternative $\lambda(\theta)$ for a two-sided test. This calculation may be completed for a single (or series of single) choices of alternative θ, for which specification of a prior distribution for θ under the alternative is unnecessary.

We see no reason the calculation cannot be generalized to composite nulls; although, we have yet to find this application in the literature. Continue to suppose the prior distribution of θ is $\lambda(\theta)$, which may be a continuous measure and does not require point mass. The prior probability of the null is redefined to be $\tau_0 = \int_{\theta\in\Theta_0} \lambda(\theta)d\theta$ for the null state space Θ_0, noting that $\tau_1 = 1 - \tau_0$ still holds. The Type I error rate α used in calculation of the pre-experimental probability of rejection must then be replaced with the average probability of rejecting the null under the null hypothesis or $1 - \bar{\beta}_0 = \int_{\theta\in\Theta_0}(1 - \beta(\theta))\lambda(\theta)d\theta$. Furthermore, bounds of integration for calculation of $1 - \bar{\beta}$ should be limited to $\theta \in \Theta_1$ for alternative state space Θ_1.

It is convenient to describe the relative size of the contributions from correct and incorrect rejections using the odds scale as

$$O_{\text{pre}} = \frac{(1-\bar{\beta})\tau_1}{\alpha\tau_0} = O_p\frac{(1-\bar{\beta})}{\alpha},$$

a.k.a. the pre-experimental rejection odds. We also define the *rejection ratio*, $R_{\text{pre}} = \frac{(1-\bar{\beta})}{\alpha}$, the term that depends on the test but not on τ_0.

Pre-experimental rejection odds are called "posterior odds for true association" by the Wellcome Trust Case Control Consortium [2007], a reconfiguration of prior odds, power, and Type I error used to calculate positive predictive value or "false-positive reporting probabilities" [Wacholder et al., 2004]. This quantity effectively predicts the odds of correctly rejecting the null (compared to incorrectly rejecting it) given prior understanding of the truth of the null and alternative. Hence, the pre-experimental rejection odds indicate how much the 'true signal' explanation of a significant result should be supported, should one occur—indicating the utility of the study, given a particular prior and commitment to using a particular test. However, this support is only given as an average over possible study results, and takes no account if, in the data at hand, a test gave a p-value marginally below α or many magnitudes lower. The authors therefore recommend that it be used only for study planning purposes.

To help interpret particular datasets, the authors instead prefer the post-experimental odds [Cox and Hinkley, 1979, pg 395], defined as

$$O_{\text{post}} = O_p \times \frac{\int_{\theta\neq\theta_0} f(x;\theta)\lambda(\theta)d\theta}{f(x;\theta_0)},$$

where $f(x;\theta)$ denotes the density of observations X indexed by parameter θ. Analogously to R_{pre}, we can define the post-experimental rejection ratio $R_{\text{post}}(x) = \frac{\int_{\theta\neq\theta_0} f(x;\theta)\lambda(\theta)d\theta}{f(x;\theta_0)}$ using observed data x; this ratio is also known as the Bayes factor (note the reciprocal of this quantity may be defined as the Bayes factor, as stated by Robert [2007, pg 231]). A more flexible definition of the Bayes factor allowing for composite null hypotheses is provided by

Robert [2007, pg 227]. This statistic may be easily manipulated to allow for nuisance parameters for which prior distributions have been defined, while frequentist statistics allowing for nuisance parameters require individual assessment of asymptotic behavior [Cox and Hinkley, 1979, pg 398].

Rewritten as $O_{\text{post}} = \frac{Pr(H_1|X)}{Pr(H_0|X)}$, we see that the post-experimental odds are the ratio of the posterior probability of the alternative to the posterior probability of the null, with no conditioning on significance. In short, given a particular dataset $X = x$, O_{post} indicates how much support should be given to the explanation that a true signal is present.

A notable simplification of the post-experimental rejection odds is available through a bound on the Bayes factor at the observed p-value. Specifically, under very general conditions,

$$R_{post} \leq \frac{1}{-ep \log p}.$$

This bound holds universally for two-sided tests, and approximately for one-sided tests [Selke, 2012]. Although formally restricted to $p \leq 1/e \approx 0.37$, the bound is useful as a device by which to tamper enthusiasm about significant results with $p < \alpha$ for practical α. When the bound is low (e.g. $R_{\text{post}} < 2.46$ for $p = 0.05$), obtaining just-significant results leads to little faith in *a priori* implausible results, noting the post-experimental reject odds are bounded above by $2.46\, O_p$ in this example. Hence, we may further interpret frequentist significance tests even if data are not available [Bayarri et al., 2016] or the Bayes factor cannot be computed analytically, as is often the case [Robert, 2007, pg 228].

For tests of $H_0 : \theta < \theta_0$ vs. $H_A : \theta \geq \theta_0$, $O_{\text{post}} = \frac{1-p_{\text{post}}}{p_{\text{post}}}$, where p_{post} is the posterior tail area, i.e. $p_{\text{post}} = \mathbb{P}[\,\theta < \theta_0 | X\,]$. With a limiting flat prior, the prior odds tends to a constant c and the post-experimental odds approximately equals $c \times \frac{1-p}{p}$, where p is the frequentist p-value. We see that, in this situation, post-experimental rejection odds do not contribute to interpretation of the results beyond re-stating the original p-value.

Thus far, we have focused on how post-experimental rejection odds may be used to better interpret p-values. However, they are useful to conduct inference themselves, i.e. significance

is reached if O_{post} exceeds a pre-specified threshold. For the remainder of this subsection, we consider use of the post-experimental rejection odds or Bayes factor themselves for inference (or as a validation of results reached using frequentist p-values). As we will show, the use of Bayes factors/post-experimental rejection odds introduces some unique challenges, and hence is not a "silver bullet" solution to the issues p-values face as an inferential tool. Typically, existing literature considers how the Bayes factor may be used for inference, although identical issues persist if the post-experimental rejection odds is used instead.

First, assignment of the prior distribution raises some challenges. With two-sided tests, the methodology requires that users choose a prior point mass at the point null, to which subsequent results are sensitive. This is a substantial limitation, albeit one shared by many other methods [Lucke, 2009]. For one-sided tests, this sensitivity is less limiting. Note this issue persists whether O_{post} is used for inference or to better interpret p-values.

When assigning a prior distribution under the alternative, investigators may be tempted to assign priors as uninformative as possible, maintaining the spirit of frequentist tests conducted. However, these priors may lead to two paradoxical problems. First, suppose an improper flat prior is assigned to θ under the alternative hypothesis, with prior mass given to the point null. Using a realistic example, Robert [2007, pg 233] displays how posterior probability in favor of the null hypothesis may have modest upper bounds, no matter the data observed, leading to frequent rejection for large enough sample size; similar problems may persist if flat priors are assigned to compact composite null and alternative hypotheses. Now suppose the prior distribution assigned to θ under the alternative is proper, with variance approaching infinity, with prior mass assigned to a point null hypothesis. Another realistic example provided by Robert [2007, pg 234] shows how, as a result of the Jeffreys-Lindley paradox, this setting may lead to posterior probabilities supporting the null hypothesis approaching one, no matter the data observed, leading Robert to conclude that assigning priors under the alternative with variance approaching infinity do not yield truly non-informative priors.

A possible resolution to these problems rests in comparison of score functions (in place

of likelihoods) under each hypothesis, emphasizing predictive capability of each hypothesis while incorporating losses from choosing to model the data using a generic alternative versus the point null hypothesis [Robert, 2014, Sprenger, 2013]. As an example, the Bayesian Reference Criterion (BRC)—the expected Kullback-Leibler divergence with respect to the posterior for θ between distributions for data under the null and alternative—does not require assignment of prior mass to point null hypotheses [Sprenger, 2013]. Unfortunately, choice of appropriate cut-offs for difference in scores to reject the null hypothesis are not provided in the theory [Robert, 2014] and may not be intuitive.

Like frequentist tests, Bayes factors/post-experimental rejection odds do not take detectable discrepancy into account when making testing decisions [Spanos, 2015]. Therefore, interpretation of non-significant results remains unclear, i.e. whether results were due to insufficient evidence/truth of the alternative or truth of the null [Wakefield, 2009, pg 159-160]. Furthermore, as the sample size increases, the null hypothesis may be rejected with high probability (depending on choice of the prior) unless it is exactly true, which will be implausible with point null hypotheses in many applications [Cox and Hinkley, 1979, pg 398]. Therefore, differences from the null of no scientific interest may be detected and prone to over-interpretation. Spanos [2013] argues that a severity analog in the Bayesian context does not exist, thereby discouraging use of the Bayes factor.

Finally, with acceptance of the null hypothesis, if the MLE $\hat{\theta}$ is in the alternative state space, the Bayes factor testing $H_0 : \theta = \hat{\theta}$ vs. $H_A : \theta \neq \hat{\theta}$ may provide convincing evidence of the truth of this new null, contradicting the initial result of the test [Spanos, 2013]; however, evidence in favor of θ taking the value of the maximum likelihood estimator, and therefore taking a value in the alternative state space, may not be contradictory to evidence in favor of the null hypothesis of particular interest to the researcher and due solely to randomness in the data [Robert, 2014].

2.2.6 Analysis of Credibility

Like pre- and post-experimental rejection odds, *Analysis of Credibility* (AnCred) [Matthews, 2018, 2019] uses Bayesian methods to assess well-testedness of hypotheses. The methodology was developed for two-sided tests with point null hypotheses, with unclear extension to one-sided tests. After conducting standard frequentist tests, AnCred introduces a reverse-Bayesian argument [Held, 2013], identifying how informative the assumed-Normal prior for θ must be for Bayesian inference to contradict the frequentist result. Bayesian inference is defined as including/excluding θ_0 in the $(1-\alpha)$ posterior credible interval. The *critical prior* that just achieves this is described through its central $1-\alpha$ region, known as the *Critical Prior Interval* (CPI). Issues with use of p-values in isolation from the broader scientific context [Nuzzo, 2014] are ameliorated by considering plausibility of values outside the CPI using prior studies and expert opinion. Applications of AnCred to *out of the blue* results, or findings for which no scientific context is available, are described later in this section.

For significant findings, the critical prior and its CPI are centered at $\theta = \theta_0$. Hence, this prior quantifies skepticism about an effect size, while acknowledging the possibility of more extreme effect sizes. The next step entails determining if there is prior belief in θ taking some value outside the CPI; while not explicitly stated, *a priori* support for effect sizes in the same direction as $\hat{\theta}$ are solely assessed in illustrative examples [Matthews, 2018]. If there is prior belief θ takes some value outside the CPI, the appropriate prior to assign should be flatter than the critical prior, changing Bayesian inference from non-significant to significant. Hence, the claim of statistical significance is deemed credible.

For non-significant frequentist findings, AnCred first requires the definition of a *substantive hypothesis*, or the one-sided hypothesis most relevant to the research question. For example, while we often use two-sided tests to assess differences in effects of treatment and placebo, we are generally interested in superiority of the treatment. (This definition has been devised from context and not explicitly stated by Matthews [2018]). The (Normal) critical prior is no longer centered at the null; instead, θ_0 is the $(\alpha/2 \times 100\%)^{\text{th}}$ or $((1-\alpha/2)\times 100\%)^{\text{th}}$

percentile of the prior, where the bulk of the prior mass supports the substantive hypothesis. The *advocacy limit* is on the same side of θ_0 as the substantive hypothesis and is moved until the prior probability the effect size is more extreme than the advocacy limit (on the same side of θ_0 as the substantive hypothesis) is $\alpha/2 \times 100\%$. Hence, the CPI is centered on the same side of θ_0 as the substantive hypothesis, with θ_0 serving as one bound.

Finally, as with significant results, if there is *a priori* support for θ outside the CPI, then claims of non-significance are deemed credible. Therefore, claims of non-significance may be challenged if the mode of the posterior distribution is in the same direction as the substantive hypothesis and effect sizes are deemed unlikely to exceed the AL; the mode in the opposite direction maintains credibility of the non-significant claim. Of special note, it is the claim of *non-significance* that is deemed credible, rather than veracity of the null. This conclusion is particularly appropriate with a point null.

Supposing the default Type I error rate of $\alpha = 0.05$, results that deserve particular attention are those with p-values in the "twilight zone", i.e. $0.05 < p < 0.10$ [Matthews, 2018]. These cases often lead to extreme ALs. In the case the AL is extreme, according to the formal description of methodology by Matthews [2018], the clear inability to claim plausibility of values beyond the AL suggest lack of credibility of the non-significant claim. However, in more informal discussion surrounding illustrative examples, the extreme value of the AL is not considered a "useful constraint on advocates of the substantive hypothesis", and as such, the credibility of the claim cannot be assessed, hence the moniker of the "twilight zone". Determining if the AL is too extreme to be helpful or extreme enough to question validity of the original non-significant inference is difficult to gauge.

For both significant and non-significant results, the amount of belief in effect sizes outside the CPI required to defend credibility of the initial result is unclear. Phrased differently, if the totality of existing scientific evidence was summarized in a distribution for θ, the amount of distinction required between that distribution and the critical prior (or CPI) required to claim credibility of the original inference is unclear; for now, allowable overlap appears to be at the user's discretion.

However, with "out of the blue" findings, no such summary of existing scientific evidence is available, and hence the process of AnCred is modified. The entirety of scientific information comes from the current data set, best summarized by the "most-probable effect size", i.e. the mode of the data-driven likelihood (the posterior distribution for the effect size with a noninformative prior). If this value lies within the CPI, then claims of significance are not *intrinsically credible*. Interestingly, p-values always yield intrinsically credible claims of significance if $p < 0.013$ for a 95% level test with normally-distributed parameters. While this finding may appear to justify lessening the p-value threshold for significance, Matthews [2018] is careful to note this threshold only indicates whether the current data set alone is sufficient to claim significance; significant results failing to meet this criteria may later be found credible through accumulation of additional evidence. However, implications for interpretation in practice remain unclear, i.e. would a non-intrinsically credible significant finding with a p-value of 0.02 be interpreted/treated differently had the result been declared non-significant?

AnCred provides useful comparison of results across studies with the same p-value via comparison of the bounds of the CPI. For instance, with significant results, it is easier to find evidence for effect sizes outside CPIs with bounds closer to the null from scientific context; therefore, larger studies with bounds closer to the null provide greater evidential weight, with more credible significance. Similar logic may be applied to non-significant results.

Matthews [2018] claims AnCred may also be used to assess discordance of significant and non-significant results, e.g. non-significant results from a large study designed to replicate the significant findings of a smaller exploratory initial study. If the mode of the data-driven likelihood for the effect size of the former is in the same direction as the latter, Matthews [2018] claims non-significance of the former is not deemed credible, and the culmination of data "point to the reality of a positive effect". Furthermore, if the 95% CI of the smaller study lies within the CPI of the larger study, non-significance is further contradicted, pointing further to concordance in conclusions. It should be noted justification for these conclusions depends in part on a meta-analysis using all data from both studies, a practice that warrants

caution [Fleming, 2010], especially if subsequent studies were planned only for significant initial results, thus inciting the "Winner's Curse" [Capen et al., 1971].

2.3 Decision-Theoretic Approach

A foundational difficulty with all the methods in Section 2.2 is the abstract nature of the metrics that determine how well a test performs. Severity, Type S error rates, exaggeration ratios, and *post hoc* power calculations are presented only as functions of *possibly*-true θ values. The latter three, with the exception of observed power, also average over behaviour in replicate datasets without any influence of the current data set (apart from significance) on the metric of reliability. Some measures, namely confidence intervals, severity, and multiple forms of *post hoc* power calculations depend on magnitudes of θ compatible with the data, which does not intuitively translate to reliability of the original inference. The Bayesian measures we have described require unrealistic point masses at point null hypotheses, or have been developed only for particular prior distributional assignments. A further difficulty is that when new metrics are introduced to assess tests, they must be calibrated in some way, and there is little guidance on how to decide, for example, what severity thresholds are appropriate, what Type S error rate can be tolerated, or how much support there must be for effect sizes outside the CPI in AnCred. Most of the existing methods we have reviewed provide point estimates of the reliability metric, failing to provide a range of suitable values or intuitive means to define such a range.

These problems can be resolved if we view testing as a decision-theoretic process, where the test result optimizes, for a given dataset and prior, the posterior expectation of some *loss function*, that ascribes costs to each combination of the decision made and the underlying truth.

This is the approach to testing laid out by Rice et al. [2020], who note that by viewing one- and two-sided significance tests as decisions about just the sign of θ, all loss functions

obeying certain regularity conditions are equivalent to the following:

$1-$ sided :

	$d =$ Above	$d =$ No decision
Loss when $\theta > \theta_0$	0	α
$\theta < \theta_0$	1	α
do d iff	$\mathbb{P}[\theta < \theta_0 \vert X] < \alpha$	$\mathbb{P}[\theta < \theta_0 \vert X] > \alpha$

(2.2)

$2-$ sided :

	$d =$ Above	$d =$ No decision	$d =$ Below
Loss when $\theta > \theta_0$	0	α	2
$\theta < \theta_0$	2	α	0
do d iff	$\mathbb{P}[\theta < \theta_0 \vert X] < \frac{\alpha}{2}$	Otherwise	$\mathbb{P}[\theta > \theta_0 \vert X] < \frac{\alpha}{2}$

(2.3)

A key condition here is that the loss for making no decision is identical for each value of θ. (In the two-sided case we have for simplicity focused on situations where either incorrect sign-declaration is equally bad, but this is straightforward to generalize.)

In both the one- and two-sided cases, the Bayes rules behave very much like classical frequentist tests: the decision to reject the null (or make no decision) rests on whether a Bayesian analog of the classical p-value is below some threshold α. Throughout the remainder of this chapter, we denote the function of the tail area used to make one- and two-sided decisions in Equations (2.2) and (2.3) respectively by $\tilde{p}$. For a one-sided test, $\tilde{p} = \mathbb{P}[\theta < \theta_0 \vert X]$ (or $\mathbb{P}[\theta > \theta_0 \vert X]$ if the direction of the null hypothesis is reversed); for a two-sided test, $\tilde{p} = 2\min\{\mathbb{P}[\theta < \theta_0 \vert X], \mathbb{P}[\theta > \theta_0 \vert X]\}$, a close analog to the procedure used to calculate the frequentist p-value. With large samples, by Bernstein-von Mises' theorem, not only does $\tilde{p}$ approximate the frequentist p-value, but α also determines the test's Type I error rate as the prior is dominated by the likelihood. However with small samples the Bayesian tail area that determines the test will—appropriately–depend more on the available prior information.

Our key observation in this section is that the losses from Equations (2.2) and (2.3) that motivate the Bayesian tests also provide metrics by which we can assess those tests. As we describe in Sections 2.3.1 below, the actual loss incurred for the decision taken is a quantity

about which we can express posterior uncertainty. The *risk*—the expectation of the loss over replicate experiments, summarizing how well the procedure might fare if repeated—is another quantity about which we can express posterior uncertainty [Parmigiani and Inoue, 2009, pg 121]. While this approach is, to our knowledge, novel in the significance-testing literature, the idea of adding a loss-estimation step to an existing decision problem has been studied for some years in decision theory literature [Robert, 2007, pp.178-180].

Throughout, the methodology, discussion and examples are provided for both one- and two-sided tests. Interpretation and usefulness of loss estimates for one- and two-sided tests are very similar. Similar trends impact risk estimates across one- and two-sided tests, with the exception of tails of the posterior distribution for risk, discussed in detail below.

2.3.1 Estimating loss of a test

Having established a loss function for a particular decision problem, estimation of that loss (for the data observed) provides an indication of the reliability of the results. Even with the same decision—a test, in our case—there may be important differences in loss incurred. This form of *loss estimation* [Robert, 2007] can itself be viewed as a decision problem, extending the decision space to estimate the loss conditional on the primary decision. For example, using η to estimate $L(\theta, d(x)) \equiv l_d$, the loss incurred in the primary decision, we may define quadratic loss $L(l_d, \eta) = (l_d - \eta)^2$ for the estimate of the primary loss incurred, which motivates the posterior mean $\mathbb{E}[\, L(\theta, d(x))|X = x\,]$ taken with respect to θ as a Bayes rule. We can extend the loss to

$$L(l_d, \eta, v) = \log(v) + \frac{(l_d - \eta)^2}{v},$$

where decision $v > 0$. This extended loss retains the posterior mean $\mathbb{E}[\, L(\theta, d(x))|X = x\,]$ as the Bayes rule for l_d, but additionally provides the posterior variance $\mathrm{Var}[\, L(\theta, d(x))|X = x\,]$ as an indicator of the reliability of the posterior mean—the term v express our desire to estimate the value of $(l_d - \eta)^2$, i.e. the squared error.

We consider sensible estimates of the loss—a.k.a. assessments—before constructing such new loss functions. For our testing rules, in which the loss for the primary testing problem can take only a few values, similar forms of assessment can be expected from many data realizations. In particular, when a null $d{=}N$ decision is returned, we know *with certainty* that loss functions (2.2) and (2.3) both return $L(\theta, d(x)) = \alpha$, so reporting any other assessment of the test would be inappropriate. When a non-null decision is returned, we know for one-sided tests the loss must be either zero (if a correct decision is made) or one (if a correct decision is made), and similarly for two-sided tests must be either zero or two. Hence for non-null decisions the posterior distribution for the loss can be stated as a single probability—say the probability that $\text{sign}(\theta)$ disagrees with the stated decision—and any loss estimation procedure which returns this probability (or a one-to-one function of it) is equivalent in practice.

This establishes that loss estimation will lead to reporting assessments of the test equivalent to the following;

$$1-\text{sided}: \quad \begin{array}{r|cc} \text{Test result:} & d=\text{Above} & d=\text{No Decision} \\ \text{Assessment reported:} & \mathbb{P}[\,\theta<\theta_0|X\,] & \alpha \end{array} \tag{2.4}$$

$$2-\text{sided}: \quad \begin{array}{r|ccc} \text{Test result:} & d=\text{Above} & d=\text{No Decision} & d=\text{Below} \\ \text{Assessment reported:} & 2\mathbb{P}[\,\theta<\theta_0|X\,] & \alpha & 2\mathbb{P}[\,\theta>\theta_0|X\,] \end{array} \tag{2.5}$$

We see that loss estimation of significance tests resolves to simply reporting the decision made in all cases, and when an active decision is made, additionally reporting $\tilde{p}$, the function of the posterior tail area used to make the decision. For non-null decisions, this accords with standard recommendation to report the actual p-value as well as whether $p < \alpha$ (Friedman et al., 2010, p.414-415). In this framework—where the sign of θ is the only quantity of interest—we have established $\tilde{p}$ fully captures the available information on how well we ascertain that quantity. Logically, therefore, assessments that go beyond reporting $\tilde{p}$ must address different quantities of interest that should be determined by the scientific context.

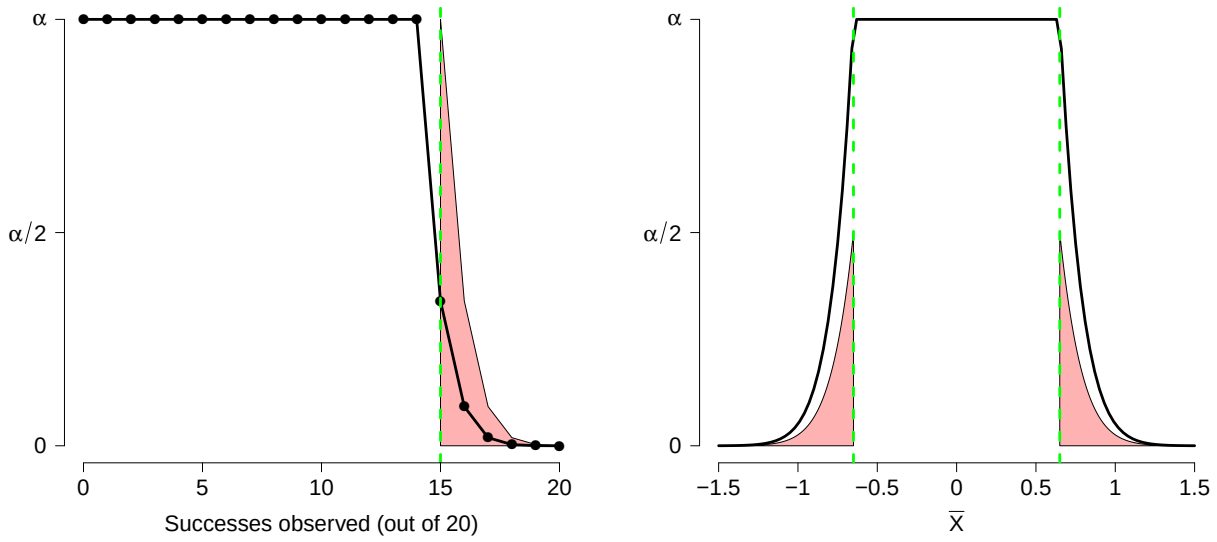
This result also establishes an important Bayesian analog of Hoenig and Heisey [2001]'s result on observed power calculations simply re-stating the p-value; for *post hoc* loss estimation of significance tests as we have developed them, in *any* setting with a univariate parameter, *any* assessment of the test's reliability in the situation at hand must state $\tilde{p}$ alone. This statement is more powerful in the Bayesian context: not only the estimate of loss a restatement of $\tilde{p}$, but due to the binary nature of the loss, any statement about the distribution of loss is a restatement of $\tilde{p}$ as well.

Figure 2.4 displays estimates of the loss and posterior support for positive loss for a one-sided test of binomial proportion θ with uniform prior in the left-hand panel (where $H_0 : \theta \leq 0.5$ vs. $H_A : \theta > 0.5$) and for a two-sided test of Normal mean θ with standard normal prior in the right-hand panel (where $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$). Estimates of loss are highest when no decision is made, which is simply the value α used to define the loss function, and are equal to $\tilde{p}$ otherwise. Posterior support for positive loss when a decision is made is equivalent to $\tilde{p}$ for one-sided tests and $\tilde{p}/2$ for two-sided tests. Therefore, not only is the original estimate a restatement of $\tilde{p}$, but any statement about the faith in that estimate is a restatement of $\tilde{p}$ as well.

2.3.2 Estimating risk of a test

As discussed in Section 2.2, several existing methods for evaluating tests rely not on their behavior with the data at hand, but on their properties averaged over an infinite series of replicate studies. These properties are therefore frequentist in nature. Unlike frequentist analyses, using Bayesian methods, we may express uncertainty about metrics that depend on true underlying values of θ. In addition, we may leverage current data to update our understanding of θ and how that might impact replicate data sets. In this section we explore what we call *risk estimation*, i.e. posterior inference for the expected values of losses (2.2) and (2.3), where the averaging is over replicate datasets, applying the Bayes rule to each of them.

This approach provides a Bayesian analog of frequentist *post hoc* power calculations—



(a) $H_0 : \theta \leq 0.5$ for success probability θ with uniform prior, $n = 20$.

(b) $H_0 : \theta = 0$ where $X \sim N(\theta, 1)$, $\theta \sim N(0, 1)$, $n = 10$.

Figure 2.4: Posterior mean (black) of loss and posterior support for positive loss when a decision is made (estimate more extreme than green dashed lines) in red. Note, loss is known be to α when no decision is made.

using the available information to infer how reliable the testing procedure might be in similar studies. This Bayesian paradigm has the ability to straightforwardly reflect all sources of uncertainty, due to both the prior's and data's impact on the posterior. More importantly, it is not subject to the same paradoxical conclusions and provides novel information beyond the original inference.

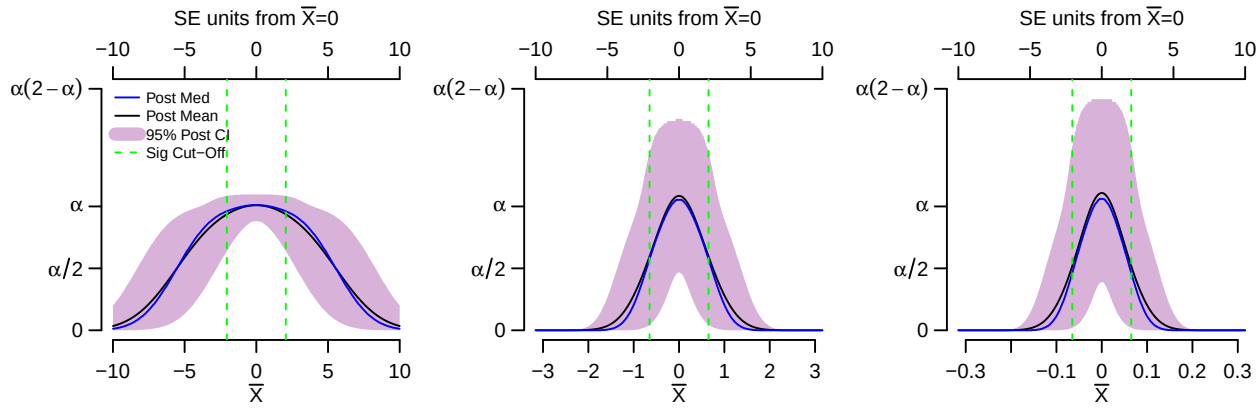
2.3.2.1 Estimating risk of a two-sided test

We first illustrate risk estimation for a Normal location problem, where a sample of size n is drawn from $N(\theta, \sigma^2 = 1)$, we have a standard Normal $N(0, v = 1)$ prior for θ, and we wish to test the null hypothesis $H_0 : \theta = 0$. We perform two-sided tests, following the Bayes rules for loss (2.3), and so have risk

$$\begin{aligned} R(\theta, d) &= \alpha\mathbb{P}[d = \text{No decision}; \theta] + 2I_{\theta<\theta_0}\mathbb{P}[d = \text{Above}; \theta] + 2I_{\theta>\theta_0}\mathbb{P}[d = \text{Below}; \theta] \\ &= \alpha\left[\Phi\left(\frac{c-\theta}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-c-\theta}{\sigma/\sqrt{n}}\right)\right] + 2I_{\theta<\theta_0}\left[1 - \Phi\left(\frac{c-\theta}{\sigma/\sqrt{n}}\right)\right] + 2I_{\theta>\theta_0}\Phi\left(\frac{-c-\theta}{\sigma/\sqrt{n}}\right). \end{aligned}$$

for standard normal distribution function $\Phi(\cdot)$ and value c such that $|\bar{X}| \geq c$ leads to rejection of the null hypothesis. In this case, $c = \sqrt{1/v + n/\sigma^2}\left(1 + \frac{\sigma^2}{vn}\right) z_{\alpha/2}$, where $1-\Phi(z_{\alpha/2}) = \alpha/2$.

Figure 2.5 gives, as a function of the observed sample mean $\bar{X}$, the posterior mean, median and 2.5% and 97.5% quantiles for the risk—all quantities that might be reported after conducting a test. These are given for three settings; a strong prior/weak data, an intermediate setting, and weak data/strong prior, where the prior grows relatively weaker as the sample size grows (from left to right). From it we see that the range of possible estimates and credible intervals for the risk varies considerably depending on both the sample mean (and hence the classic p-value) and the sample size. In a strong prior/weak data setting, with non-extreme observed $\bar{X}$, we should have strong belief that the true risk is not much less than α, the cost for making no-decision. With more extreme observations our beliefs generally support lower values of risk but with wide uncertainty. In contrast the larger sample size results show us that with weaker prior/stronger data, with non-extreme data

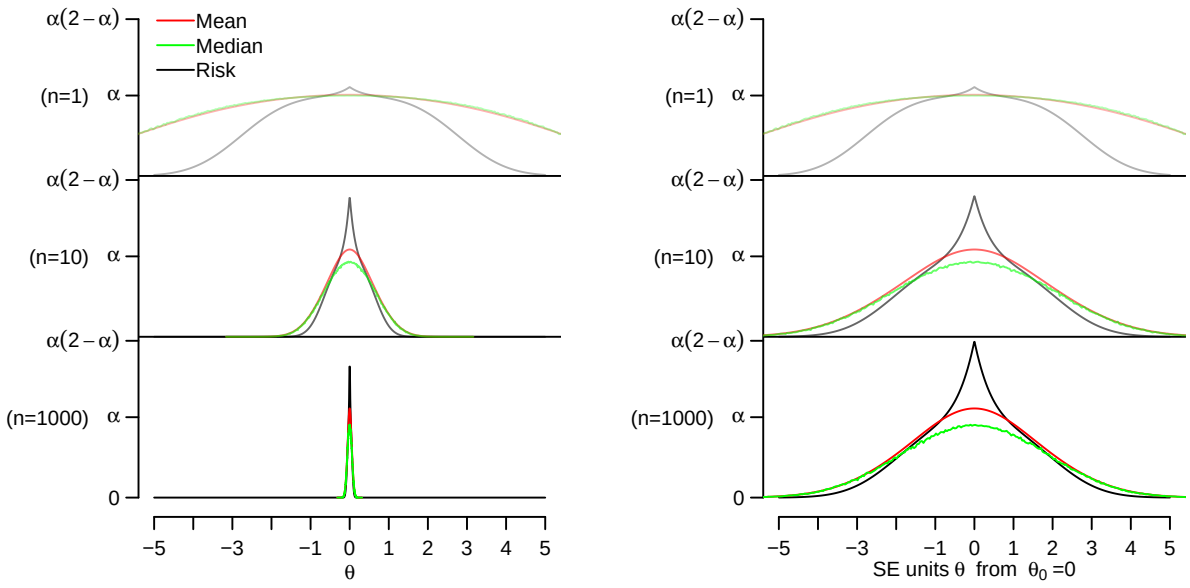


(a) $n = 1$ (b) $n = 10$ (c) $n = 1000$

Figure 2.5: Posterior mean (black) and median (blue) of risk for a two-sided normal location problem with $N(\theta, 1)$ data and standard normal prior for θ. Sample means beyond the green lines lead to rejection of the null hypothesis. 95% posterior credible intervals (equal tails) for risk are provided in purple.

(such as just-significant results), there is very wide uncertainty about the risk, ranging from almost $\alpha(2-\alpha)$, the risk incurred at $\theta = 0$ (the maximum value possible with a limiting flat prior), down to close to zero.

To illustrate broad validity of the posterior's information about the risk, Figure 2.6 shows the expected value (averaging over replicate datasets) of the posterior mean risk and median risk. While for weak data/strong prior settings these estimates of risk over-estimate the true risk (due to the large influence of the prior centered at null value θ_0, where the true risk is greatest), for stronger data/weaker priors, the expectations of these estimates track closely with the true risk for non-null values of θ. An interesting exception to this is the region approximately one standard error either side of θ_0 where the true risk 'spikes' up to near $\alpha(2-\alpha)$, but the estimates are biased low. This region contains parameter values where non-null decisions, when taken, often have the wrong sign, and consequently the size of this region shrinks as the information in the data increases. As it shrinks at the same rate as the uncertainty in the posterior, the posterior mean and median risk are calculated with



(a) (b)

Figure 2.6: Expected calculation of posterior mean (in red) and median (in green) of risk vs. risk for a fixed value of θ and values of θ fixed SE units away from $\theta = 0$ for a normal-normal two-sided test.

high contributions from regions within the 'spike', and smaller contributions outside it. This biases the estimates low.

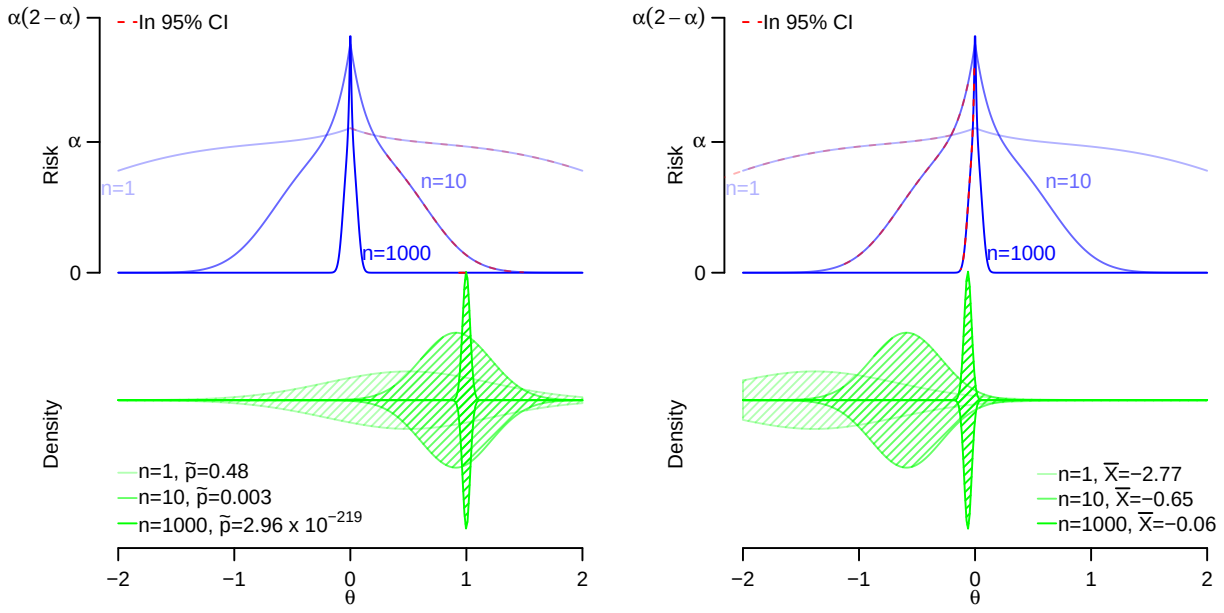
Supposing θ takes a value in the alternative (e.g. θ=1) and estimates approximate the true value of θ across multiple studies with increasing sample size, estimates of risk perform as expected (see Figure 2.7, panel (a) for the Normal location problem with strong, intermediate and weak priors as above): both the posterior mean (i.e. estimate) of risk shrinks towards zero and assuredness in the value of risk increases, mimicking similar shrinking of the posterior distribution about $\theta = 1$. With both increased assuredness in θ being a member of the alternative state space and consistency of the testing procedure, estimated probability of correctly rejecting the null hypothesis increases, thereby reducing estimates of the risk.

As stressed in Section 2.1, our focus is on methods for assessing testing decisions, not

point estimates. It is therefore of interest to assess how the information on risk changes as a function of the observed testing decision—specifically as a function of $\tilde{p}$ discussed in Section 2.3.1 that behaves very much like a classical p-value. It is pertinent to note in Panels (b)-(d) of Figure 2.7 show for just-significant, significant, and non-significant results, while the posterior variance for θ shrinks, the range of values of θ for which the risk spikes shrinks. The relative speed with which these two quantities change with respect to strength of the data (i.e. sample size) dictates change in posterior support for low and high values of risk. We also note that $R(\theta_0, d)$ increases as $n \to \infty$; therefore, some values of high risk are not possible with low sample sizes, restricting inclusion of high values of risk in posterior credible intervals with low sample sizes.

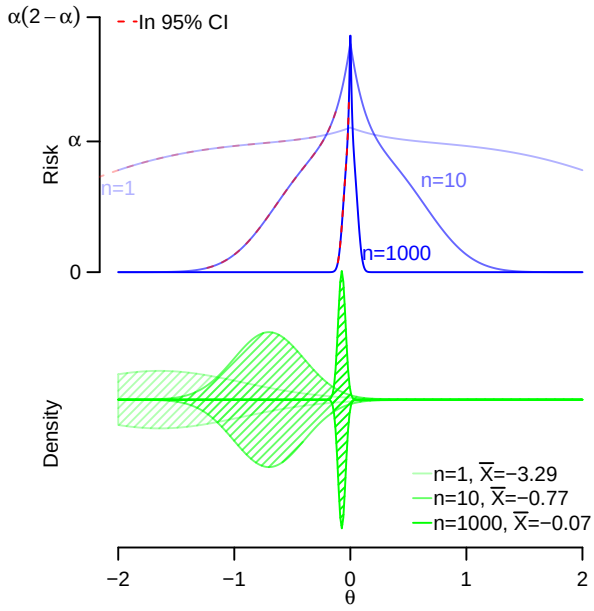
To see the general impact of the prior on estimates of risk and 95% posterior credible intervals, see Figure 2.8. Recall that for a given data set, the posterior mean approximates the posterior median as shown in Figure 2.5; the former was arbitrarily chosen to plot here. As $\tilde{p}$ gets smaller (i.e. with stronger signals) our estimate of the risk also shrinks, reflecting a more reliable testing setup. Also, for fixed $\tilde{p}$, the estimate of risk is monotonic in sample size, with the order switching around $p = 0.50$. This trend results in approximately frequentist inference (with $n = 1000$) yielding the highest estimates of risk with large $\tilde{p}$, and the lowest estimates of risk with small $\tilde{p}$. With increasing sample sizes, we see a somewhat counter-intuitive result: while there is increased precision in data-driven estimates of θ with larger sample sizes, the precision in risk estimation lessens, resulting in wider posterior credible intervals. This phenomenon is due to the competing shrinking of the risk curve 'spike' and posterior variance discussed above.

We are particularly interested in where risk is "high", which we define to be substantial support for the risk exceeding α, or the cost of making no decision no matter the data observed. Recall loss in excess of α is only achieved when an incorrect declaration is made about the sign of θ. Hence, a test with risk exceeding α may be deemed *futile*, since the risk of making no decision regardless of the data is less than the risk incurred by collecting data and attempting a decision. Such a distinction is imperative to motivate or dissuade future

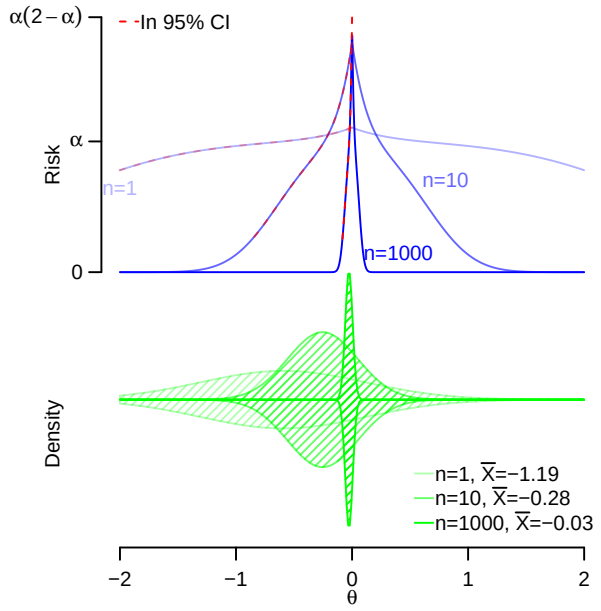


(a) $\bar{X} = 1$

(b) $\tilde{p} = \alpha = 0.05$



(c) $\tilde{p} = 0.02 < \alpha$



(d) $\tilde{p} = 0.40 > \alpha$

Figure 2.7: For a normal location problem where $X \sim N(\theta, \sigma^2 = 1)$ and $\theta \sim N(0, 1)$ testing $H_0 : \theta = 0$, risk (in blue) and posterior density of θ (in green) with given observed sample means $\bar{X}$ or value of $\tilde{p}$, one of which is fixed in each panel. Values of risk included in the 95% posterior credible interval for risk (equal tails) are covered by the black dashed line. Deeper colors signify larger sample sizes and weaker priors relative to the data ($n = 1, 10, 1000$).

testing. Hence, using Figure 2.8, we see that investigators should be wary of futile future testing with non-significant results in general or when the strength of the prior outweighs the data (i.e. the sample size is small). However, in the former case, the larger the sample size, the more support there is for non-futile future testing.

For small $\tilde{p}$, in particular those where the null is rejected, increased reliability of the test is also empirically observed as the relative strength of the data increases. However, after small sample sizes there are rapidly-diminishing returns; compare the red/blue lines and blue/black lines. Posterior support for very low values of risk is only acquired with notable strength in the data. In addition, for very small values of $\tilde{p}$, testing settings with stronger data have little support for futile testing compared to testing settings with stronger priors. The posterior credible interval of risk with approximately frequentist analysis stretches nearly to α for values of $\log(\tilde{p}) \geq \log(0.005)$, suggesting lack of futility, and therefore reliability, of tests with $\tilde{p} \approx p$ below this threshold. In these cases, a true signal is nearly assured. This conclusion is consistent with Johnson [2013]'s suggested p-value threshold of 0.005; his p-value threshold of 0.001 for "highly significant" results may also be justified using risk estimation as near-assurance the test is not futile. The same argument using credible intervals nearly holds for $\tilde{p} = 0.01$, approximating the threshold for significant p-values guaranteeing intrinsic credibility of significance due to Matthews [2018].

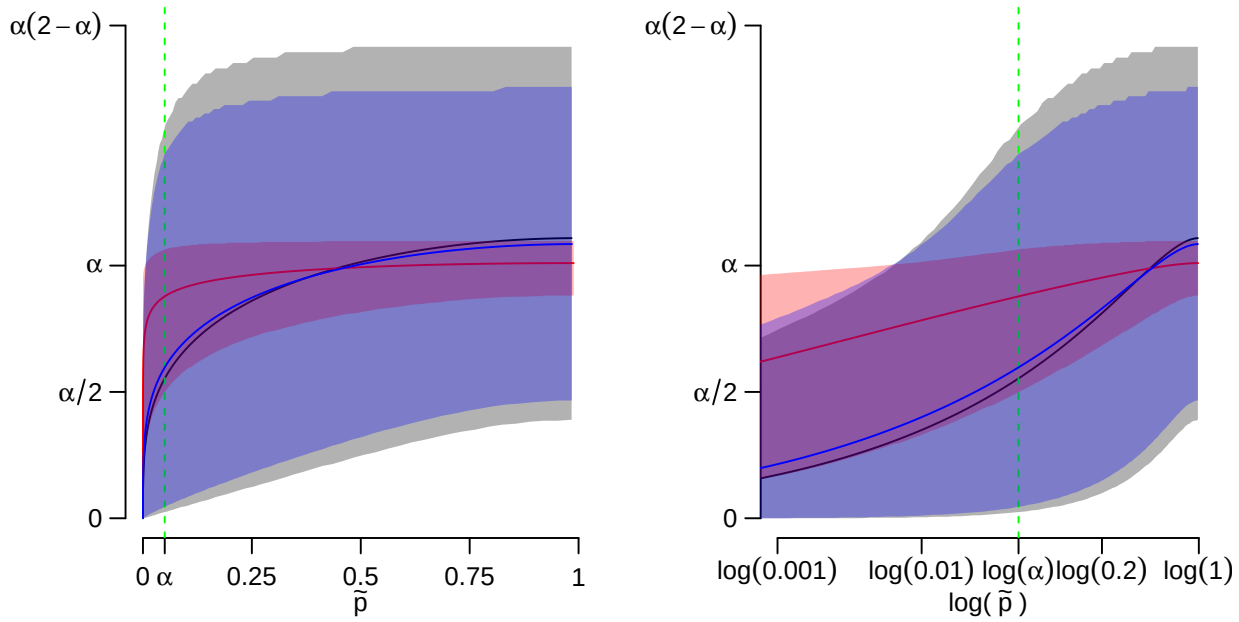
Risk estimation may also be related to Shafer [in press]'s nomenclature for testing by betting. Suppose the price of a bet on the alternative is dictated by the expected value of a statistic under the null. If evidence leads to retrieval of five times the money bet, the test "merits attention". Applying analogous logic to risk estimation, we may suppose a test "merits attention" when the risk is a fifth of a futile test, i.e. risk is $\alpha/5$. Aforementioned thresholds for $\tilde{p}$ with reasonable strength in the data correspond to point estimates of risk approximately equal to $\alpha/5$ or less. However, if the merit of attention is earned with strong posterior support for risk below $\alpha/5$, we can see it is difficult to achieve, and impossible in some cases. This comparison is discussed in greater depth by Krakauer and Rice [in press].

With just-non-significant results, while estimates of risk lie below the futility threshold

of α, there is substantial support both for values of risk well in excess of α and values near 0. Therefore, the merit of such a test remains substantially uncertain. This contrasts sharply with the practice of describing just-non-significant results enthusiastically, as a "trend toward significance" [Wood et al., 2014], indicating that a replicate study (or slightly more data) would yield significant results with high probability. Our result shows that the same results are compatible with future tests for which a considerable fraction of significant result are due to incorrect sign declarations. Thus, there is support for future significant test results being in the wrong direction, if not failing to reach significance at all. Naturally, the same holds for just-significant results. Hence, it is entirely plausible we are in a low-risk setting (and appropriately making a decision about the sign of θ) *or* a futile one, in which making a decision is not worthwhile and the incorrect sign for θ has been declared.

For non-significant results, we see that risk estimation provides little information over $\tilde{p}$. Estimated risk remains relatively constant with similarly near-constant upper bounds near the maximum value of risk that is reasonably plausible with moderately informative data; with stronger priors, posterior support is very concentrated around α. As $\tilde{p}$ approaches 1, the intervals for the mildly- and weakly-informative prior settings indicate that low values of risk (say below $\alpha/2$) can be ruled out, but maximal values of risk remain plausible. This finding is not particularly helpful with interpretation of results, and hence there is little information gained by knowing whether a tail area was just above α or was close to 1 for a fixed sample size.

For more information on posterior distributions of the risk, see Appendix B.1. In particular, note that while posterior credible intervals widen in both directions (i.e. include more values of risk at both extremes) as the sample size increases with just-significant results, these intervals change due to both shifting of the posterior distribution to have increased relative support for lower values of risk *and* skewness of the distribution, resulting in longer right-hand tails (e.g. Figure B.1). Unfortunately, the posterior distribution of the risk does not have a closed form; therefore, understanding how support for certain values of risk shifts with the increase in sample size (beyond inclusion/exclusion from a posterior credible interval)



(a) (b)

Figure 2.8: 95% posterior credible intervals for risk by twice minimum posterior tail area (denoted by $\tilde{p}$) on the left, and $\log(\tilde{p})$ on the right with posterior mean of risk, n=1, 10, and 1000 in red, blue, and black, respectively. The risk is estimated for a normal location problem with $X \sim N(\theta, \sigma^2 = 1)$ and $\theta \sim N(0, 1)$ testing $H_0 : \theta = 0$.

will need to be replicated for any other $\tilde{p}$ of interest.

2.3.2.2 Estimating risk of one-sided tests

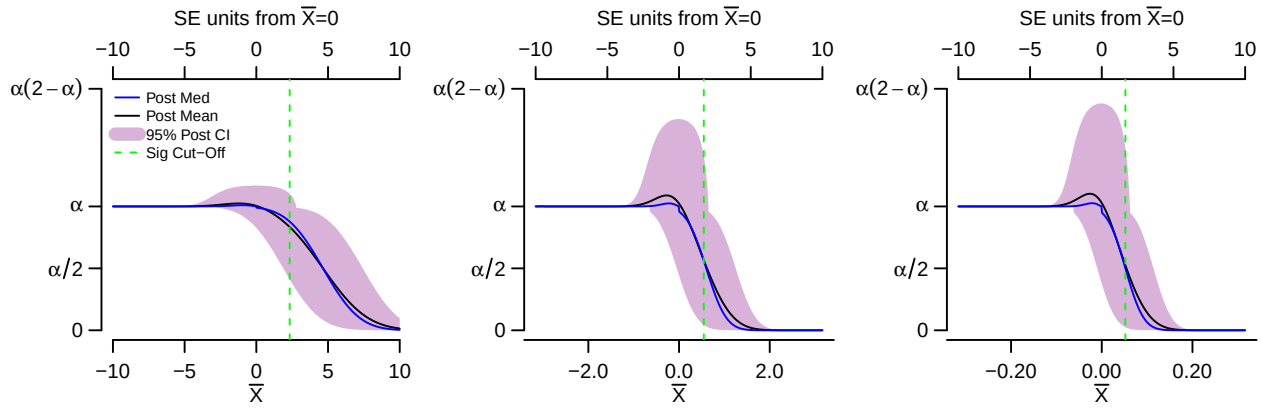
To explore risk estimation with a one-sided test, we continue with the Normal location problem, but replacing the point null hypothesis with $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$ where $\theta_0 = 0$ without loss of generality. The prior and data distributions remain the same. Using the Bayes rule for loss 2.2, the risk for the test becomes

$$\begin{aligned} R(\theta, d) &= \alpha Pr(d = \text{No decision}; \theta) + I_{\theta<\theta_0} Pr(d = \text{Above}; \theta) \\ &= \alpha \Phi\left(\frac{c' - \theta}{\sigma/\sqrt{n}}\right) + I_{\theta<\theta_0}\left[1 - \Phi\left(\frac{c' - \theta}{\sigma/\sqrt{n}}\right)\right], \end{aligned}$$

where $\bar{X} \geq c'$ leads to rejection of the null hypothesis. In this case,
$c' = \sqrt{1/v + n\sigma^{-2}}\left(1 + \frac{\sigma^2}{vn}\right) z_\alpha$ where $1 - \Phi(z_\alpha) = \alpha$. We note that, unlike the two-sided case, the risk is no longer symmetric about θ_0; see Figure 2.10 for some illustrative examples.

Lack of symmetry of the risk curve about θ_0 leads to lack of symmetry in estimates of the risk across values of $\bar{X}$, as shown in Figure 2.9. For any $\bar{X}$, as with two-sided tests, changes in estimates of risk with increasing strength in the data reflect changes in shape of the risk curve for values of θ near $\bar{X}$. For instance, more values of $\bar{X} < 0$ give estimated risk approximately equal to α (noting the change in scale of the horizontal axis, and hence the "hump" is closer to 0). These estimates reflect the widening plateau in the risk curve from the increased probability in no active decision being made for any fixed $\theta < \theta_0$. With $\bar{X} > 0$, two trends are apparent: (1) the range of positive values of $\bar{X}$ yielding non-negligible positive estimates of risk shrinks and (2) the "spike" in estimated risk (and related upper credible interval bound) takes place over a narrower range of values of θ. As with two-sided tests, upper bounds of credible intervals within this spike increase since the maximum risk $R(\theta_0, d)$ increases as $n \to \infty$.

The bias of the estimation technique—assessed by comparing expected calculation of risk

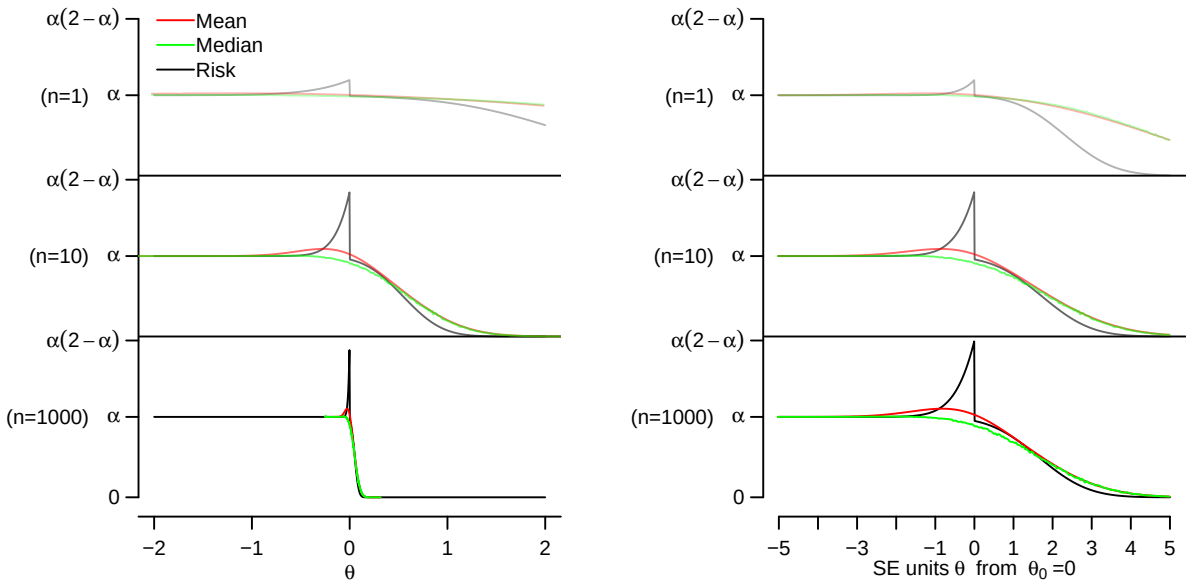


(a) $n = 1$ (b) $n = 10$ (c) $n = 1000$

Figure 2.9: Posterior mean (black) and median (blue) of risk for a one-sided normal location problem with $N(\theta, 1)$ data and standard normal prior for θ. Sample means greater than the green lines lead to rejection of the null hypothesis. 95% posterior credible intervals (equal tails) for risk are displayed in purple.

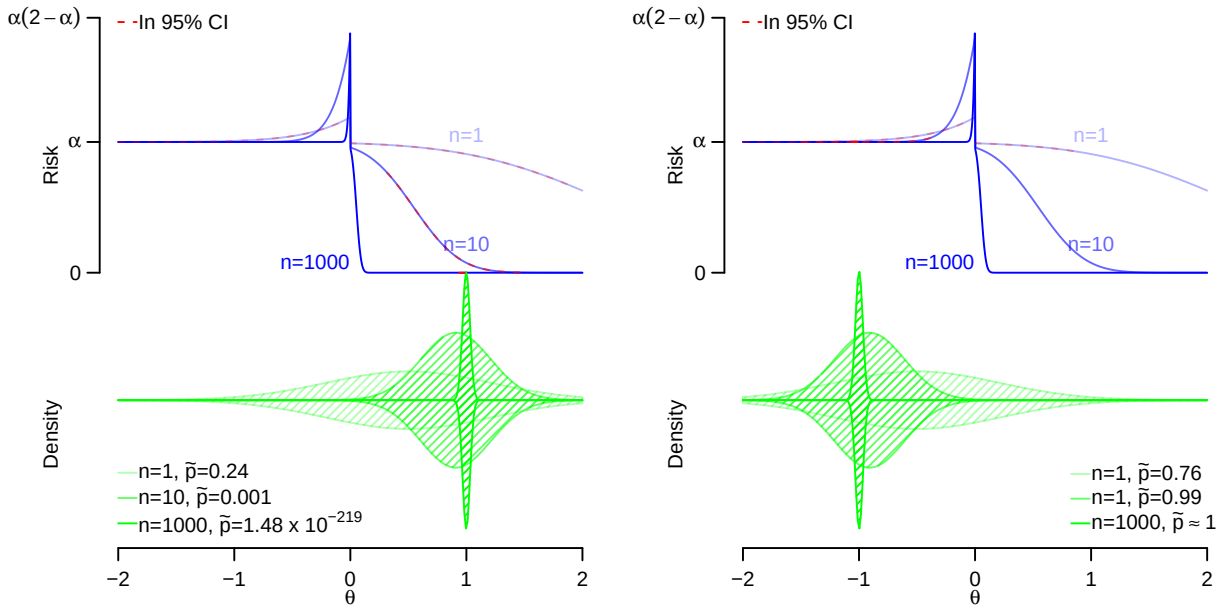
estimates (posterior means and medians of risk) to the risk curve—is presented in Figure 2.10. Bias for θ in the alternative state space is similar to the two-sided test: overestimation of risk for extreme θ is remedied by stronger data. In general, estimation of risk is quite accurate in the null parameter space (i.e. $\theta < \theta_0$), a by-product of frequent lack of decision for values of θ in the null space. Nearly identically to the two-sided testing scenario, estimates of risk are not expected to capture the spike in true risk within one standard error unit from θ_0, the one interval in the null state space where there is non-negligible probability the null will be inappropriately rejected. However, unlike two-sided tests, the spike is present only for values of θ in the null space; this result is due to loss greater than α only being plausible for $\theta \in \Theta_0$. This spike is less extreme than for two-sided tests, noting the increase in risk with one-sided tests is relative to α, the lowest possible risk possible with $\theta < \theta_0$, as opposed to near 0 with two-sided tests.

With $\theta \in \Theta_A$, estimates of risk behave as expected as the data provide more precise estimates of θ, as shown in the left-hand panel of Figure 2.11: supposing $\theta = 1$ and $\bar{X} \approx \theta$



(a) (b)

Figure 2.10: Expected calculation of posterior mean (in red) and median (in green) of risk vs. risk (black) for a fixed value of θ and values of θ fixed SE units away from $\theta = 0$ for a one-sided Normal location problem where $X \sim N(\theta, \sigma^2 = 1)$, $\theta \sim N(0, 1)$, and $H_0 : \theta < 0$.



(a) $\bar{X} = 1$ (b) $\bar{X} = -1$

Figure 2.11: Risk (in blue) and posterior density of θ (in green) with given observed sample means $\bar{X}$. Values of risk included in the 95% posterior credible interval for risk (equal tails) are covered by the red dashed line. Deeper colors signify larger sample sizes ($n = 1, 10, 1000$).

across all experiments, risk estimates universally approach 0. Unique to one-sided tests, we may consider the same trend for $\theta \in \Theta_0$ (in the right-hand panel of Figure 2.11), a scientifically plausible eventuality (as opposed to the "straw man" point null). Note that if $\theta \in \Theta_0$, $R(\theta, d) \to \alpha$ as $n \to \infty$, and $R(\theta, d) \geq \alpha$ for any n. Suppose we continue to have $\bar{X}$ approximating θ across experiments and $\theta \in \Theta_0$, e.g. $\theta = -1$. Despite increased precision in estimation of θ, estimates of risk approach $\alpha > 0$ with near certainty as the data become stronger relative to the prior, reflecting loss incurred for failing to make a decision, despite increased assurance that $\theta \in \Theta_0$. Such loss is a unique attribute to significance tests, as opposed to hypothesis tests.

As functions of $\tilde{p}$, estimates of risk and precision of those estimates behave similarly for one- and two-sided tests, as displayed in Figure 2.12. Differences in risk estimates across

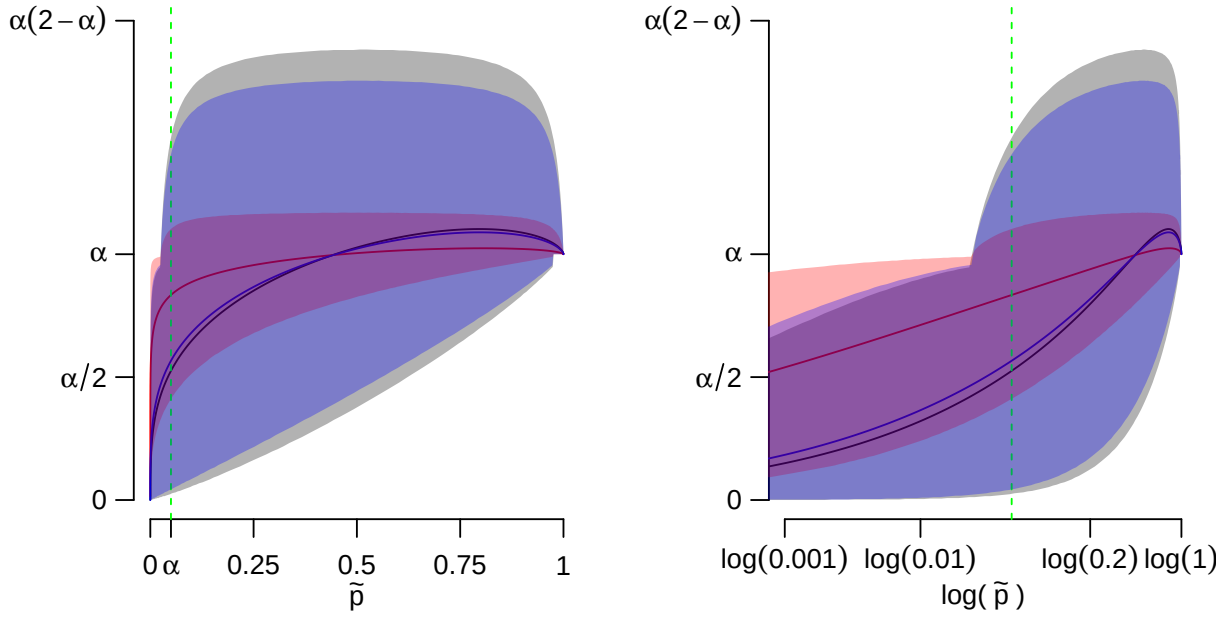
both relative strengths of the data and across different values of $\tilde{p}$ (including both significant and non-significant results) are generally quite similar. However, unlike two-sided tests, as posterior assurance that $\theta \in \Theta_0$ increases, the posterior probability that repeated experiments with the same strength of data will lead to failure to reject the null hypothesis (resulting in loss α) increases. Hence, with one-sided tests, support for small values of risk with non-significant results decreases sharply as $\tilde{p}$ increases, where the lower bound for the 95% posterior credible interval approaches α as $\tilde{p}$ approaches 1. Furthermore, this increased assurance that the risk is close to α for large $\tilde{p}$ also leads to a drop in the upper bound for the posterior credible interval, in contrast to consistent support for very high values of risk for nearly every non-significant two-sided test.

As with two-sided tests, the interpretation of a "trend toward significance" [Wood et al., 2014] should be used with trepidation, noting there is substantial evidence future tests might either appropriately or inappropriately reject the null due to support for low and high risk (relative to α), respectively. Slightly higher values of $\tilde{p}$ retain substantial support for risk below the futility threshold of α compared to two-sided tests—certainly as high as $\log(0.01)$, even with stronger priors. Thus, Johnson [2013]'s p-value threshold recommendations could perhaps be relaxed with one-sided tests according to this metric. With moderate strength of data, values of $\tilde{p}$ that "merit attention" (using Shafer [in press]'s language of betting scores) are similar to those that merit attention from two-sided tests. The same issue arises if merit is earned with posterior support for risk below $\alpha/5$.

For an example of the posterior distribution of risk for a given $\tilde{p}$, see the right-hand panel of Figure B.1 in Appendix B.1 where $\tilde{p} = \alpha$. This figure shows sharp changes in distribution of risk for the one-sided case near higher values of the distribution, contributing to bumps in upper bounds of posterior credible intervals in Figure 2.12.

2.4 Discussion

The skeptical reader may call the necessity of the research described above into question, arguing that standard practice dictates calculation of power before a test is conducted,



(a) (b)

Figure 2.12: 95% posterior credible intervals for risk by the posterior tail area (denoted by $\tilde{p}$) on the left and $\log(\tilde{p})$ on the right with posterior mean of risk, n=1, 10, and 1000 in red, blue, and black, respectively. Risk is estimated for a normal location problem where $X \sim N(\theta, 1)$ and $\theta \sim N(0, 1)$ testing $H_0 : \theta \leq 0$. Significance is reached if $\tilde{p} < \alpha = 0.05$, denoted by the green dashed line.

thereby removing the necessity to evaluate the merit of a test. However, in reality, the faith that may be placed in results from any study, even those for which power analyses were previously conducted, are consistently questioned, e.g. the controversial results of the ANDROMEDA-SHOCK study discussed by Spiegelhalter [2020]. Despite usefuleness of power calculations to plan a study, they require an assumed value of the effect size; therefore, with no knowledge of the effect size being tested, there is no certainty the desired power of the test was reached. As a result, the astute investigator will question if significance was due to detection of a discrepancy from the null that is practically meaningless or a chance false positive; we may also question if a non-significant result motivates future data collection or is a call to cease investigation due to veracity (or near veracity) of the null.

Every decision with policy or treatment implications raises inherent questions of cost, i.e. how much we are willing to risk losing with each decision made. Therefore, even if severe consequences are not entirely assured, relative costs dictates specific actions. A decision-theoretic approach to testing provides a mathematically-justified rule, taking probabilities of incorrect decisions and cost of each decision into account. Rice et al. [2020] showed how, when only the sign of θ is of interest, significance tests based on posterior tail areas can be straightforwardly motivated as decision rules. Hence, we may incorporate prior knowledge about the parameter of interest or approximate frequentist inference. Additionally, the posterior tail area takes the definition often inappropriately given to the p-value, e.g. $Pr(\theta < \theta_0)$.

This new testing procedure allows for new nomenclature to quantity reliability of significance tests: loss and risk. Unlike many previous methods, the loss function provides natural calibration for levels of these quantities that should be sought, which we have called the *futility threshold*. The posterior distribution of each quantity may be estimated and functionals reported. As with *post hoc* observed power calculations, estimates of loss fail to provide information beyond posterior tail areas used to conduct the initial inference.

On the other hand, estimates of risk provide novel information about a test. We have seen that with fixed relative strength of the data and prior, risk estimates decrease monoton-

ically with evidence in favor of the alternative. However, unlike observed power, the exact relationship depends on the relative strength of prior and data. We have also seen that as data overwhelms the prior, precision in our estimates of risk lessens. This finding has important implications for just-non-significant results, for which there is substantial evidence future testing will yield significance—in-keeping with a "trend toward significance" [Wood et al., 2014]—albeit, possibly in the wrong direction. With similar assessment of the reliability of just-significant results, enthusiasm for such results should be dampened. Instead, our recommended thresholds for tests that warrant enthusiasm—e.g. frequentist p-values below approximately 0.005—are somewhat consistent with recommendations given by Johnson [2013], Matthews [2019], and to a lesser extent, Shafer [in press].

Future research may begin by further exploring the relationship between the frequentist p-value and estimates of risk, in particular if posterior quantiles and/or means of risk are the *same* one-to-one function of p-values—as seen in Appendix B.2—for all priors yielding identical frequentist and Bayesian inference. Extensions of estimates of risk to scenarios with multiple unknown parameters and nuisance parameters would be ideal.

Finally, future researchers may consider defining the posterior distribution of power, especially because the risk function is a weighted average of multiple power functions. Such a calculation would be a posterior version of Bayesian assurance [Chen and Chen, 2018, Spiegelhalter, 2004], a calculation generally completed to find the necessary sample size for a study. Finding the posterior distribution, and therefore variance, of power might negate issues of over-confidence in observed power calculations [Gelman, 2019b] and allow for continued estimation of a quantity of general familiarity to researchers.

Chapter 3

DECISION THEORETIC JUSTIFICATION FOR WINNER'S CURSE REDUCTION

Abstract

Default estimators often fail to retain their ideal properties (e.g. unbiasedness, admissibility, etc.) when used in non-regular settings. A classic example is the Winner's Curse bias—upward bias in estimates when reports are limited to significant results. Improvements to default estimators, such as Zhong and Prentice [2008]'s amelioration of the Winner's Curse bias, often distill to *shrinkage*, or the movement of a complex decision to a simple or parsimonious one. We introduce a single family of loss functions that motivates existing improvements to default estimators using a decision-theoretic framework. The loss function quantifies the trade-off between accuracy and parsimony. Hence, estimators created for different purposes may newly be united in this trade-off, thus motivating assessment of how well all estimators in this newly defined class improve a single quality, e.g. reducing Winner's Curse bias. This Bayesian framework further allows for incorporation of prior skepticism or information about a parameter being tested. Our particular focus is on decision-theoretic justification for Zhong and Prentice [2008]'s estimators, its ability to reduce the Winner's Curse bias compared to other estimators in this newly-defined class, and how prior skepticism affects bias reduction.

3.1 Introduction

Estimators of parameters of interest θ may be selected for multiple default reasons—maximizing a likelihood, unbiasedness, consistency, etc. In regular settings default methods are optimal, or as close to optimal that no further justification is needed. But this is not case in non-

regular scenarios, for which default estimates were not originally intended, and where these estimates may not have their default properties. A common example is a form of selection bias called the *Winner's Curse* [Capen et al., 1971], a bias away from the null (used to define a hypothesis about θ) when point estimates are only reported following statistically significant test results.

It is therefore natural to want to improve default estimates to retain their default properties in non-regular settings. Existing modifications include [Zhong and Prentice, 2008]'s reduction of bias due to the Winner's Curse, Stein [1956]'s admissible estimator of three or more Normal locations estimated simultaneously, and Thompson [1968]'s shrinkage of the minimum variance linear unbiased estimator of a population mean that reduces mean squared error. Despite their disparate motivations, all these estimators (among others) boil down to *shrinkage* of the original point estimate. *Shrinkage* may generally be defined as the process of moving an estimate or decision from a complex value towards a simple or parsimonious one. These estimators prove the quality of an estimator might be improved by making it *simpler*, even if it is not the direct intent of the creator.

The extent of shrinkage may be constant, i.e. pre-determined and not affected by the data observed, or *data-adaptive* instead. The shrinkage of data-adaptive approaches may be a continuous function of the data, or discontinuous. Our particular focus is on two data-adaptive shrinkage estimators due to Zhong and Prentice [2008] that shrink point estimates of θ toward the null to ameliorate Winner's Curse bias.

With the understanding a bevy of disparately-motivated estimators all effectively *shrink* point estimates, we may question how well *any* shrinkage estimator improves a default estimator falling short in one respect. For example, the simpler and analytically-tractable estimators due to Stein [1956] or Thompson [1968] might have similar or superior Winner's Curse bias amelioration to estimators due to Zhong and Prentice [2008], despite separate motivations for their genesis. Before using any shrinkage estimator, it is vital to have some interpretation of the shrinkage conducted, ensuring modifications to the original estimate are sensible. Furthermore, with only frequentist justifications available for many shrinkage

estimators, it is not clear how prior information about θ might be incorporated in a Bayesian analysis.

Both issues can be resolved by having the degree of shrinkage be a Bayes rule to an explicit loss function in a Bayesian decision-theoretic analysis. We define one broad family of such loss functions below, which happens to yield Bayesian analogs of all the aforementioned data-adaptive shrinkage estimators, and others. Our family of loss functions makes explicit trade-offs between *accuracy* and *parsimony*—i.e. the amount of inaccuracy one is willing to trade for some amount of simplification of the estimator—and thus facilitates interpretation of the degree of shrinkage across disparately-motivated estimators. If an estimator defined with the intent of improving a property of a default estimator is justified using this class of loss functions—i.e. shrinks by trading some degree of parsimony for accuracy—it is therefore sensible to consider how well other estimators within this newly-defined class improve the same property. Thus, the aforementioned comparison across shrinkage estimators is highly motivated. Our exploration of how this framework allows prior information to be used focuses on the estimators due to Zhong and Prentice [2008], and the extent to which decision-theoretic shrinkage and prior-based shrinkage interact.

The rest of the chapter's structure is as follows. In Section 3.2, we will quantify the magnitude of Winner's Curse biases, then define and illustrate two estimators developed by Zhong and Prentice [2008] to ameliorate this bias. Next, Section 3.3 defines a class of loss functions yielding data-adaptive shrinkage Bayes rules, and give methods to help devise loss functions that retroactively justify particular shrinkage functions. These methods are used to justify Bayesian analogs of the estimators due to Zhong and Prentice [2008] in Section 3.4, motivating the comparison of its ability to ameliorate the Winner's Curse to other shrinkage estimators justified using the same loss function family, e.g. Stein [1956]'s and Thompson [1968]'s estimators. Furthermore, we may compare their inherent parsimony/accuracy trade-offs. Shrinkage from prior specification alone and interaction of the prior with estimators from Zhong and Prentice [2008]—newly applicable to Bayesian inference—will also be explored in Section 3.4.1. Based on our primary findings, we also show how a simple "step-function"

shrinkage (i.e. constant shrinkage across specified ranges of data) is capable of similar bias correction from Zhong and Prentice [2008]'s estimators in Section 3.5. Further exploration reveals how completely unbiased estimates may be achieved, albeit with nonsensical choices of shrinkage that are no longer justified using the aforementioned family of loss functions, thus showing the need for realistic tradeoffs to motivate methods, instead of seeking only unbiasedness.

3.2 The Winner's Curse and Zhong and Prentice [2008]'s Corrected Estimates

The term *Winner's Curse* was coined for economic use by Capen et al. [1971] to describe money lost to the "winners" of drilling rights auctions, due to overstating the value of a reserve in order to guarantee its acquisition—the winning bid invariably over-estimates the value of a reserve. However, this loss by the winner may be prevented if all bids are systematically "shrunk" [Young et al., 2008].

In statistics, *Winner's Curse* has been adopted to more generally describe the bias of estimated effect sizes away from the null if (1) only estimates from significant tests are reported and (2) the same data are used to conduct the original test and estimate the effect size. It is therefore a form of regression to the mean. For target of estimation θ, the bias equates to stating that

$$\mathbb{E}\left[|\hat{\theta} - \theta_0| \mid \text{significant test result assessing } \mathrm{H}_0 : \theta = \theta_0\right] > |\theta - \theta_0|,$$

where θ_0 is invariably zero in practice.

This form of selection bias is present across many fields, including genome-wide association studies (GWAS), in which only significant estimated associations between single nucleotide polymorphisms (SNPs) and an outcome are typically reported. Consequently, SNP associations failing to reach the significance threshold are not studied further, and reported estimates of association are biased away from the null.

This strong preference for significance of results in any field—as shown and/or discussed

by Krzyzanowska et al. [2003], Stern and Simes [1997], and Chavalarias et al. [2016], among others—biases its estimates away from the null. The degree of bias is one way to quantify the impact of "publication bias" [van Zwet and Cator, 2020, Gelman and Carlin, 2014]. Its impact may be non-trivial, e.g. published inflated results affecting clinical treatment decisions may not be refuted for years following the initial publication [van Zwet and Cator, 2020].

Unfortunately, the extent of Winner's Curse bias is dictated by the unknown true value of θ, as demonstrated below. Suppose $\hat{\theta}$ is asymptotically normally distributed and a two-sided Wald test $H_0 : \theta = \theta_0$ is conducted. (Throughout we assume standard error σ of $\hat{\theta}$ is estimated with $\hat{\sigma}$, but this estimate may be simply replaced by σ if known in any formulation.) Due to the assumed-normal distribution of the test statistic, denoted by $Z = (\hat{\theta} - \theta_0)/\hat{\sigma}$, the test result is significant at level α if $|Z| > \Phi^{-1}(1 - \alpha/2) \equiv c$, where $\Phi(\cdot)$ denotes the CDF of a standard normal variable. For the usual case where $\theta_0 = 0$, the asymptotic distribution of the *reported* effect size estimate is a truncated normal [Garner, 2007] with expected value

$$\mathbb{E}_{\hat{\theta}\left|\left|\frac{\hat{\theta}}{\hat{\sigma}}\right|\geq c\right.}(\hat{\theta};\theta) = \theta + \underbrace{\sigma\frac{\phi\left(\frac{\theta}{\sigma} - c\right) - \phi\left(-\frac{\theta}{\sigma} - c\right)}{\Phi\left(\frac{\theta}{\sigma} - c\right) + \Phi\left(-\frac{\theta}{\sigma} - c\right)}}_{\text{Bias}}. \tag{3.1}$$

Equation 3.1 shows how the bias depends on the test's chosen level α, the unknown value of θ, and random variability as reflected in σ. The black curves in Figure 3.1 show the bias of $\hat{\theta}$ if reports are limited to significant results. Empirically, it is clear that the bias shrinks to zero as the true value of θ gets further from the null and is particularly problematic for values of θ near the null. Interestingly, while the level α and values θ and σ are all determinants of power of the test, the bias is seen to be non-monotonic with respect to power, albeit not far from this property. Monotonicity does hold true in some cases, e.g. for estimates of differences in success probabilities, with significance determined by a Pearson χ^2 test [Xiao and Boehnke, 2009]. For a fixed value of θ, empirically we can see bias increases as the level decreases (i.e. power decreases). These trends suggest particularly severe biases in GWAS, which to control family-wise Type I errors typically use extremely small Type I error rates

for individual tests, but aim to detect very small effect sizes (i.e. the true value of θ is close to θ_0). As reflected by scaling of the vertical axis, increasing the variability (i.e. increasing power) will decrease the bias for a fixed value of θ.

Zhong and Prentice [2008] devised effective and intuitive strategies to ameliorate Winner's Curse biases, using the formulation given in Equation (3.1). Our focus is on two of their three adjustments to be used in one-stage designs, as opposed to multi-stage designs that filter covariates in earlier stages with reports of effect sizes limited to subsequent independent stages.

Their first approach corrects the acquired estimate $\hat{\theta}_{\text{observed}}$ to

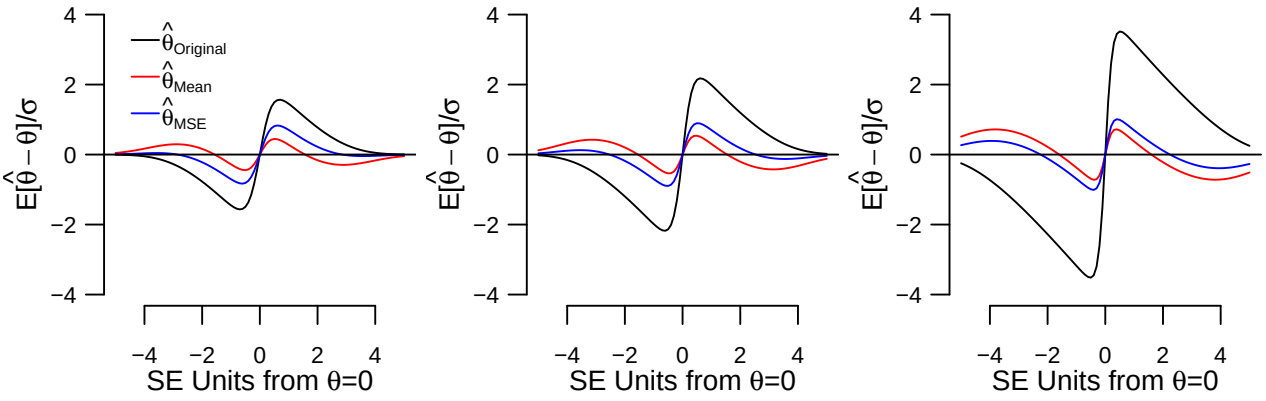
$$\hat{\theta}_{\text{Mean}} = \delta : \mathbb{E}_{\hat{\theta}||Z|\geq c}(\hat{\theta}; \delta) = \hat{\theta}_{\text{observed}}, \tag{3.2}$$

which can be seen as a solution to a plug-in version of Equation (3.1). Throughout the remainder of this chapter, we will often denote $\hat{\theta}$ by $\hat{\theta}_{\text{observed}}$ to differentiate it from corrected estimators; however, $\hat{\theta}$ will continue to denote the original uncorrected estimate of θ throughout, and will be used when its meaning is clear. The third of three estimators defined by Zhong and Prentice [2008]—not considered here—can be similarly derived by plugging $\hat{\theta}$ into equations defining the median of the reported results.

Assuming the asymptotic truncated normal density $f(\hat{\theta}||Z| > c; \theta)$ for reported estimates of effect sizes, $\hat{\theta}_{\text{Mean}}$ may also be motivated by noting that

$$\hat{\theta}_{\text{Mean}} = \text{argmax}_{\theta} f(\hat{\theta}||Z| > c; \theta), \tag{3.3}$$

as proven by Zhong and Prentice [2008, Theorem 3.1]. In simulations of adjusted odds ratio estimates in GWAS, this estimator was found to over-correct with moderate effect sizes. This also occurs with tests of normal location θ, as seen in Figure 3.1's red curves. For fixed θ, the over-correction becomes more extreme as α decreases, which is particularly problematic for GWAS.



(a) $\alpha = 0.05$ (b) $\alpha = 0.01$ (c) $\alpha = 10^{-5}$

Figure 3.1: Bias of original estimates (in black) and two corrected estimators ($\hat{\theta}_{\text{Mean}}$ in red and $\hat{\theta}_{\text{MSE}}$ in blue) when estimates of normal location $\hat{\theta}$ are reported only from significant results.

To account for the over-correction, Zhong and Prentice [2008] developed another estimator $\hat{\theta}_{\text{MSE}}$ that is a weighted average of $\hat{\theta}_{\text{Original}}$ and $\hat{\theta}_{\text{Mean}}$, with weight on the former dictated by the approximate percent of MSE of $\hat{\theta}_{\text{Original}}$ from the variance. Specifically,

$$\hat{\theta}_{\text{MSE}} = \hat{K}\hat{\theta}_{\text{Original}} + (1 - \hat{K})\hat{\theta}_{\text{Mean}}, \tag{3.4}$$
$$\text{where } \hat{K} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (\hat{\theta}_{\text{Original}} - \hat{\theta}_{\text{Mean}})^2}.$$

While not considered here, [Zhong and Prentice, 2008] apply a similar adjustment to $\hat{\theta}_{\text{Median}}$, replacing $\hat{\theta}_{\text{Mean}}$ in Equation 3.4 with the median-motivated corrected estimator mentioned above.

Weighting in this way means giving a correction to $\hat{\theta}_{\text{Original}}$ that is always less extreme than $\hat{\theta}_{\text{Mean}}$. With these particular weights, if $|\hat{\theta}_{\text{Original}} - \hat{\theta}_{\text{Mean}}|$ is small, $\hat{\theta}_{\text{MSE}}$ favors the original estimate and adjusts by less than $\hat{\theta}_{\text{Mean}}$'s adjustment. As illustrated by Figure 3.1's blue curves, stopping the over-correction for larger θ values is at the cost of larger bias with values of θ near the null, particularly for moderate values of α.

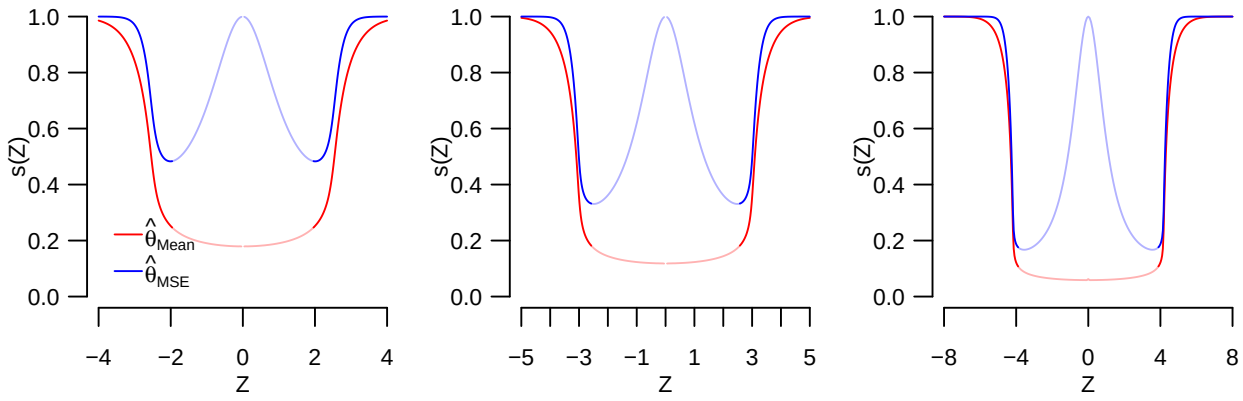
Clearly, in order to address a bias away from the null, both $\hat{\theta}_{\text{Mean}}$ and $\hat{\theta}_{\text{MSE}}$ shrink $\hat{\theta}_{\text{Original}}$ towards the null. Suppose the degree of shrinkage is some value $s(x) \in \{0, 1\}$ that depends on the data X, i.e. $\hat{\theta}_{\text{Mean}} = s_{\text{Mean}}(x)\hat{\theta}_{\text{Original}}$ and $\hat{\theta}_{\text{MSE}} = s_{\text{MSE}}(x)\hat{\theta}_{\text{Original}}$. (We note this function may depend on other fixed parameters, including the level α of the test.) Denoting $\hat{\theta}_{\text{Original}}$ as $\hat{\theta}_{\text{O}}$ and s_{Mean} as s_{E}, $s_{\text{E}}(x)$ is the solution to

$$\begin{aligned}
\hat{\theta}_{\text{O}} &= s_{\text{E}}(x)\hat{\theta}_{\text{O}} + \sigma \frac{\phi\left(\frac{s_{\text{E}}(x)\hat{\theta}_{\text{O}}}{\sigma} - c\right) - \phi\left(-\frac{s_{\text{E}}(x)\hat{\theta}_{\text{O}}}{\sigma} - c\right)}{\Phi\left(\frac{s_{\text{E}}(x)\hat{\theta}_{\text{O}}}{\sigma} - c\right) + \Phi\left(-\frac{s_{\text{E}}(x)\hat{\theta}_{\text{O}}}{\sigma} - c\right)}, \\
\implies Z &= s_{\text{E}}(x)Z + \frac{\phi\left(s_{\text{E}}(x)Z - c\right) - \phi\left(-s_{\text{E}}(x)Z - c\right)}{\Phi\left(s_{\text{E}}(x)Z - c\right) + \Phi\left(-s_{\text{E}}(x)Z - c\right)}.
\end{aligned}$$

This reparameterization shows that, while an explicit solution to $s_E(x)$ does not exist, its value depend on the data only through the Z-score. Therefore, $s_{\text{Mean}}(\cdot)$ may be reparameterized to be a function of Z; we will do this throughout. The dependence on Z alone also holds for $\hat{\theta}_{\text{MSE}}$—shortened to $\hat{\theta}_{\text{M}}$—as we see by noting that

$$\begin{aligned}
\hat{K} &= \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + (\hat{\theta}_{\text{O}} - \hat{\theta}_{\text{E}})^2} \\
&= \frac{1}{1 + Z^2(1 - s_{\text{E}}(x))^2} \\
\implies \hat{\theta}_{\text{M}} &= \hat{K}\hat{\theta}_{\text{O}} + (1 - \hat{K})\hat{\theta}_{\text{E}} \\
&= \hat{\theta}_{\text{O}}\left[\frac{1 - s_{\text{E}}(x)}{1 + Z^2(1 - s_{\text{E}}(x))^2} + s_{\text{E}}(x)\right] \\
&= \hat{\theta}_{\text{O}}\left[\frac{1 + Z^2 s_{\text{E}}(x)(1 - s_{\text{E}}(x))^2}{1 + Z^2(1 - s_{\text{E}}(x))^2}\right] \\
&\equiv \hat{\theta}_{\text{O}}\, s_{\text{MSE}}(x),
\end{aligned} \tag{3.5}$$

with the last term on the right-hand side s_{MSE} being the $\hat{\theta}_{\text{MSE}}$-specific shrinkage factor, a function of the data only through Z. Figure 3.2 displays both shrinkage factors for various α levels as functions of Z; focus is placed on shrinkage used in practice for significant Z scores, denoted with opaque lines.



(a) $\alpha = 0.05$ (b) $\alpha = 0.01$ (c) $\alpha = 10^{-5}$

Figure 3.2: Shrinkage value $s(Z)$ for a given Z-score across different levels α for $\hat{\theta}_{\text{Mean}}$ (red) and $\hat{\theta}_{\text{MSE}}$ (blue). Ligher lines indicate values of Z where shrinkage is irrelevant because results are not significant, and no estimate would be reported.

While the limit of $s_{\text{MSE}}(Z)$ is 1 as $Z \to 0$, we note that—while visually imperceptible—the resulting W-shape does yield non-monotonic shrinkage as *significant* values of the Z score grow further from the null for values of $\alpha > 0.0307$. However the decreasing values of $s_{\text{MSE}}(Z)$ occur for only a small range of values past significance, so the practical impact seems minimal.

To formally show the motivation for monotonicity of any shrinkage with respect to Z, we consider tests yielding varying Z-scores, for which the resulting p-value is "just-significant" (i.e. the p-value is α or some $\alpha - \epsilon$ value just below it). For any fixed θ, lower α results in larger c, and therefore increased selection bias, as proven in Theorem 1 below.

Theorem 1. *For any fixed θ, the absolute bias of $\hat{\theta}$ conditional on significance $|\mathbb{E}[\hat{\theta}\,||Z| \geq c] - \theta|$ increases as the critical value c increases.*

Proof. Due to absolute bias being symmetric about $\theta = 0$, suppose without loss of generality $\theta \geq 0$. Also suppose without loss of generality $\sigma = 1$; if proven, the result holds for any fixed

$\frac{\theta}{\sigma}$. Since bias for any $\theta > 0$ is positive, we need not consider absolute bias. We note that

$$\frac{\partial}{\partial c}\Big[\mathbb{E}[\hat{\theta}||Z| \geq c] - \theta\Big] = \frac{(\theta - c)\phi(\theta - c) + (\theta + c)\phi(-\theta - c)}{\Phi(\theta - c) + \Phi(-\theta - c)} + \frac{[\phi(\theta - c) + \phi(-\theta - c)][\phi(\theta - c) - \phi(-\theta - c)]}{[\Phi(\theta - c) + \Phi(-\theta - c)]^2}.$$

Now, fix function $h_1(c, \theta) = [\Phi(\theta - c) + \Phi(-\theta - c)]\frac{\partial}{\partial c}\Big[\mathbb{E}[\hat{\theta}||Z| \geq c] - \theta\Big]$, such that if $h_1(c, \theta) > 0$, the same holds true for $\frac{\partial}{\partial c}\Big[\mathbb{E}[\hat{\theta}||Z| \geq c] - \theta\Big]$. The function $h_1(c, \theta)$ may be written as

$$h_1(c, \theta) \equiv \frac{\theta}{\phi(\theta - c) - \phi(-\theta - c)} - \frac{c}{\phi(\theta - c) + \phi(-\theta - c)} + \frac{1}{\Phi(\theta - c) + \Phi(-\theta - c)}.$$

Noting that

$$\begin{aligned} h_1(c, \theta) > h_2(c, \theta) &\equiv \frac{\theta}{\phi(\theta - c) + \phi(-\theta - c)} - \frac{c}{\phi(\theta - c) + \phi(-\theta - c)} + \frac{1}{\Phi(\theta - c) + \Phi(-\theta - c)} \\ &= \frac{\theta - c}{\phi(\theta - c) + \phi(-\theta - c)} + \frac{1}{\Phi(\theta - c) + \Phi(-\theta - c)}, \end{aligned}$$

we consider two separate cases: $\theta > c$ and $\theta \leq c$.

First, if $\theta > c$, then clearly $h_2(c, \theta) > 0$. Since $h_1(c, \theta) > h_2(c, \theta) > 0$, and therefore $\frac{\partial}{\partial c}\Big[\mathbb{E}[\hat{\theta}||Z| \geq c] - \theta\Big] > 0$, the proof is complete for this case.

For the second case, i.e. $\theta - c < 0$, $h_2(c, \theta)$ may be rewritten so that all normal densities and distributions are functions of positive elements:

$$h_2(c, \theta) = \frac{\theta - c}{\phi(c - \theta) + \phi(c + \theta)} + \frac{1}{[1 - \Phi(c - \theta)] + [1 - \Phi(\theta + c)]}.$$

Gordon [1941] proved that for any $x > 0$,

$$\frac{1 - \Phi(x)}{\phi(x)} < \frac{1}{x} \implies \phi(x) > x(1 - \Phi(x)),$$

which implies $h_2(c,\theta) > h_3(c,\theta)$ where

$$h_3(c,\theta) \equiv \frac{\theta - c}{(c-\theta)[1-\Phi(c-\theta)] + (c+\theta)[1-\Phi(c+\theta)]} + \frac{1}{[1-\Phi(c-\theta)] + [1-\Phi(c-\theta)]}.$$

To determine if $h_3(c,\theta) > 0$, initially suppose $h_3(c,\theta) < 0$ for some pair (c,θ). Then,

$$\begin{aligned} \frac{\theta - c}{(c-\theta)[1-\Phi(c-\theta)] + (c+\theta)[1-\Phi(c+\theta)]} &< \frac{-1}{[1-\Phi(c-\theta)] + [1-\Phi(c-\theta)]} \\ \implies \Phi(c+\theta) &> 1. \end{aligned}$$

The final contradiction proves $h_3(c,\theta) > 0 \ \forall (c,\theta)$. Since we have $h_1(c,\theta) > h_2(c,\theta) > h_3(c,\theta) > 0$ when $c - \theta > 0$, the proof that $\frac{\partial}{\partial c}\left[\mathbb{E}[\hat{\theta}||Z| \geq c] - \theta\right] > 0$ is complete. □

With this proven monotonicity, it is reasonable to induce similar monotonicity in shrinkage as Z grows further from the null with fixed significance threshold c is reasonable. For this reason, we only consider use of $s_{\text{MSE}}(Z)$ where its shrinkage is monotonic in Z, which in turn means only considering $\alpha < 0.0307$.

3.3 Unifying Decision-Theoretic Approach to Shrinkage

As noted in Section 3.2, both of Zhong and Prentice [2008]'s estimators ameliorate the Winner's Curse by shrinking point estimates towards the null by an amount determined by the data. The amount of shrinkage for these estimators and many others depends on data only through the absolute Z-score. A family of loss functions yielding Bayes rules with data-adaptive shrinkage depending on data only through the Bayesian analog of the Z-score will be developed below, through intuitive loss function concatenation and weighting. We then develop a method reverse-engineering a loss function in this family that justifies a specific existing shrinkage estimator. This methodology will then be employed to motivate Zhong and Prentice [2008]'s shrinkage estimators, to compare behavior of these estimators with others, and in Section 3.4, to show how estimators from Zhong and Prentice [2008] may be used in a Bayesian paradigm.

3.3.1 Defining the Family of Loss Functions

Motivation for a family of loss functions yielding data-adaptive shrinkage Bayes rules (defined by the Z-score) uses multiple concepts. We first introduce a simpler "all-or-nothing" shrinkage, much like the interpretation of hypothesis test results. Let indicator $h = 1$ denote an active decision about parameter θ of dimension p based on data, $h = 0$ otherwise. For each $h \in \{0, 1\}$, a decision d must be determined; to do so, we consider the following loss function;

$$L(d, h, \Sigma) = \log|\Sigma| + \begin{cases} (\theta - d)^T \Sigma^{-1} (\theta - d) + \gamma & \text{if } h = 1 \\ (\theta_0 - d)^T \Sigma^{-1} (\theta_0 - d) + (\theta - \theta_0)^T \Sigma^{-1} (\theta - \theta_0) & \text{if } h = 0 \end{cases}. \tag{3.6}$$

With decision $h = 1$—what we shall call an *active* decision—we use squared error loss, weighted by positive definite matrix Σ. Each element of Σ^{-1} dictates the trade-off between losses for individual components of θ by adding $(\Sigma^{-1})_{ij}(\theta_i - d_i)(\theta_j - d_j)$ to the loss. When $i = j$, we see there is penalty for lack of *accuracy*. With regard to decision Σ, we note that if Σ^{-1} has determinant near 0, then some projection of $\theta - d \neq 0$ must be essentially unpenalized. Hence, proximity of determinant of Σ^{-1} to 0 is penalized via $\log|\Sigma|$. Additional loss γ is added to making active decisions with $h = 1$ "expensive".

Bayes rules for d, Σ, and h are stated and proven in Theorem 2 below. The Bayes rule for d will depend on data only through the Z-score, in-keeping with the final loss function family of interest (to justify estimators due to Zhong and Prentice [2008]). With flat priors approximating frequentist analysis, the decision may be based on a typical Z-score. However, with an informative prior for θ, shrinkage is based on a natural Bayesian analog to the Z-score, $Z^2 = \mathbb{E}[\theta - \theta_0]^T \mathrm{Var}^{-1}(\theta) \mathbb{E}[\theta - \theta_0]$. Notation Z will be used to denote this Bayesian analog throughout, noting the typical Z-score may be approximated using the Bayesian analog with flat priors.

Theorem 2. *For loss in Equation* (3.6)*, Bayes rules for* d*,* h*, and* Σ *are*

Condition	Decision		
	h	d	Σ
$Z^2 > e^\gamma - 1$	1	$\mathbb{E}[\,\theta\,]$	$Var[\,\theta\,]$
$Z^2 < e^\gamma - 1$	0	θ_0	$Var[\,\theta\,] + \mathbb{E}[\,\theta - \theta_0\,]\mathbb{E}[\,\theta - \theta_0\,]^T$

Proof. Consider minimization of $\mathbb{E}[L(d, h, \Sigma)|Y] = \mathbb{E}[L|Y]$ for arbitrary observed data set Y. For $h = 1$, the expected value may be written as

$$\begin{aligned} \mathbb{E}[\,L(\cdot, h = 1) \mid Y\,] &= \log|\Sigma| + \mathbb{E}[\,(\theta - d)^T\Sigma^{-1}(\theta - d)\,] + \gamma \\ &= \log|\Sigma| + tr(\Sigma^{-1}\mathbb{E}[\,(\theta - d)(\theta - d)^T\,]) + \gamma. \end{aligned}$$

Set $\Sigma = \mathbb{E}[\,(\theta - d)(\theta - d)^T\,]A^{-1}$ for positive definite matrix A. Then,

$$\mathbb{E}[\,L(\cdot, h = 1) \mid Y\,] = \log\left|\mathbb{E}[\,(\theta - d)(\theta - d)^T\,]\right| - \log|A| + tr(A) + \gamma.$$

For eigenvalues $\lambda_1, ..., \lambda_K$ of matrix A,

$$-\log|A| + tr(A) = \sum_{i=1}^{n}[-\log(\lambda_i) + \lambda_i].$$

The sum may be minimized with respect to each λ_i by setting first derivative $-\frac{1}{\lambda_i} + 1 = 0 \implies \lambda_i = 1$; with positive second derivative, $A = I_{p\times p}$ minimizes the sum for identity matrix I. Using this value of A,

$$\mathbb{E}[\,L(\cdot, h = 1) \mid Y\,] = \log\left|\mathbb{E}[\,(\theta - d)(\theta - d)^T\,]\right| + p + \gamma.$$

To minimize with respect to d, it remains to consider

$$\begin{aligned}\log\left|\mathbb{E}[\,(\theta-d)(\theta-d)^T\,]\right|+p+\gamma &= \log\left|\mathbb{E}[\,(\theta-\mathbb{E}[\,\theta\,]+\mathbb{E}[\,\theta\,]-d)(\theta-\mathbb{E}[\,\theta\,]+\mathbb{E}[\,\theta\,]-d)^T\,]\right| \\ &\quad +p+\gamma \\ &= \log\left|\mathrm{Var}[\,\theta\,]+(\mathbb{E}[\,\theta\,]-d)(\mathbb{E}[\,\theta\,]-d)^T\,]\right|+p+\gamma,\end{aligned}$$

and by the matrix determinant lemma this is

$$\log|\mathrm{Var}[\,\theta\,]|+\log\left(1+(\mathbb{E}[\,\theta\,]-d)^T\mathrm{Var}[\,\theta\,]^{-1}(\mathbb{E}[\,\theta\,]-d)\right)+p+\gamma,$$

which is minimized by setting $d=\mathbb{E}[\,\theta\,]$ when $h=1$, which in turn means setting $\Sigma=\mathrm{Var}[\,\theta\,]$ when $h=1$. This Bayes rule achieves minimized posterior loss
$\mathbb{E}[\,L(\cdot,h=1)\,]=\log|\mathrm{Var}[\,\theta\,]|+p+\gamma$.

When $h=0$,

$$\begin{aligned}\mathbb{E}[\,L(\cdot,h=0)|Y\,] &= \log|\Sigma|+\mathbb{E}[(\theta_0-d)^T\Sigma^{-1}(\theta_0-d)+(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0)] \\ &= \log|\Sigma|+tr(\Sigma^{-1}\mathbb{E}[\,(\theta_0-d)(\theta_0-d)^T\,])+tr(\Sigma^{-1}\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T\,]) \\ &= \log|\Sigma|+tr(\Sigma^{-1}\{\mathbb{E}[\,(\theta_0-d)(\theta_0-d)^T\,]+\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]\}).\end{aligned}$$

Now suppose $\Sigma=(\mathbb{E}[\,(\theta_0-d)(\theta_0-d)^T\,]+\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T])A^{-1}$ for positive definite matrix A. Then,

$$\mathbb{E}[\,L(\cdot,h=0)|Y\,]=\log|\mathbb{E}[\,(\theta_0-d)(\theta_0-d)^T\,]+\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]|-\log|A|+tr(A).$$

By the same reasoning used above, $A=I_{p\times p}$ for identity matrix I. Therefore, the expected

value of the loss is

$$\begin{aligned}\mathbb{E}[\, L(\cdot, h=0)|Y\,] &= \log|\mathbb{E}[\,(\theta_0-d)(\theta_0-d)^T\,] + \mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]| + p \\ &= \log\left(\left[1+(\theta_0-d)^T\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T](\theta_0-d)\right]|\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]|\right)+p \\ &= \log\left(1+(\theta_0-d)^T\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T](\theta_0-d)\right)+\log|\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]| \\ &\quad + p,\end{aligned}$$

with the penultimate line due to the matrix determinant lemma. Since $\mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T]$ is positive definite, $d=\theta_0$ when $h=0$. Plugging this decision into the formulation for Σ, we now have $\Sigma = \mathbb{E}[\,(\theta-\theta_0)(\theta-\theta_0)^T] = \mathrm{Var}[\theta] + (\mathbb{E}[\theta]-\theta_0)(\mathbb{E}[\theta]-\theta_0)^T$ when $h=0$.

Hence, the minimized loss when $h=0$ is

$$\log\left|\mathrm{Var}[\theta]+(\mathbb{E}[\theta]-\theta_0)(\mathbb{E}[\theta]-\theta_0)^T\right|+p = \log\left(1+(\mathbb{E}[\theta]-\theta_0)^T\mathrm{Var}^{-1}(\theta)(\mathbb{E}[\theta]-\theta_0)\right)+\log|\mathrm{Var}(\theta)|+p$$

due to the matrix determinant lemma. Recall, the minimized loss when $h=1$ is

$$\log|\mathrm{Var}(\theta)| + p + \gamma.$$

Therefore, decide $h=1$ when

$$\begin{aligned}\log|\mathrm{Var}(\theta)| + p + \gamma &< \log\left(1+(\mathbb{E}[\theta]-\theta_0)^T\mathrm{Var}^{-1}(\theta)(\mathbb{E}[\theta]-\theta_0)\right)+\log|\mathrm{Var}(\theta)|+p, \text{ i.e.} \\ \gamma &< \log\left(1+Z^2\right).\end{aligned}$$

So, choose $h=1$ if and only if $Z^2 > e^\gamma - 1$.

□

If an active decision is made, the Bayes rule for θ is $\mathbb{E}[\theta]$. However, with decision $h=0$—what we shall call an *inactive* decision—the Bayes rule is $d=\theta_0$, which can be viewed as shrinking all the way to the null. In this case, loss depends on how far the truth is away

from this simplified conclusion. This type of loss reflects what we call *embarrassment* about making a 'null' choice for θ, and is particularly costly when the truth is far from the null. The choice between an active and inactive decision is based on data only through Z, much like typical significance tests.

Combining the two terms in Equation (3.6) we can obtain a Bayes rule that balances the losses incurred with active and inactive decisions. Bayes rules from sums of loss functions are often weighted averages of the decisions from each loss being added [Zellner, 1994], with weights dictating the trade-off rate for incurring each type of loss. Using fixed weight/trade-off rate $s \in (0,1)$ the combined loss function may be written as

$$\begin{aligned} L(d,\Sigma) = \quad & s\left(\log|\Sigma| + (\theta - d)^T\Sigma^{-1}(\theta - d)\right) + \\ & (1-s)\left(\log|\Sigma| + (\theta_0 - d)^T\Sigma^{-1}(\theta_0 - d) + (\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0)\right) \\ = \log|\Sigma| + s & \underbrace{(\theta - d)^T\Sigma^{-1}(\theta - d)}_{\text{inaccuracy}} \\ & + (1-s)\underbrace{\left((\theta_0 - d)^T\Sigma^{-1}(\theta_0 - d) + (\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0)\right)}_{\text{embarrassment}}. \end{aligned} \tag{3.7}$$

Trade-off rate s accounts for the relative penalty incurred for making an active decision versus an inactive one; hence γ is now absent. The Bayes rule for $d = s\mathbb{E}[\theta] + (1-s)\theta_0$ is defined as anticipated; with $s \in \{0,1\}$, s dictates how much the estimate using squared error loss should shrink towards the null.

To make the amount of shrinkage a data-adaptive decision—i.e. s as a function of data through Z—we further generalize loss (3.7). Using positive function $f(s)$ to weight terms in the loss, we consider

$$\begin{aligned} L(d,s,\Sigma) = \log|\Sigma| + \frac{f(s)^{1/p}}{(1-s)^{1/p}}\Bigg(& s\underbrace{(\theta - d)^T\Sigma^{-1}(\theta - d)}_{\text{inaccuracy}} \\ & + (1-s)\underbrace{\left((\theta - \theta_0)^T\Sigma^{-1}(\theta - \theta_0) + (d - \theta_0)^T\Sigma^{-1}(d - \theta_0)\right)}_{\text{embarrassment}}\Bigg), \end{aligned} \tag{3.8}$$

where we note that s is now a decision, along with d and Σ. Each trade-off rate s dictates relative loss attributable to inaccuracy and embarrassment, made clear by the individual terms within the parentheses. The overall loss for a given trade-off rate s is scaled by the coefficient $\frac{f(s)^{1/p}}{(1-s)^{1/p}}$. Hence, the function $f(s)$ dictates the cost of each trade-off rate $s \in (0,1)$, and therefore how much each trade-off rate is valued.

As proven in Theorem 3 below, the Bayes rule for d remains unchanged, setting $d = s\mathbb{E}[\theta] + (1-s)\theta_0$. As previously mentioned, the extent of the shrinkage s is data-adaptive depending on the data through Z^2 alone; the functional form of this dependence is determined by the chosen weighting function $f(s)$.

Theorem 3. *The Bayes rule* (d, Σ, s) *for the loss function in Equation* (3.8) *is*

$$\begin{aligned} d &= s\mathbb{E}[\theta] + (1-s)\theta_0, \\ \Sigma &= \frac{f(s)^{1/p}}{(1-s)^{1/p}} \left(Var(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T \right), \\ s &= \underset{s}{\operatorname{argmin}} \frac{f(s)}{(1-s)} \left(1 + (1-s^2)Z^2\right). \end{aligned}$$

Proof. Consider minimization of $\mathbb{E}[L(d,s,\Sigma)|Y] = \mathbb{E}[L|Y]$ for arbitrary observed data set Y. The expected value may be written as

$$\begin{aligned} \mathbb{E}[L|Y] = \log|\Sigma| + \frac{f(s)^{1/p}}{(1-s)^{1/p}} \Bigg(& \operatorname{trace}\{s\Sigma^{-1}\mathbb{E}[(\theta-d)(\theta-d)^T]\} \\ & + \operatorname{trace}\{(1-s)\Sigma^{-1}\mathbb{E}[(\theta-\theta_0)(\theta-\theta_0)^T]\} + \operatorname{trace}\{(1-s)\Sigma^{-1}\mathbb{E}[(d-\theta_0)(d-\theta_0)^T]\} \Bigg). \end{aligned}$$

Let

$$\Sigma = \frac{f(s)^{1/p}}{(1-s)^{1/p}}\Bigg(s\mathbb{E}[(\theta-d)(\theta-d)^T] + (1-s)\mathbb{E}[(\theta-\theta_0)(\theta-\theta_0)^T] + (1-s)\mathbb{E}[(d-\theta_0)(d-\theta_0)^T]\Bigg)A^{-1}$$

for positive definite matrix A. Then,

$$\mathbb{E}[L|Y] = \log\left|\frac{f(s)^{1/p}}{(1-s)^{1/p}}\Bigg(s\mathbb{E}[(\theta-d)(\theta-d)^T] + (1-s)\mathbb{E}[(\theta-\theta_0)(\theta-\theta_0)^T] + (1-s)\mathbb{E}[(d-\theta_0)(d-\theta_0)^T]\Bigg)\right| - \log|A| + \text{trace}(A).$$

For eigenvalues $\lambda_1, ..., \lambda_K$ of matrix A,

$$-\log|A| + \text{trace}(A) = \sum_{i=1}^{n}[-\log(\lambda_i) + \lambda_i].$$

The sum may be minimized with respect to each λ_i by setting first derivative $-\frac{1}{\lambda_i} + 1 = 0 \implies \lambda_i = 1$; with positive second derivative, $A = I_{p\times p}$ minimizes the sum for identity

matrix I. Using this value of A,

$$\begin{aligned}
\mathbb{E}[L|Y] &= \log\left(\frac{f(s)}{1-s}\right) + p + \log\Big| s\mathbb{E}[(\theta-d)(\theta-d)^T] \\
&\qquad + (1-s)\mathbb{E}[(\theta-\theta_0)(\theta-\theta_0)^T] + (1-s)\mathbb{E}[(d-\theta_0)(d-\theta_0)^T]\Big| \\
&= \log\left(\frac{f(s)}{1-s}\right) + p + \log\Big| \left(d - s\mathbb{E}[\theta] - (1-s)\theta_0\right)\left(d - s\mathbb{E}[\theta] - (1-s)\theta_0\right)^T \\
&\qquad + \mathrm{Var}(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T\Big| \\
&= \log\left(\frac{f(s)}{1-s}\right) + p \\
&\quad + \log\left(1 + (d - s\mathbb{E}[\theta] - (1-s)\theta_0)^T\Delta^{-1}(d - s\mathbb{E}[\theta] - (1-s)\theta_0)\right) + \log|\Delta|\,,
\end{aligned}$$

where $\Delta = \mathrm{Var}(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T$ and the last step uses the matrix determinant lemma. Since $\left[\mathrm{Var}(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T\right]^{-1}$ is positive definite, $\mathbb{E}[L|Y]$ is minimized with respect to d when $d = s\mathbb{E}[\theta] + (1-s)\theta_0$. With this decision,
$\Sigma = \frac{f(s)^{1/p}}{(1-s)^{1/p}}\left(\mathrm{Var}(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T\right)$ and

$$\begin{aligned}
\mathbb{E}[L|Y] &= \log\left(\frac{f(s)}{1-s}\right) + p + \log\left|\mathrm{Var}(\theta) + (1-s^2)\mathbb{E}[\theta-\theta_0]\mathbb{E}[\theta-\theta_0]^T\right| \\
&= \log\left(\frac{f(s)}{1-s}\right) + p + \log\left(1 + (1-s^2)\mathbb{E}[\theta-\theta_0]^T\mathrm{Var}^{-1}(\theta)\mathbb{E}[\theta-\theta_0]\right) + \log|\mathrm{Var}(\theta)|
\end{aligned}$$

and therefore the decision $s = s(Z^2)$ minimizes

$$\begin{aligned}
&\frac{f(s)}{1-s}\left(1 + (1-s^2)\mathbb{E}[\theta-\theta_0]^T\mathrm{Var}^{-1}(\theta)\mathbb{E}[\theta-\theta_0]\right) \\
&= \frac{f(s)}{1-s}\left(1 + (1-s^2)Z^2\right),
\end{aligned}$$

using the Bayesian analog of the Wald statistic Z. □

3.3.2 Justifying Shrinkage Estimators Using the Loss Function Family

Finding a function $f(s)$ for which the Bayes rule of the loss function in Equation (3.8) provides a Bayesian analog of an existing estimator can be non-trivial, especially when the relevant $s(Z^2)$ has no analytic solution. However, using some elementary properties of differential equations, this derivation is simple if the target shrinkage function is a one-to-one function of Z^2, as is typical (e.g. the estimators of Stein [1956]–via Baranchik [1970]–and Thompson [1968]).

Recall the Bayes rule for the functional form of $s(Z^2)$ minimizes

$$g(s) \equiv f(s)\left(\frac{1}{1-s} + (1+s)Z^2\right), \tag{3.9}$$

with minimization of Equation (3.9) achieved when $\frac{\partial g(s)}{\partial s} = 0$. Manipulating the resulting partial derivative at the assumed equality yields

$$\frac{\partial f(s)}{\partial s} = f(s)\frac{-[1+(1-s)^2Z^2]}{1-s+(1-s)^2(1+s)Z^2}. \tag{3.10}$$

If $s(Z^2)$ is a 1:1 function of Z^2, then a unique solution to $f(s)$ may be found using the following two steps:

(1) Find $\lambda(s) \equiv s^{-1}(Z^2)$ and replace Z^2 in Equation (3.10):

$$\frac{\partial f(s)}{\partial s} = f(s)\frac{-[1+(1-s)^2\lambda(s)]}{1-s+(1-s)^2(1+s)\lambda(s)} \tag{3.11}$$

(2) Finally, a *unique* solution to $f(s)$ (up to a constant coefficient, which does not impact Bayes rules in Equation (3.8)) may be found, using the result of Ahmad and Ambrosetti

[2014, pg 7]:

$$f(s) = c \exp\left[-\int_s^{s_{\max}} q(w)dw\right], \text{ where} \tag{3.12}$$
$$q(w) \equiv \frac{-[1 + (1-w)^2\lambda(w)]}{1 - w + (1-w)^2(1+w)\lambda(w)}.$$

The solution to $f(s)$ should be appropriately scaled for computational ease, but as previously mentioned, this scaling does not impact the Bayes rule for $s(Z^2)$. Hence, if the antiderivative for $q(w)$ is available, the investigator need only evaluate the antiderivative at s. Otherwise, numerical integration may be used either as written (where $s_{\max} < 1$ is the maximum value of s for which the value of $f(s)$ will be evaluated) or by switching the bounds of integration, changing the sign of the integral, and having a lower bound equal to minimum possible value of s for a given shrinkage estimator (e.g. for the mean-based shrinkage estimator of Zhong and Prentice [2008], $\lim_{Z \to 0} s_{\text{Mean}}$).

This formulation does not guarantee a *minimum* to $g(s)$ is achieved with respect to s for the derived $f(s)$ when $s(Z^2)$ takes the target formulation, and this must be checked before proceeding. Given that the optimization problem for s is one-dimensional, we have found visual empirical checks sufficient for this purpose, and particularly helpful when $f(s)$ is not analytically tractable. An advantage to following this procedure is that if the aforementioned solution to $f(s)$ does not yield a minimum (with respect to s) to $g(s)$ when evaluated at the target $s(Z^2)$, since Equation (3.12) is the unique solution to Equation 3.11, no suitable solution to $f(s)$ exists, and justification for the target $s(Z^2)$ using the loss function in Equation (3.8) is known to be impossible. This result might be helpful to investigators interested in motivating another shrinkage estimator other than those considered here with loss (3.8). Furthermore—and essential to the application of shrinkage estimators due to Zhong and Prentice [2008]—this formulation enables numerical approximation to values of $f(s)$ when an analytic form of $\lambda(s)$ is unavailable.

3.4 Zhong and Prentice [2008]'s Estimators as Bayes Rules

Shrinkage estimator s_{Mean} is always a one-to-one function of Z^2 as illustrated in Figure 3.2, albeit one without an analytic form. As discussed in Section 3.2, shrinkage estimator s_{MSE} is a one-to-one function of Z^2 for significant results only if $\alpha < 0.0307$, with motivation for monotonicity discussed in the same section. We therefore only consider values of $\alpha < 0.0307$, and replace the default $\alpha = 0.05$ with $\alpha = 0.03$. Under these conditions, the method of Section 3.3.2 may be used to find losses for which the corresponding Bayes rules are Bayesian analogs of s_{Mean} and s_{MSE}. As no analytic form of either shrinkage estimator exists, we use numerical approximations to calculate the relevant $f(s_{\text{Mean}})$ and $f(s_{\text{MSE}})$.

For comparison, we also justify the estimators of Stein [1956] and Thompson [1968] using the same family of loss functions. As discussed in Section 3.1, justification for these shrinkage estimators using the same family of loss functions illustrates similar behavior of these estimators (i.e. shrinking point estimates towards a null using inaccuracy/embarrassment trade-offs), motivating comparison of Winner's Curse bias amelioration. Also note estimators of Stein [1956] and Thompson [1968] are analytically much simpler than those due to Zhong and Prentice [2008]; hence, we may question if the complexity is "worth it". The former estimators are similar to each other in that they were both originally created to reduce mean squared error, translating to (sometimes composite) quadratic loss in decision-theoretic nomenclature. Rao [1980] warned quadratic loss yields estimators that are engineered to shrink extreme estimates that may rarely occur, so that optimality is not assured for moderate estimates, thereby possibly leading to more absolute bias than default estimators. This emphasis works to our favor when only significant results are reported; we wish to curtail very extreme estimates, with no need to consider performance of small or moderate estimates. As Rao [1980] recommends, we do carefully choose tuning parameters for these estimators (as described below), and explicitly assess absolute bias (of significant results) before recommending their use. Recall neither Stein [1956]'s nor Thompson [1968]'s estimators were originally intended for application to significant results only, and we therefore believe use of

these estimators for amelioration of the Winner's Curse is uniquely considered here.

Table C.1 in Appendix C.1 gives explicit statements of these estimates' shrinkage of point estimates, and weighting functions $f(s)$ defining the loss functions for which they are Bayes rules. The tuning parameters in Stein [1956]'s and Thompson [1968]'s estimators have been chosen so that bias of these estimators approximates bias of the s_{MLE} estimator near the null (to be discussed further below). However, under restrictions defined in Table C.1, parameters may instead be fixed so that these estimators' biases equal those of s_{Mean} or s_{MLE} for a single specified θ value. While Thompson [1968]'s estimator is a smooth function of Z, our analog of Stein [1956]'s estimator shrinks any estimate with sufficiently-small Z all the way to the null, and hence is not smooth. However, with our focus on significant results, the choice of tuning parameters to approximate the bias-correction of s_{MLE} and s_{Mean} precludes having to consider this. For visual simplicity, we limit figures to a single value of tuning parameters, but discuss how behavior changes with altered tuning parameters later.

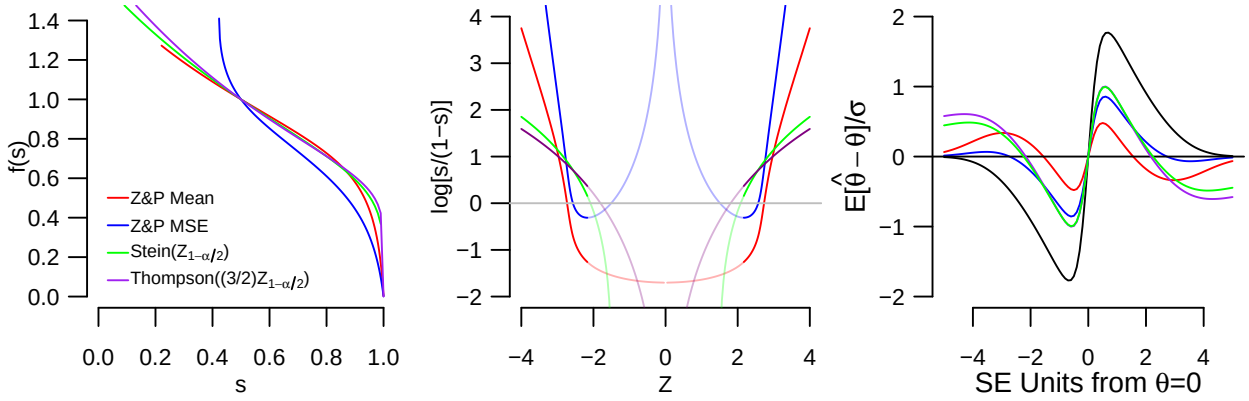
We compare the performance of the more-classical shrinkage estimators to those of Zhong and Prentice [2008] using three illustrations: (1) justification using the family of loss functions allows for comparison of inherent valuation of each trade-off rate s via $f(s)$ for each estimator, as well as (2) the relative cost of embarrassment and inaccuracy for a given data set; finally, we may (3) compare how well the Winner's Curse bias is reduced across all four estimators.

As a function of $s(\cdot)$, the form of corresponding $f(s)$ does not depend on the prior assigned to θ. Hence, the solution is functionally identical for informative and flat priors, the latter providing a motivation for frequentist methods. Panel (a) of Figure 3.3 displays functions $f(s)$ scaled so that $f(0.5) = 1$ for s_{Mean} and s_{MSE} when $\alpha = 0.03$, as well as Stein [1956]'s and Thompson [1968]'s shrinkage estimators with identical scaling. The range over which $f(s)$ is defined is the range of shrinkage values s for each estimator, and hence is not identical across estimators. Across values of s in the ranges of all shrinkage estimators, the value of $f(s)$ is quite similar. The choices of parameters defining Stein [1956]'s and Thompson [1968]'s estimators therefore give similar weight to individual trade-off rates s to Zhong and Prentice [2008]'s estimator.

To further visualize the inherent trade-offs made with each aforementioned shrinkage estimator, we may plot the relative influence of accuracy and embarrassment, or $s/(1-s)$, on the loss in Equation (3.8). While this ratio may be plotted for any shrinkage estimator, justification for the shrinkage estimator by Equation (3.8) gives the new specific interpretation of the ratio as the trade-off between inaccuracy and embarrassment. This ratio is equivalent to the number of units of embarrassment one is willing to trade for each unit of inaccuracy.

Panel (b) of Figure 3.3 displays these trade-off rates for the same $\alpha = 0.03$ (on the log scale to show multiplicative change). Also, with s restricted between 0 and 1, a direct trade-off occurs when $s = 0.50$, meaning reports of the estimated effect size are half the original point estimate. Overall, loss of accuracy is extremely expensive compared to embarrassment for most significant Z-scores, and is always the case across all significant Z-scores for the displayed estimators from Stein [1956] and Thompson [1968]. For a small range of Z-scores just past significance, embarrassment is of greater concern than accuracy for s_{MSE}. This concern is exacerbated across nearly the same range of Z-scores with s_{Mean}, i.e. for just-significant results, s_{Mean} is the shrinkage estimator most concerned with embarrassment. It stands to reason the relative concern with embarrassment and inaccuracy should be monotonic in extremity of the data, as we see with all shrinkage estimators here. Hence, there is stronger motivation for s being monotonic in Z, inducing the same monotonicity in the plotted ratio.

The aforementioned particularly high concern of s_{Mean} with embarrassment for just-significant results corroborates with optimal bias-correction for θ near the null with s_{Mean} compared to other shrinkage estimators considered here (supposing $\theta_0 = 0$ without loss of generality), displayed in panel (c) of Figure 3.3. Here, we suppose priors are flat, motivating frequentist inference (use of informative priors follows in Section 3.4.1). Despite similar bias correction of s_{MSE} and both Stein [1956]'s and Thompson [1968]'s estimators near the null, these estimators have far worse behavior in the tails—even worse than the original uncorrected estimate—though the bias of all estimates eventually vanishes for extreme-enough θ (not plotted). Also note that the choice of tuning parameters to define both Stein [1956]'s and Thompson [1968]'s estimators is a trade-off: higher parameter values lead to better bias-



(a) $f(s)$ (b) $\log\left(\frac{s}{1-s}\right)$ (c) Bias

Figure 3.3: $f(s)$ (scaled so that $f(0.5) = 1$), $\log[s/(1-s)]$, and bias of shrunken estimates using s_{Mean}, s_{MSE}, and Stein($a = Z_{1-\alpha/2}$) and Thompson($k = (3/2)Z_{1-\alpha/2}$) shrinkage estimates with $\alpha = 0.03$. Panel (c) displays bias in estimating normal location θ, where sample mean $\bar{X} \sim N(\theta, \sigma^2)$ and θ is not assigned a prior distribution; the black curve displays bias of uncorrected estimates.

correction near the null, but even worse in the tails. As such, keeping behavior near the null fixed, superior bias-correction in the tails is a strength of Zhong and Prentice [2008]'s estimators. Therefore, the analytically-complex form of Zhong and Prentice [2008]'s estimators appears worthwhile.

3.4.1 Interaction of Prior and Loss Functions

All comparisons of bias in Section 3.4 assume use of flat priors/frequentist inference with normal data. Naturally, with the justification for shrinkage functions using decision theory, influence of prior distribution specification on behavior of the shrinkage estimator should be explored. With prior specification, with the use of a Bayesian Wald statistic to define the shrinkage, neither panels (a) nor (b) of Figure 3.3 will change. However, lingering bias after shrinkage as a function of the true value of θ is subject to change with altering prior specification, a topic further explored here. We begin by considering how well the choice

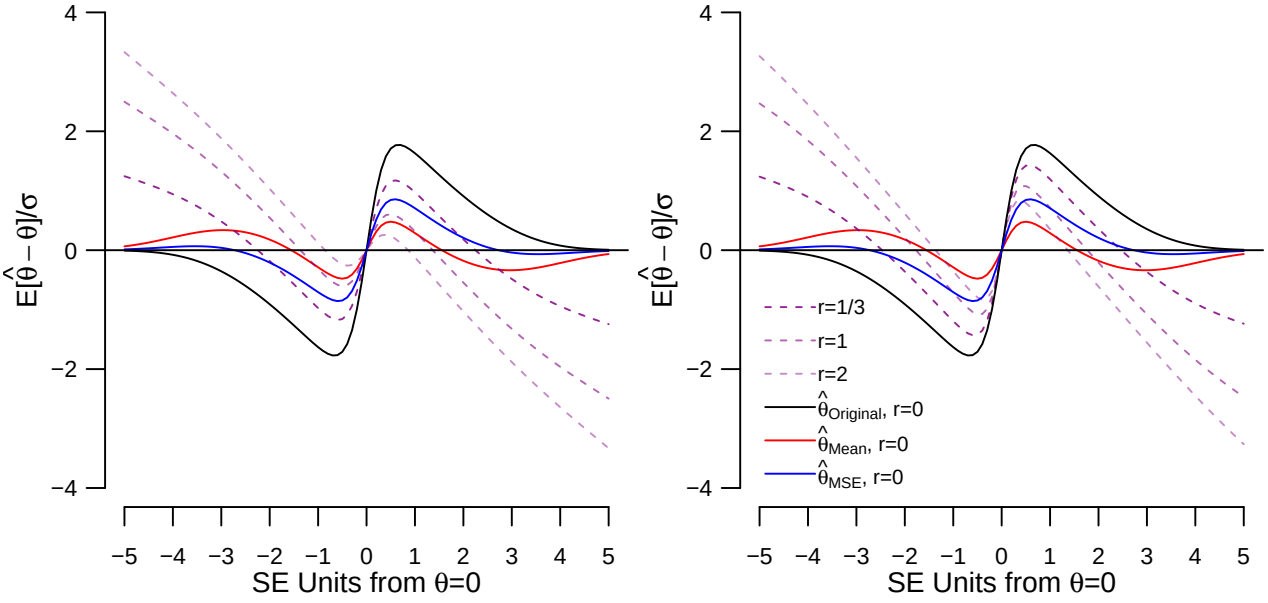
of prior distribution alone (without using a shrinkage function) corrects for selection bias; poor performance will motivate use of a shrinkage estimator in addition to prior specification. Later, we consider the interaction on shrinkage between assigning a prior and using a shrinkage-inducing loss function.

Using simple squared-error loss—as in Equation (3.6)—the Bayes rule for the estimate of θ is the posterior mean $\mathbb{E}[\theta|X] = \mathbb{E}[\theta]$; this estimate can also be motivated informally as just a reasonable summary of the posterior. In many cases, the posterior mean is a shrunken version of a point estimate, landing between a prior mean and an estimate acquired from data alone. For example, suppose $\bar{X} = \hat{\theta} \sim N(\theta, \sigma^2)$ and θ has prior $N(\theta_0, v^2)$. The posterior mean is a weighted average of the prior mean and sample mean $\hat{\theta}$;

$$\mathbb{E}[\theta|X] = \hat{\theta}\left[\frac{1/\sigma^2}{1/\sigma^2 + 1/v^2}\right] + \theta_0\left[\frac{1/v^2}{1/\sigma^2 + 1/v^2}\right].$$

Fixing the prior mean at the null θ_0 ensures shrinking of $\hat{\theta}$ towards the null. Assuming for simplicity $\theta_0 = 0$, we see shrinkage of $\hat{\theta}$ towards the null by a factor of $\left[\frac{1/\sigma^2}{1/\sigma^2+1/v}\right]$ is not dictated by the Z-score, but rather through specification of the prior and data variance. Note that the shrinkage estimator is defined by these two quantities alone and therefore, for a given design and prior, any significant estimate $\hat{\theta}$ is shrunk an identical amount.

To see how well prior assignment ameliorates the Winner's Curse bias, we allow v^2 to vary, supposing σ^2 is fixed at any value. For the sake of generality, we plot the bias for a variety of sample sizes r of the prior *relative* to the data in Figure 3.4. For example, if $\sigma^2 = 1$ and $\nu^2 = 0.5$, the prior is twice as informative as the data on the posterior, and hence the prior sample size relative to the data is 2. Panel (a) assumes results deemed significant using *frequentist inference* are reported, while panel (b) assumes Bayesian inference is used (i.e. the value of the Bayesian Wald statistic analog relative to the same critical value used in frequentist inference). Note the Bayesian analog of the Wald statistic shrinks the standard frequentist Wald statistic by a factor of $\left(\frac{1}{r+1}\right)^{1/2}$; therefore, Bayesian inference requires more extreme data to reject the null, with increasing extremity required as r grows (i.e. the prior



(a) Frequentist inference

(b) Bayesian inference

Figure 3.4: Bias with shrinkage due to prior assignment alone (purple) with a variety of prior relative sample sizes r vs. a flat prior and use of s_{Mean} (red), s_{MSE} (blue), or reporting the original estimate (black) with $\alpha = 0.03$, testing $H_0 : \theta = 0$ for normal location θ where $\hat{\theta} \sim N(\theta, \sigma^2)$ and $\theta \sim N(\theta_0 = 0, \sigma^2/r)$. Estimates are limited to those from significant tests using frequentist inference in panel (a), and using Bayesian inference in panel (b).

becomes more informative relate to the data). In either case, these figures show how prior assignment alone is not effective at ameliorating the Winner's Curse, compared to using a flat prior and either s_{Mean} or s_{MSE}. The introduction of a steadily-increasing bias as θ grows further from 0 is particularly problematic, far exceeding the bias incurred from the original point estimate.

Rejecting use of the prior alone to correct for Winner's Curse, we now consider how assignment of a prior distribution affects bias-reducing capabilities of our Bayesian analogs of the estimators from Zhong and Prentice [2008].

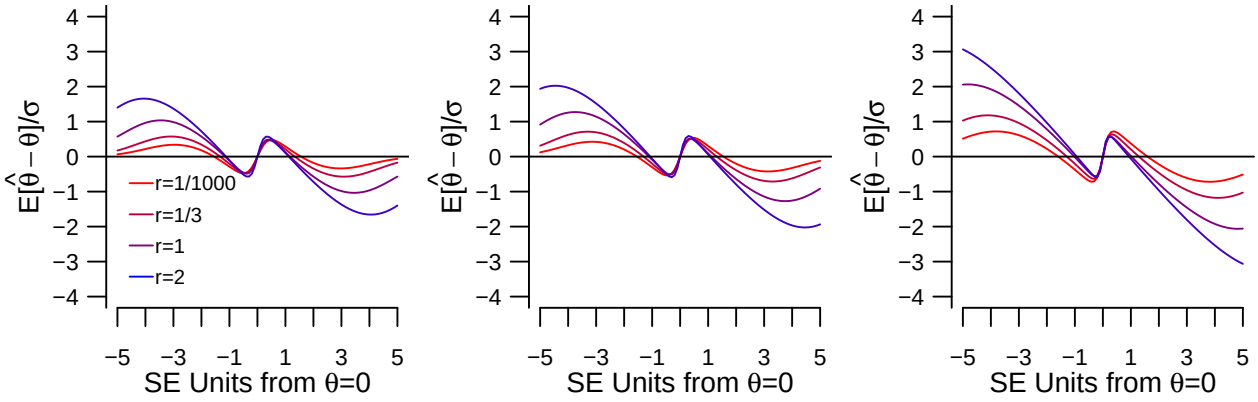
In what follows we use Bayesian inference (using the Bayesian analog of the Wald statistic) throughout. In particular, the Bayesian analogs of the shrinkage estimators from Zhong and

Prentice [2008] are now used exclusively to determine the amount of shrinkage, i.e. the arguments of $s_{\text{Mean}}(Z)$ and $s_{\text{MSE}}(Z)$. Recall the Bayesian analog of the Z-score shrinks the standard Wald statistic by a factor of $\left(\frac{1}{r+1}\right)^{1/2}$. Also recall shrinkage becomes more extreme using either shrinkage estimator as Z grows closer to the null. Therefore, assignment of a prior reduces the value of the argument to the shrinkage function, inducing more extreme shrinkage towards the null for a given data set as the prior becomes more informative.

Figure 3.5 shows the effect of this dual shrinkage on bias correction for a variety of prior sample sizes r relative to the data using s_{Mean}; the same is displayed in Figure 3.6 using s_{MSE}. The least-informative prior with $r = 1/1000$ approximates frequentist inference, in which the prior does not contribute. The most informative prior, $r = 2$, yields a prior that gives twice the precision as the available data, meaning the prior mean (zero) receives two-thirds of the weight in the posterior mean.

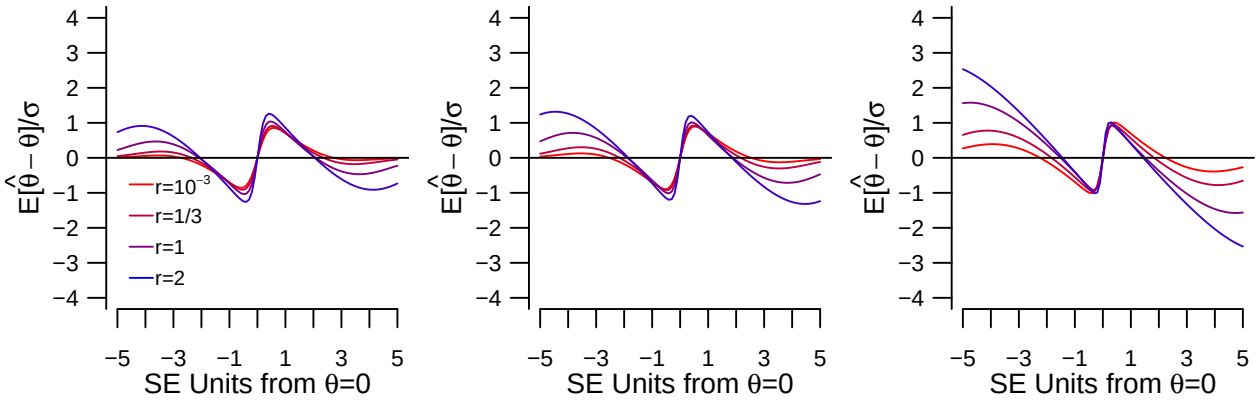
Overall, when shrinkage using either shrinkage estimator with noninformative priors is insufficient, bias is nearly identical across more informative priors. The divergence in bias occurs when the shrinkage estimator with diffuse priors is expected to *over-correct* the estimate. In other words, over-correction becomes worse as the prior becomes more informative. This finding is unsurprising because shrinkage factors approach 0 for any fixed $\hat{\theta}$ as the prior grows increasingly informative due to the dual shrinkage. Bias is more sensitive to the prior precision for smaller α, a noteworthy warning to GWAS investigators wishing to use very small α. Trends are extremely similar for s_{Mean} and s_{MSE}, with the latter generally being less sensitive to changes in the prior distribution. This may be due to, for a given Z-score, there being a smaller range of values of θ with expected over-correction of realized estimates $\hat{\theta}$ and less-extreme shrinkage due to s_{MSE} in general. Across all priors considered, s_{MSE} continues to incur more bias by failing to shrink enough for values of θ near the null.

Based on these results, it seems that while the practice of incorporating prior knowledge about θ is important for inference, the prior should *not* be chosen just to induce shrinkage ameliorating the Winner's Curse. In fact, informative priors centered about the null may exacerbate over-correction from existing shrinkage estimators, particularly s_{Mean}.



(a) $\alpha = 0.03$ (b) $\alpha = 0.01$ (c) $\alpha = 10^{-5}$

Figure 3.5: Bias using s_{Mean} for a variety of relative sample sizes r of the prior to the data. We suppose $\hat{\theta} \sim N(\theta, \sigma^2)$ and $\theta \sim N(\theta_0 = 0, \sigma^2/r)$.



(a) $\alpha = 0.03$ (b) $\alpha = 0.01$ (c) $\alpha = 10^{-5}$

Figure 3.6: Bias using s_{MSE} for a variety of relative sample sizes r of the prior to the data. We suppose $\hat{\theta} \sim N(\theta, \sigma^2)$ and $\theta \sim N(\theta_0 = 0, \sigma^2/r)$.

These results may appear to discourage any use of informative priors centered at the null. However, such priors may convey legitimate skepticism about θ. While their use may induce biases at larger values of θ, as a subjective statement of what the data tell us, they still have value. Furthermore, despite over-correction for more extreme values of θ, the shrunken estimates may still represent meaningful departures from the null (e.g. with $\alpha = 10^{-5}$ and $r = 2$, if the true value of θ is 5σ, $\mathbb{E}[\hat{\theta}_{\text{Mean}}] = 2\sigma$), and thus a potentially nonneglible departure from the null will be reported. Finally, shrinkage is particularly important with small α and values of θ near the null (see Figure 3.1), as in e.g. GWAS studies. The assignment of an informative prior centered at the null when using s_{Mean} or s_{MSE} will therefore not seriously detract from bias correction when small α is used and values of θ are close to the null, as are typical in GWAS studies.

In summary, with our new decision-theoretic justification for both s_{Mean} and s_{MSE}, astute investigators can consider how choice of informative prior for θ can induce bias into estimates for significant findings. They can also understand effectiveness of these shrinkage estimators in mitigating selection bias both with and without the use of informative priors.

3.5 Step-Function Alternative

In Section 3.4, we investigated sacrifices to Winner's Curse amelioration by using simpler estimators than Zhong and Prentice [2008]'s estimators. We'll repeat the process here with a particular class of shrinkage estimators, a subclass of which are Bayes rules to the loss in Equation (3.8). We will begin by assessing estimators in this subclass, followed by shrinkage estimators outside of this subclass.

A particularly simple form of shrinkage estimate—with appealing connections to statistical testing—is one where shrinkage is a step-function, not shrinking at all when $|Z|$ exceeds some threshold, and shrinking by a fixed amount otherwise.

Formally, we consider

$$\hat{\theta}_{\text{Step}}(Z, \hat{\theta}_{\text{Original}}; s, k) = \begin{cases} s\hat{\theta}_{\text{Original}}, & c \leq |Z| \leq k \\ \hat{\theta}_{\text{Original}}, & |Z| > k, \end{cases} \tag{3.13}$$

where c is the critical value of the significance test. We can equivalently define shrinkage function

$$s_{\text{Step}}(Z; s, k) = \begin{cases} s, & c \leq |Z| \leq k \\ 1, & |Z| > k. \end{cases} \tag{3.14}$$

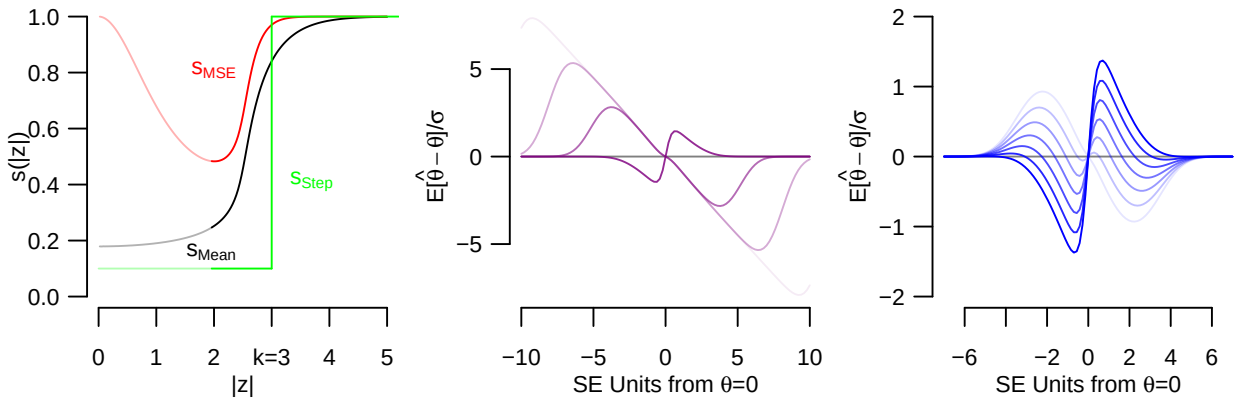
Written in this way, we may also interpret this form of shrinkage as deeming sufficiently-significant results as "trustworthy", with no pressing need for correction. In contrast, a value of s well below 1 may express that, while results are statistically significant at some level, there should be considerable skepticism about the value of θ, estimates of which may be substantially biased.

Figure 3.7(a) gives an example of s_{Step} with $s = 0.10$, $k = 3$, and $\alpha = 0.05$, comparing it to shrinkage functions s_{Mean} and s_{MSE} for the same α-level. Figures 3.7(b) and 3.7(c) illustrate how choices of s and k affect bias-correction across θ values, for fixed values of k and s respectively.

Seeing the impact of different choices of s and k motivates the search for optimal values. We choose to assess different choices by minimax bias, i.e. we seek to minimize the maximum absolute bias across any θ. Formally, the optimal choices are

$$(s_{\text{optim}}, k_{\text{optim}}) = \underset{k \in (c,\infty),\, s \in [a,b]}{\text{argmin}} \left(\max_{\theta \in \mathbb{R}} \mathbb{E}[\, |\hat{\theta}_{\text{Step}}(Z, \hat{\theta}_{\text{Original}}; s, k) - \theta|; \theta\,] \right),$$

where a and b, the bounds on potential s values, are real-valued. For now we limit our search to $s \in [0, 1]$, but will later relax this. Figures 3.8(a) and (b) show two views of the -log(maximum absolute bias) as a function of (s, k) with $\alpha = 0.05$. While a ridge exists,

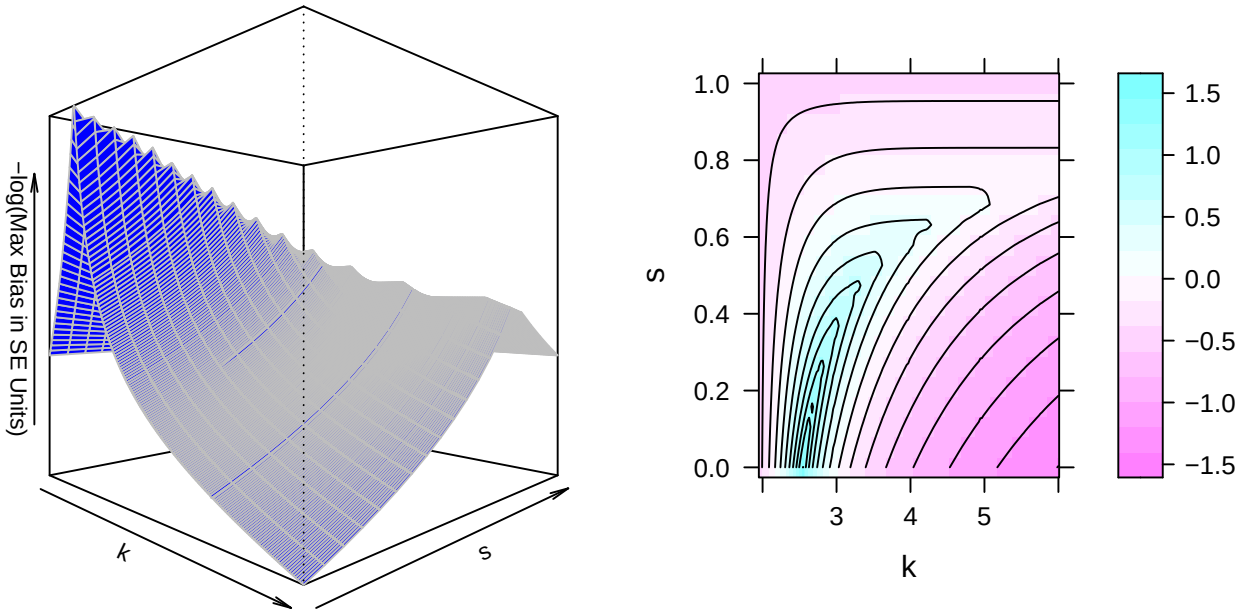


(a) Shrinkage factors (b) Fixed $s = 0.10$ (c) Fixed $k = 3$

Figure 3.7: (a) with $\alpha = 0.05$, example of $s_{\text{Step}}(Z; s, k)$ with $k = 3$ and $s = 0.10$ compared to $s_{\text{Mean}}(Z)$ and $s_{\text{MSE}}(Z)$, where translucent values are not used in practice, because significance has not been reached; in normal location problem, bias of $\hat{\theta}_{\text{Step}}$ with $\alpha = 0.05$ with (b) decreasing values of $k \in \{11, 8, 5, 2\}$ as the lines get darker with fixed $s = 0.10$ and (c) increasing values of $s \in \{0, 0.15, 0.3, 0.45, 0.60, 0.75, 0.90\}$ as the lines get darker with fixed $k = 3$.

along which the maximum bias is almost constant, the optimum choice occurs with $s = 0$ and for a value $k = 2.549$ (equivalent to $\alpha_2 = 0.011$). This result suggests it is optimal to report estimates equal to the null value when the p-value is between 0.011 and 0.05, retaining the original point estimate if p is smaller. Optimality at the boundary value of $s = 0$ (i.e. shrinking point estimates all the way to 0) also holds for other α values. For example, if $\alpha = 0.01$, then $(s_{\text{optim}}, \alpha_{2,\text{optim}}) = (0, 0.002)$. If $\alpha = 10^{-5}$, then $(s_{\text{optim}}, \alpha_{2,\text{optim}}) = (0, 1.79^{-5})$.

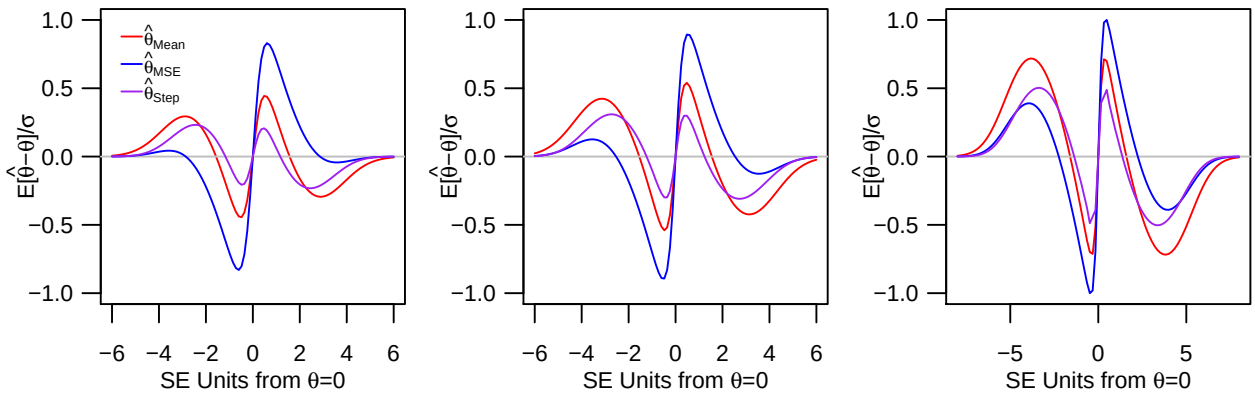
With this optimal shrinkage value of 0, the threshold value k takes on a new meaning. If $|Z| > k$, we can interpret the result as significant enough that any Winner's Curse bias may reasonably be ignored. For significant results not achieving $|Z| > k$, while the test result is significant we may interpret the result as an indication that no point estimate should be attempted; deviation of θ from the null can not be reliably assessed. Note that with $\alpha = 0.05$, the p-value threshold beyond which results are "believable" using these criteria and the interpretation of results is close to Johnson [2013]'s suggested significance threshold of 0.005.



(a) 3-D

(b) Contour

Figure 3.8: Two views of -log(maximum absolute bias)/σ with (s,k) pair with $\alpha = 0.05$ in normal location problem; minimum values plotted for k and s in left-most panel are 0.



(a) $\alpha = 0.05$ (b) $\alpha = 0.01$ (c) $\alpha = 10^{-5}$

Figure 3.9: Bias for normal location problem of $\hat{\theta}_{\text{Mean}}$, $\hat{\theta}_{\text{MSE}}$, and $\hat{\theta}(\cdot; s_{\text{optim}}, k_{\text{optim}})$, with $s_{\text{optim}} \in [0, 1]$.

A further interpretation of the threshold k is that it defines when the exaggeration ratio–expected exaggeration of the estimated effect size given significance [Gelman and Carlin, 2014]–is so minimal, they should not be a concern.

It is natural to question how well these simple "step" shrinkage estimates ameliorate Winner's Curse biases, compared to the estimators defined by Zhong and Prentice [2008]. Using Figure 3.9, we empirically find minor differences, at most. In fact, for many combinations of θ and α the simpler estimator corrects better; granted, the improvement over Zhong and Prentice [2008]'s estimators is minimal at small α, which is arguably the setting where corrections have most impact.

This simpler estimator when $s = 0$ may be justified using the loss function family in Equation (3.8), the same one justifying the Bayesian analog of Zhong and Prentice [2008]'s estimators; see Table C.1. Unsurprisingly, the function $f(s_{\text{Step}})$ when $s = 0$ with the optimal choice of k is nearly identical to $f(s_{\text{Mean}})$ (not pictured), indicating similar valuation given to each trade-off value. Furthermore, justification using this family of loss functions makes this very simple estimator a "reasonable" correction to the default estimator to ameliorate the Winner's Curse due to the trade-off between accuracy and parsimony.

Our final exploration of step-function alternatives to s_{Mean} and s_{MSE} allows s to be negative. We recognize that permitting the signs of estimates to be reversed is not intuitively sensible. Recall this estimator cannot be justified using loss in Equation (3.8), because s_{Step} is no longer restricted to be between 0 and 1. A more rigorous argument notes that with $\alpha = 0.05$, the probability of incorrect sign declaration conditional on significance (known as a *Type S error*) is always below 50% [Gelman and Carlin, 2014], so reversing signs of significant results would, on average, make estimates worse. Yet despite this, extending the axes in Figure 3.8 we see that *better* minimax biases are available if s can, in some situations, take carefully-selected negative values. For example, Figure 3.10 shows how, compared to optimal choices of $s \in [0, 1]$ and k for $\alpha = 0.05$, bias correction is almost universally improved with $s = -4.47$ and $\alpha_2 = 0.04$. This 'better' estimate does not shrink point estimates for which the p-value is less than 0.04, but for those where $0.04 < p < 0.05$, multiplies the estimate by a factor of -4.47. The same figure displays an even more extreme example, where bias is nearly 0 across all θ if all point estimates remain unchanged apart from those yielding p-values between 0.049 and 0.050, which are multiplied by a factor of -58.15.

Taking this trend to its extreme, as $\alpha_2 \to \alpha$, $\mathbb{E}[\hat{\theta}_{\text{Step}} - \theta] \to 0$ as long as $s \to -\infty$ at a specific rate, defined in Theorem 4 below. This means that an exactly unbiased estimate is achieved, in the limit, by leaving alone almost all point estimates but wildly reversing those that are only just significant, i.e. when $p = \alpha - \epsilon$—giving them an unbounded estimate with the opposite sign to that observed.

Theorem 4. $\lim_{k \to c^+} \mathbb{E}[\hat{\theta}_{Step} - \theta] = 0$ *for any fixed* θ *if shrinkage value* $s = s(c, k)$ *is chosen based on the threshold for shrinkage* k *and critical value* c, $\lim_{k \to c^+} s(k, c) = -\infty$ *and* $\frac{\partial}{\partial k} \frac{1}{s(c,k)} = -c$ *for any critical value* c.

Proof. Suppose s takes values in accordance with trends/values assumed true in the statement of the theorem—for example, $s(c, k) = \frac{1}{c(c-k)}$–and that $\sigma = 1$ without loss of generality.

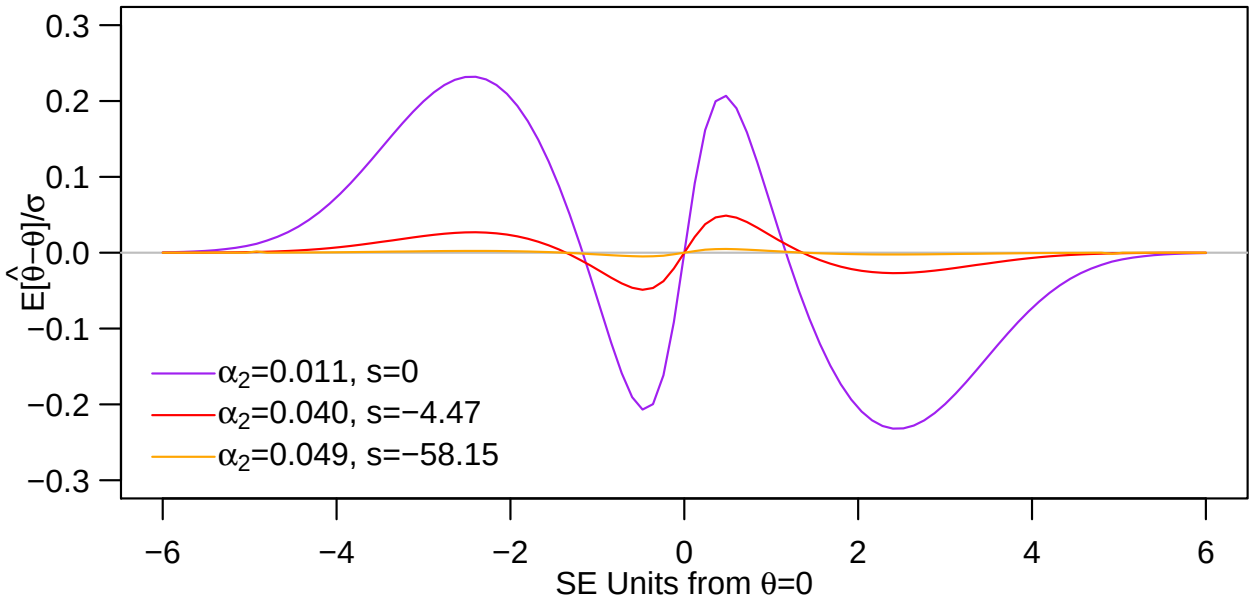


Figure 3.10: With the normal location problem, bias with optimal choice of s and α_2 with non-negative s, compared to bias with optimal choices of s, which may be negative, for given α_2.

Consider the following expression for $\mathbb{E}[s\hat{\theta}_{\text{Step}};\theta] - \theta$:

$$\mathbb{E}[s_{\text{Step}}\hat{\theta}_{\text{Original}};\theta] - \theta = \frac{A + B + C + D - \theta\Omega}{\Omega}, \text{ where}$$

$$A = \int_{-\infty}^{-k} \hat{\theta}\phi(\hat{\theta} - \theta)d\hat{\theta},$$

$$B = \int_{-k}^{-c} s\hat{\theta}\phi(\hat{\theta} - \theta)d\hat{\theta},$$

$$C = \int_{c}^{k} s\hat{\theta}\phi(\hat{\theta} - \theta)d\hat{\theta},$$

$$D = \int_{k}^{\infty} \hat{\theta}\phi(\hat{\theta} - \theta)d\hat{\theta}, \text{ and}$$

$$\Omega = \Phi(\theta - c) + \Phi(-\theta - c).$$

Using simplifications of truncated normal expectations, individual components above may

be rewritten as

$$
\begin{aligned}
A &= \theta\Phi(-k-\theta) - \phi(-k-\theta), \\
B &= s\{\theta[\Phi(-c-\theta) - \Phi(-k-\theta)] + \phi(-k-\theta) - \phi(-c-\theta)\}, \\
C &= s\{\theta[\Phi(k-\theta) - \Phi(c-\theta)] + \phi(c-\theta) - \phi(k-\theta)\}, \\
D &= \theta\Phi(\theta-k) + \phi(k-\theta).
\end{aligned}
$$

Next, limiting behavior of each individual component is considered in turn. Note no limits depend on k approaching c from a specified direction; hence, such specification is removed from the limit. First for B,

$$
\begin{aligned}
\lim_{k\to c} B &= \lim_{k\to c} \frac{\theta[\Phi(-c-\theta) - \Phi(-k-\theta)] + \phi(-k-\theta) - \phi(-c-\theta)}{1/s} \\
&= \frac{0}{0} \\
\implies \lim_{k\to c} B &= \frac{\theta\phi(-c-\theta) - (c+\theta)\phi(-c-\theta)}{-c} \text{ by L'Hopital's rule} \\
&= \phi(-c-\theta).
\end{aligned}
$$

The same procedure with C yields $\lim_{k\to c} C = -\phi(c-\theta)$. Limits of A and D are as follows:

$$
\begin{aligned}
\lim_{k\to c} A &= \theta\Phi(-c-\theta) - \phi(-c-\theta) \\
\lim_{k\to c} D &= \theta\Phi(\theta-c) + \phi(c-\theta).
\end{aligned}
$$

Therefore, the bias in the limit is

$$\begin{aligned}
\lim_{k\to c}\left[\mathbb{E}[s_{\text{Step}}\hat{\theta}_{\text{Original}};\theta]-\theta\right] &= \lim_{k\to c}\frac{A+B+C+D-\theta\Omega}{\Omega} \\
&= \frac{\theta\Phi(-c-\theta)+\theta\Phi(\theta-c)-\theta[\Phi(\theta-c)+\Phi(-\theta-c)]}{\Omega} \\
&\quad + \frac{-\phi(-c-\theta)+\phi(c-\theta)+\phi(-c-\theta)-\phi(c-\theta)}{\Omega} \\
&= \frac{0}{\Omega} = 0.
\end{aligned}$$

□

As indicated above, this estimator is not intended for practical use. However, its existence does serve to show that the goal of removing Winners' Curse biases—the primary goal of Zhong and Prentice [2008] and related literature—is not sufficient to define sensible statistical procedures. This is contrasted with our decision-theoretic approach, which is normative: assuming the relevant loss function states our costs of particular decisions, then adopting the Bayes rule, or something close to it, is essentially the only rational course of action. Defining an estimator without such justification may lead to a completely unreasonable estimator.

3.6 Discussion

Despite prolific claims of its necessity for "good" estimates, we have shown that manipulating a readily-interpretable, simple estimator to be *unbiased* may yield a nonsensical one. However, the assessment of an estimate as "sensible" may not be simple, but is made easier by having some mechanism that helps us intuitively understand how the shrinkage estimator works. Quantifying inherent assumptions made, such as the trade-off made for accuracy in favor of parsimony (or vice versa), is one such mechanism, made possible by defining a loss function in the class defined by Equation (3.8) for which the shrinkage function is the Bayes rule.

Notably, the estimators from Zhong and Prentice [2008]—very effectively designed to

ameliorate the Winner's Curse—are essentially Bayes rules to loss functions in this class. Therefore, this shrinkage estimator now belongs to a class of other estimators, similarly shrinking point estimates towards the null by trading accuracy for parsimony. It is therefore newly worthwhile to investigate how these other estimators account for selection bias, despite motivations for their creation far from such a purpose. While more-classical estimators had worse performance than Zhong and Prentice [2008]'s, a simple step-function estimator that is readily interpretable performed remarkably well when appropriately parameterized.

Perhaps the most important consequence of this decision-theoretic justification is the novel ability to apply Zhong and Prentice [2008]'s estimators to Bayesian inference. Assignment of an informative prior alone is ineffective in removing the Winner's Curse, providing further motivation for Zhong and Prentice [2008]'s approach. Use of both an informative prior and Zhong and Prentice [2008]'s estimators may yield inferior bias-correction compared to use of the latter alone, especially for more extreme values of θ. Therefore, both feasible values of θ should be considered as well as strong scientific motivation for assignment of a prior.

BIBLIOGRAPHY

Shair Ahmad and Antonio Ambrosetti. *A Textbook on Ordinary Differential Equations*. Springer, 2014.

Paul Allison. Why i don't trust the hosmer-lemeshow test for logistic regression. `https://statisticalhorizons.com/hosmer-lemeshow`, March 2013.

Yanik J Bababekov and David C Chang. Post hoc power: a surgeon's first assistant in interpreting "negative" studies. *Annals of surgery*, 269(1):e11–e12, 2019.

Alvin J Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, pages 642–645, 1970.

MJ Bayarri, Daniel J Benjamin, James O Berger, and Thomas M Sellke. Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72:90–103, 2016.

Jolene Birmingham, Andrea Rotnitzky, and Garrett M Fitzmaurice. Pattern–mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):275–297, 2003.

Paula R Blasi, Chloe Krakauer, Melissa L Anderson, Jennifer Nelson, Terry Bush, Sheryl L Catz, and Jennifer B McClure. Factors associated with future dental care utilization among low-income smokers overdue for dental visits. *BMC oral health*, 18(1):183, 2018.

Barbara Bloom, Patricia F Adams, Robin A Cohen, and Catherine Simile. Smoking and oral health in dentate adults aged 18-64. *NCHS data brief*, (85):1–8, 2012.

SL Botman, TF Moore, CL Moriarity, and VL (National Center for Health Statistics) Parsons. Design and estimation for the national health interview survey, 1995-2004. *Vital Health Stat*, 2(130), 2000.

Edward C Capen, Robert V Clapp, William M Campbell, et al. Competitive bidding in high-risk situations. *Journal of petroleum technology*, 23(06):641–653, 1971.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

CDC. Five reasons why calling a quitline can be key to your success. `https://www.cdc.gov/tobacco/campaign/tips/quit-smoking/quitline/index.html`, March 2020.

Kwun Chuen Gary Chan. Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika*, 100(1):269–276, 2013.

David Chavalarias, Joshua David Wallach, Alvin Ho Ting Li, and John PA Ioannidis. Evolution of reporting p values in the biomedical literature, 1990-2015. *Jama*, 315(11):1141–1148, 2016.

Ding-Geng Chen and Jenny K Chen. Statistical power and bayesian assurance in clinical trial design. In *New Frontiers of Biostatistics and Bioinformatics*, pages 193–200. Springer, 2018.

Philip E Cheng. Nonparametric estimation of mean functionals with data missing at random. *Journal of the American statistical association*, 89(425):81–87, 1994.

National Research Council et al. *The prevention and treatment of missing data in clinical trials*. National Academies Press, 2010.

David Roxbee Cox. *Principles of statistical inference*. Cambridge university press, 2006.

David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. Chapman and Hall/CRC, 1979.

Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Chapman and Hall/CRC, 2008.

Marie Davidian. Multiple imputation methods under mar. `https://www4.stat.ncsu.edu/~davidian/st790/notes/chap4.pdf`, 2017.

Susan K Drilea, Britt C Reid, Chien-Hsun Li, Jeffrey J Hyman, and Richard J Manski. Dental visits among smoking and nonsmoking us adults in 2000. *American journal of health behavior*, 29(5):462–471, 2005.

Ronald A Fisher. Inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 26(4):528–535, 1930.

Garrett M Fitzmaurice, Nan M Laird, and Lucy Shneyer. An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in medicine*, 20(7):1009–1021, 2001.

Thomas R Fleming. Clinical trials: discerning hype from substance. *Annals of internal medicine*, 153(6):400–406, 2010.

Lawrence M Friedman, Curt Furberg, David L DeMets, David M Reboussin, Christopher B Granger, et al. *Fundamentals of clinical trials*, volume 4. Springer, 2010.

Chad Garner. Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 31(4):288–295, 2007.

Andrew Gelman. Comment on "post-hoc power using observed estimate of effect size is too noisy to be useful". *Annals of surgery*, 270(2):e64, 2019a.

Andrew Gelman. Don't calculate post-hoc power using observed estimate of effect size. *Annals of surgery*, 269(1):e9–e10, 2019b.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

Robert D Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4): 337–350, 2016.

David J Hand. *Dark Data: Why What You Don't Know Matters*. Princeton University Press, 2020.

Jan Hannig. On generalized fiducial inference. *Statistica Sinica*, 19(2), 2009.

Jan Hannig. Generalized fiducial inference via discretization. *Statistica Sinica*, pages 489–514, 2013.

Jan Hannig, Hari Iyer, Randy CS Lai, and Thomas CM Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515): 1346–1361, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Leonhard Held. Reverse-bayes analysis of two common misinterpretations of significance tests. *Clinical Trials*, 10(2):236–242, 2013.

John M Hoenig and Dennis M Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

Joseph W Hogan and Nan M Laird. Mixture models for the joint distribution of repeated measures and event times. *Statistics in medicine*, 16(3):239–257, 1997.

David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, 9(10):1043–1069, 1980.

Joseph G Ibrahim, Stuart R Lipsitz, and Nick Horton. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Applied Statistics*, pages 361–373, 2001.

Joao Jesus and Richard E Chandler. Estimating functions and the generalized method of moments. *Interface focus*, 1(6):871–885, 2011.

Valen E. Johnson. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1313476110.

Niko A Kaciroti and Trivellore Raghunathan. Bayesian sensitivity analysis of incomplete data: bridging pattern-mixture and selection models. *Statistics in medicine*, 33(27):4841–4857, 2014.

Jae Kwang Kim and Cindy Long Yu. A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493): 157–165, 2011.

Chloe Krakauer and Kenneth Rice. Contribution to the discussion of testing by betting: A strategy for statistical and scientific communication by Chloe Krakauer and Kenneth Rice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, in press.

Monika K Krzyzanowska, Melania Pintilie, and Ian F Tannock. Factors associated with failure to publish large randomized trials presented at an oncology meeting. *JAMA*, 290 (4):495–501, 2003.

Russell V Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001.

Dennis V Lindley. Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1958.

Roderick J Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., 2002.

Joseph F Lucke. A critique of the false-positive report probability. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(2):145–150, 2009.

Xiaodong Luo and Wei Yann Tsai. A proportional likelihood ratio model. *Biometrika*, 99 (1):211–222, 2012.

Robert AJ Matthews. Beyond 'significance': principles and practice of the analysis of credibility. *Royal Society Open Science*, 5(1):171047, 2018.

Robert AJ Matthews. Moving towards the post $p < 0.05$ era via the analysis of credibility. *The American Statistician*, 73(sup1):202–212, 2019.

Deborah Mayo and Richard D Morey. A poor prognosis for the diagnostic screening critique of statistical tests. `http://osf.io/ps38b`, Jul 2017.

Deborah G Mayo. *Statistical inference as severe testing*. Cambridge: Cambridge University Press, 2018.

Deborah G Mayo and Aris Spanos. Severe testing as a basic concept in a neyman–pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(2):323–357, 2006.

Jennifer B McClure, Terry Bush, Melissa L Anderson, Paula Blasi, Ella Thompson, Jennifer Nelson, and Sheryl L Catz. Oral health promotion and smoking cessation program delivered via tobacco quitlines: the oral health 4 life trial. *American journal of public health*, 108(5):689–695, 2018.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press, 2011.

Frederick Mosteller. A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics*, pages 58–65, 1948.

KW Newey and Daniel McFadden. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245, 1994.

Regina Nuzzo. Scientific method: statistical errors. *Nature News*, 506(7487):150, 2014.

Office of the Surgeon General and National Institute of Dental and Craniofacial Research (US). *Oral health in America: a report of the Surgeon General*. US Public Health Service, Department of Health and Human Services, 2000.

Daniel J O'Keefe. Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Communication methods and measures*, 1(4):291–299, 2007.

Art Owen. Thoughts on "Statistical Inference as Severe Testing" by Deborah Mayo. `https://statmodeling.stat.columbia.edu/2019/04/12/several-reviews-of-`

deborah-mayos-new-book-statistical-inference-as-severe-testing-how-to-get-beyond-the-statistics-wars/, Jan 2019.

Giovanni Parmigiani and Lurdes Inoue. *Decision theory: Principles and approaches*, volume 812. John Wiley & Sons, 2009.

Joost DJ Plate, Alicia S Borggreve, Richard van Hillegersberg, and Linda M Peelen. Post hoc power calculation: observing the expected. *Annals of surgery*, 269(1):e11, 2019.

C Radhakrishna Rao. Some comments on the minimum mean square error as a criterion of estimation. Technical report, University of Pittsburgh Institute for Statistics and Applications, 1980.

Kenneth Rice, Tyler Bonnett, and Chloe Krakauer. Knowing the signs: a direct and generalizable motivation of two-sided tests. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2020.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Christian Robert. severe testing:beyond statistics wars. https://xianblog.wordpress.com/2019/01/07/severe-testing-beyond-statistics-wars/, Jan 2019.

Christian P Robert. On the jeffreys-lindley paradox. *Philosophy of Science*, 81(2):216–232, 2014.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

Andrea Rotnitzky and James Robins. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in medicine*, 16(1):81–102, 1997.

Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

Mauricio Sadinle and Jerome P Reiter. Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*, 106(4):889–911, 2019.

Daniel Scharfstein, Aidan McDermott, William Olson, and Frank Wiegand. Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Statistics in Biopharmaceutical Research*, 6(4):338–348, 2014.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

David Schruth. *caroline: A Collection of Database, Data Structure, Visualization, and Utility Functions for R*, 2013. URL `https://CRAN.R-project.org/package=caroline`. R package version 0.7.6.

Thomas M Sellke. On the interpretation of p-values. Technical report, Tech. Rep. Department of Statistics, Purdue University, 2012.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, in press.

Jun Shao and Lei Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 2016.

Michelle Shardell, Eleanor Simonsick, Gregory E Hicks, Barbara Resnick, Luigi Ferrucci, and Jay Magaziner. Sensitivity analysis for nonignorable missingness and outcome misclassification from proxy reports. *Epidemiology*, 24(2):215, 2013.

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

Aris Spanos. Revisiting bayes' rule and evidence: Notional events vs. real data. Technical report, Virginia Tech, working paper, 2015.

D. J Spiegelhalter. *Bayesian approaches to clinical trials and health-care evaluation*. Statistics in practice (Chichester, England). John Wiley & Sons, Chichester, 2004. ISBN 0470092599.

David Spiegelhalter. Andromeda and 'appalling science': a response to hardwicke and ioannidis. `https://medium.com/wintoncentre/andromeda-and-appalling-science-a-response-to-hardwicke-and-ioannidis-a79458efdba1`, Jan 2020.

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

C Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution: Berkeley. *Edited by: Neyman J*, pages 197–206, 1956.

Jerome M Stern and R John Simes. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *Bmj*, 315(7109):640–645, 1997.

Niansheng Tang, Puying Zhao, and Hongtu Zhu. Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica*, 24(2):723, 2014.

James R Thompson. Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63(321):113–122, 1968.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 1998.

Erik van Zwet and Eric Cator. The significance filter, the winner's curse and the need to shrink. *arXiv preprint arXiv:2009.09440*, 2020.

Sholom Wacholder, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6):434–442, 2004.

Jon Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(1):79–86, 2009.

Jon Wakefield. *Bayesian and frequentist regression methods*. Springer Science & Business Media, 2013.

Chenguang Wang, MJ Daniels, Daniel O Scharfstein, and S Land. A bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial. *Journal of the American Statistical Association*, 105(492):1333–1346, 2010.

Ronald L Wasserstein, Nicole A Lazar, et al. The asa's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

Robert WM Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32, 1976.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

John Wood, Nick Freemantle, Michael King, and Irwin Nazareth. Trap of trends to statistical significance: likelihood of near significant p value becoming more significant with extra data. *BMJ*, 348:g2215, 2014.

Rui Xiao and Michael Boehnke. Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(5):453–462, 2009.

Neal S Young, John PA Ioannidis, and Omar Al-Ubaydli. Why current publication practices may distort science. *PLoS Med*, 5(10):e201, 2008.

Arnold Zellner. Bayesian and non-bayesian estimation using balanced loss functions. In *Statistical decision theory and related topics V*, pages 377–390. Springer, 1994.

Hui Zhao, Pu-Ying Zhao, and Nian-Sheng Tang. Empirical likelihood inference for mean functionals with nonignorably missing response data. *Computational Statistics & Data Analysis*, 66:101–116, 2013.

Hua Zhong and Ross L Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9(4):621–634, 2008.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 1

A.1 Missing at Random and Equal Marginal Means

Using auxiliary marginal information, in Section 1.7.3.1, we found our estimated $\mathbb{E}[Y \mid R = 0]$ to be equal to the estimated $\mathbb{E}[Y \mid R = 1]$. It may be tempting to use this equivalence to claim that, supposing $\mathbb{E}[Y \mid R = 0] = \mathbb{E}[Y \mid R = 1]$, that data are missing at random, rendering the new methodology obsolete. However, we will demonstrate that neither (1) MAR implies $\mathbb{E}[Y \mid R = 0] = \mathbb{E}[Y \mid R = 1]$ nor (2) the reverse.

The first point is readily apparent from marginalization of $f(y \mid x, r)$ to derive $f(y \mid r)$. In general,

$$f(y \mid R = 1) = \int_x f(y \mid x, R = 1) f(x \mid R = 1) dx,$$
$$f(y \mid R = 0) = \int_x f(y \mid x, R = 0) f(x \mid R = 0) dx.$$

Recall that under MAR, $f(y \mid x, R = 1) = f(y \mid x, R = 0)$. Therefore, if $f(x \mid R = 1) \neq f(x \mid R = 0)$, then $f(y \mid R = 1)$ may not equal $f(y \mid R = 0)$, possibly resulting in different marginal means among observed and unobserved individuals.

The reverse is particularly important for us to address—i.e. showing $\mathbb{E}[Y \mid R = 0] = \mathbb{E}[Y \mid R = 1]$ does not imply MAR—to motivate use of nonignorable missingness models when these marginal means are (estimated to be) equal, as in the example in Section 1.7. Consider the following example of this principal, which is a simplified version of the model used in Section 1.7. Suppose $[Y \mid x, R = 1]$ is a Bernoulli random variable with success probability defined by a logistic regression on single binary covariate X (and no intercept)

with coefficient λ so that

$$\mathbb{P}[Y = y \mid x, R = 1] = \left(\frac{\exp(\lambda x)}{1 + \exp(\lambda x)}\right)^{y} \left(\frac{1}{1 + \exp(\lambda x)}\right)^{1-y}.$$

Suppose $\mathbb{P}[Y = y \mid x, R = 0]$ is defined via an exponential tilt of $\mathbb{P}[Y = y \mid x, R = 1]$ indexed by coefficient β. $[Y \mid x, R = 0]$ is now also Bernoulli with success probability defined by a logistic regression logit$\mathbb{P}[Y = 1 \mid x, R = 0] = \lambda x + \beta$ because

$$\begin{aligned}\mathbb{P}[Y = y \mid x, R = 0] &= \mathbb{P}[Y = y \mid x, R = 0]\frac{\exp(\beta y)}{\mathbb{E}[\exp(\beta Y)|x, R = 1]} \\ &= \left(\frac{\exp(\lambda x + \beta)}{1 + \exp(\lambda x + \beta)}\right)^{y} \left(\frac{1}{1 + \exp(\lambda x + \beta)}\right)^{1-y}, \qquad \text{(A.1)}\end{aligned}$$

where the first line defines the exponential tilt, which is simplified in the last line. The value of β denotes the dependence of the missingness mechanism on Y, and therefore the departure from MAR. If $\beta \neq 0$, then data are MNAR. Now, $\mathbb{P}[Y = y \mid R = r]$ may be derived by marginalizing over $\mathbb{P}[Y = y, X = x \mid R = r]$. Doing so, and setting $\mathbb{P}[Y = y \mid R = 1]$ equal to $\mathbb{P}[Y = y \mid R = 0]$, we find the equivalence holds if

$$\begin{aligned}\mathbb{P}[X = 1 \mid R = 1]\left\{\frac{1}{1 + \exp(\lambda)} - \frac{1}{2}\right\} + \frac{1}{2} = \\ \mathbb{P}[X = 1 \mid R = 0]\left\{\frac{1}{1 + \exp(\lambda + \beta)} - \frac{1}{1 + \exp(\beta)}\right\} + \frac{1}{1 + \exp(\beta)}. \qquad \text{(A.2)}\end{aligned}$$

There are many $(\beta \neq 0, \lambda)$ pairs for which some pair of probabilities $(\mathbb{P}[X = 1 \mid R = 1], \mathbb{P}[X = 1 \mid R = 0])$ satisfies this equivalence, e.g. with $\beta = -2$ and $\lambda = 10$, for any $\mathbb{P}[X = 1 \mid R = 1] \in (0, 1)$, there exists a $\mathbb{P}[X = 1 \mid R = 0] \in (0, 1)$ satisfying Equation (A.2). Therefore, we may have $\mathbb{E}[Y \mid R = 0]$ equal to $\mathbb{E}[Y \mid R = 1]$ despite data being missing not at random. Hence, we are motivated to use our novel methodology that allows for MNAR to assess the data, despite estimated equivalence of $\mathbb{E}[Y \mid R = 0]$ and $\mathbb{E}[Y \mid R = 1]$.

A.2 Model Checking: Binary Outcomes

In Section 1.7.3.2, we suppose the probability of seeking professional dental care among respondents is linear in the same covariates included in the original regression model used by Blasi et al. [2018] on the logit scale. These covariates were chosen to ensure their inclusion in our analysis to enable comparison to the original results; the logit link was chosen for the sake of convenience, since it is the canonical link for a binary outcome (with distribution in the class of generalized linear models). However, since the novel method is valid only for correctly-specified verifiable densities, these assumptions should be empirically verified in practice. In this section, we recommend strategies for verification of modeling choices with binary outcomes using the verifiable density defined in Section 1.7.3.2 for illustration.

Model-checking with binary outcomes is notoriously-difficult; existing methods for other discrete or continuously-measured outcomes are generally not suitable. In particular, with residuals equal to 0 or 1, simple plots of (deviance) residuals are typically not interpretable. To address this issue, some methods have been developed to directly test assumptions used to model binary outcomes.

Hosmer and Lemeshow developed a goodness-of-fit test specifically for multiple-logistic regression models [Hosmer and Lemesbow, 1980]. It is a Pearson chi-square test, comparing binned predicted probabilities of the outcome using the regression model to the proportion of individuals in that bin with the observed outcome. The null hypothesis is that the probability distribution is the same; hence, a well-fitted model will result in a large p-value. The original method was tested using bins equal to the number of covariates plus one. Following this recommendation, we acquire a p-value of 0.96 when applying the goodness-of-fit test to our assumed verifiable density. Not only is the interpretation of p-values using this statistic somewhat problematic, but also the sensitivity of the method to the number of bins [Allison, 2013]. For instance, by using the default 10 number of bins, the p-value for our verifiable density falls just below 0.50.

Instead, Gelman and Hill [2006, pg 97] recommend comparing these binned predictions

to proportions of individuals with the outcome visually, a categorized version of standard residuals used to assess regression models. In addition to plotting *binned* predicted values vs. *average* residuals—either on the logit or probability scale—Gelman and Hill [2006] include 95% prediction bounds, i.e. roughly 95% of binned residuals should be within those bounds. The width of the bound on either side of 0 is $2\sqrt{\frac{p(1-p)}{n_b}}$, where p is the predicted probability and n_b is the number of individuals in a given bin. As Figure A.1 illustrates, some individuals appear to cause over-estimation of probability for a large number of individuals; however, our modeling choice does not appear to be a particularly bad fit. Gelman and Hill [2006] recommend other model checks, including error rates [Gelman and Hill, 2006, pg 99] and average predictive comparisons [Gelman and Hill, 2006, pg 100-103].

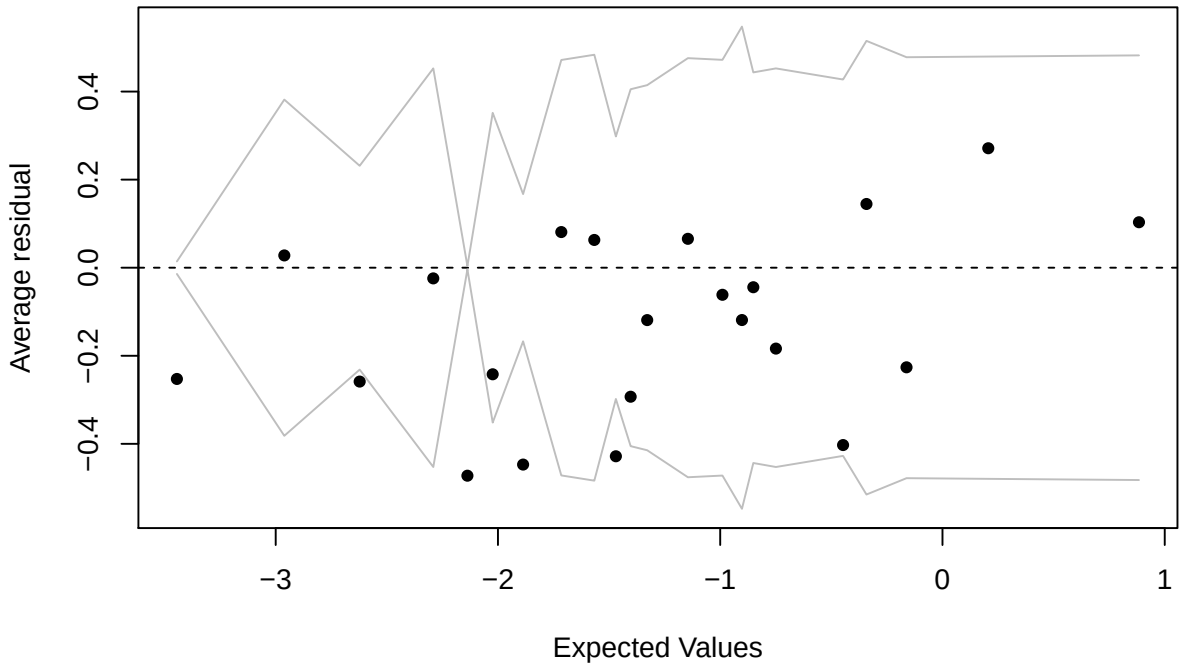


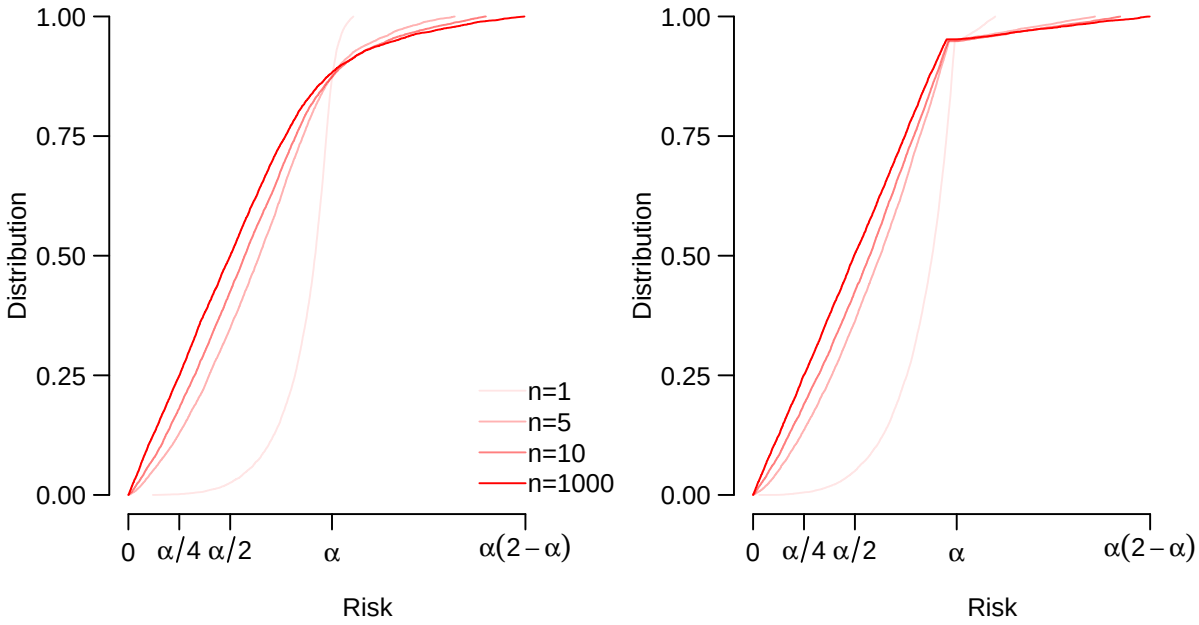
Figure A.1: Binned residuals (on the logit scale) and 95% prediction bounds for $f(y \mid x, R = 1)$.

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

B.1 Posterior Distribution of the Risk

Some results in Section 2.3.2 may be counter-intuitive, particularly the increased spread in the posterior distribution of the risk with increased sample size for a fixed value of $\tilde{p}$. Figure B.1 displays the posterior distribution of the risk when $\tilde{p} = \alpha$ for one- and two-sided tests, this illustrating the underlying mechanism of those spread. Subtle shifts in the distribution can lead to substantial shifts in posterior credible intervals. For both one- and two-sided tests, the posterior distribution of the risk shifts slightly to the left as the sample size increases, thus providing more support for lower values of risk. The distribution simultaneously flattens for values of risk above α so that the range of the distribution appropriately extends to more values of risk (approaching $\alpha(2-\alpha)$). Thus, left- and right-hand tails are longer and credible intervals are wider.



(a) $H_0 : \theta \neq \theta_0$ (b) $H_0 : \theta \leq \theta_0$

Figure B.1: Posterior distribution of risk for when $\tilde{p} = \alpha$ for Normal location problems where $X \sim N(\theta, \sigma^2 = 1)$, $\theta \sim N(\theta_0, 1)$, and $\alpha = 0.05$.

B.2 p-Value Prior

Frequent inability to explicitly define the posterior distribution of risk makes it difficult to understand the relative impact of the prior and data on this distribution. An interesting prior to consider is defined by Cox and Hinkley [1979, pg 395-399]; we refer to this prior as the "p-value prior", following Wakefield [2009].

The p-value prior is motivated by the disagreement between inference from a single dataset using Bayes factors and p-values. In some cases, such as Normal location problems discussed in Sections 2.3.1 and 2.3.2 with Normal priors for θ, the extent of disagreement can be severe. The Cox-Hinkley family of priors address the discrepancy for these problems by forcing Bayes factors (and posterior rejection odds) to be one-to-one functions of p-values, thereby leading to identical inference (with specifically chosen thresholds) using either method.

Conditional on the alternative being true, Cox and Hinkley [1979] assign a $N(\theta_0, v)$ prior to θ, where $v = \frac{d\sigma^2}{n}$ for constant d, i.e. the prior standard deviation is proportionate to the standard deviation of $\bar{X}$. Cox and Hinkley [1979] recommend using $d = 10$. Separately, they assign a non-zero probability p_0 of the null hypothesis being true to enable calculation of the posterior rejection odds [Cox and Hinkley, 1979, pg 397]; see Section 2.2.5 for further details on posterior rejection odds. Despite convenient equality of inference, dependence of the prior distribution on n is antithetical to the Bayesian philosophy [Cox and Hinkley, 1979, pg 397].

For our purposes, when we consider use of the this prior for two-sided tests with point nulls, we do not suppose truth of the null is scientifically plausible (see discussion by Rice et al. [2020]), and hence consider only a special case of the Cox-Hinkley p-value prior where $\theta \sim N(\theta_0, v)$ always with no assignment of point mass at the null. With this prior assignment, $\tilde{p}$ for both one- and two-sided tests are one-to-one functions of the Z-score indexed by the value of d. Hence, $\tilde{p}$ is a one-to-one function of the frequentist p-value.

Consider use of the loss in Equation (2.3) to conduct inference supposing $\theta_0 = 0$, which is displayed in Figure B.2. The same figure displays an interesting relationship between the risk curve and sample size with this prior: the risk is identical for any value of θ the same

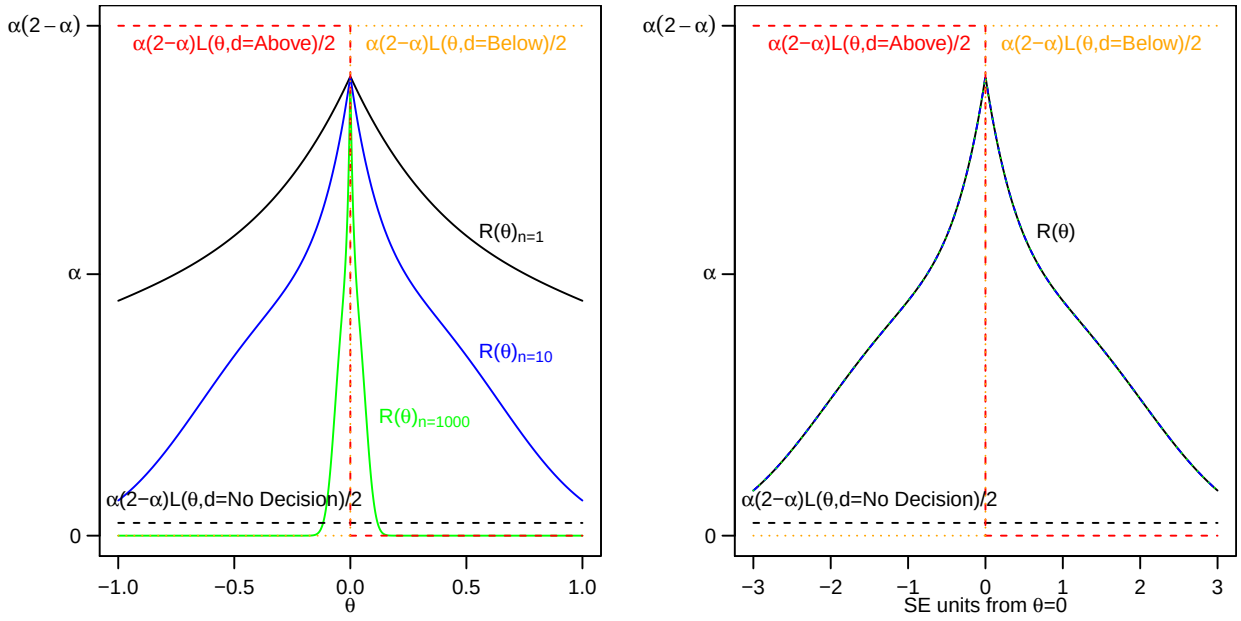


Figure B.2: Risk and loss (scaled) for the p-value prior for a Normal location problem where $X \sim N(\theta, \sigma^2 = 1)$ and $\theta \sim N\left(\theta_0, \frac{d\sigma^2}{n}\right)$ with $d = 10$. Loss does not depend on sample size; sample size is denoted in subscript for risk. All risk curves are superimposed when plotted on the SE scale (on the right).

number of SE units away from $\theta = 0$. Risk also happens to monotonically decrease as the sample size grows for any fixed value of θ, which we have not seen with other studied priors. Finally, the risk is identical at $\theta = 0$ across sample sizes. These results may perhaps be intuitively explained by the constantly proportionate impact of the prior distribution and data on the posterior distribution for θ, and hence the risk.

Figure B.3 shows how the posterior mean of risk is identical across sample sizes for fixed $\tilde{p}$, making this estimate of the risk a one-to-one function of $\tilde{p}$, and hence the p-value. Also note posterior credible intervals for risk are identical across sample sizes, not prone to the same widening of posterior credible intervals as sample size increases in Figure 2.8 with arbitrary prior variance assignments.

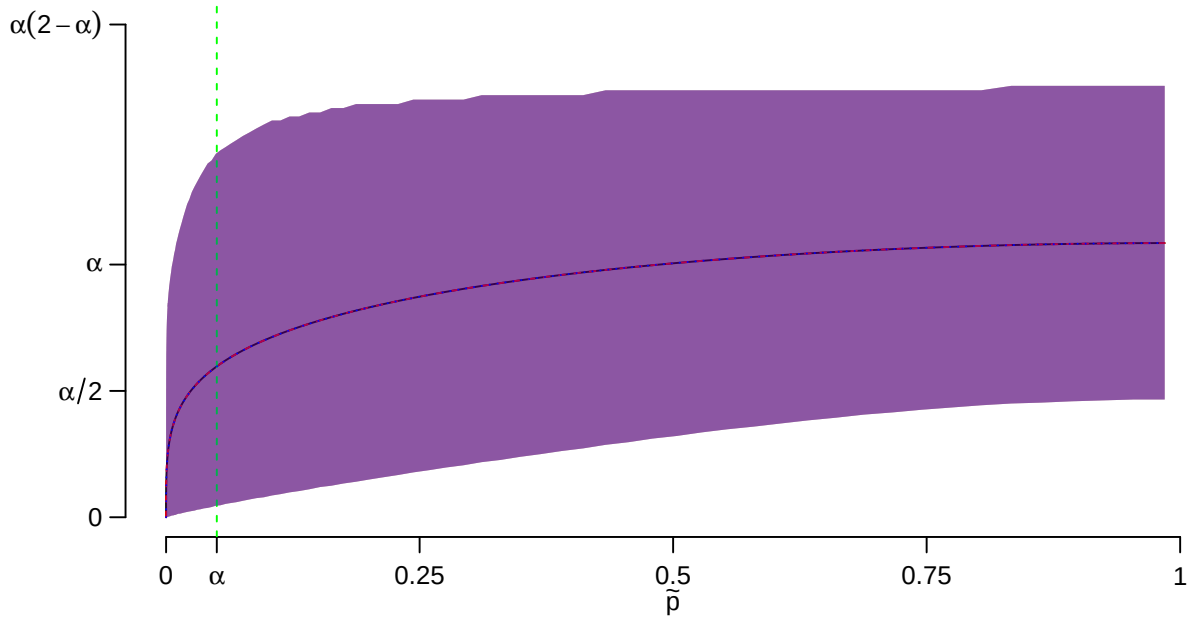


Figure B.3: Posterior mean and 95% posterior credible interval of risk for a given value of double the minimum posterior tail area (denoted by $\tilde{p}$) for $n = 1$ (in red), $n = 10$ (in blue), and $n = 1000$ (in grey); posterior means and credible intervals identical across sample sizes.

In conclusion, with this prior, all aforementioned results suggest increasing sample size is guaranteed to decrease risk, much like the same guaranteed increase in power. However, estimates of reliability of inference with the same $\tilde{p}$ across studies with different sample sizes is identical, and hence cannot be improved by changing the sample size. These results mirror the one-to-one relationship between the p-value and observed power, which are not surprising given equality of inference using this prior and frequentist significance tests.

A topic of future research will be uncovering if this result holds across all priors yielding identical frequentist and Bayesian inference, and therefore if there is information to be gained about the risk of a frequentist test beyond the p-value.

B.3 Observed Power as a Function of p-Values

The universal claim by Hoenig and Heisey [2001] has been reiterated by numerous authors [O'Keefe, 2007, Lenth, 2001, Plate et al., 2019]. It does not appear to hold, exactly, when the

measure of precision defining the test statistic depends on the null hypothesis—for example, a score test for disease prevalence exceeding a given value.

Consider a one-sided score test of binomial proportion θ, i.e. $H_0 : \theta \leq \theta_0$ vs. $H_A : \theta > \theta_0$, $\theta_0 \neq 0$ where $X \sim \text{Bin}(n, \theta)$ with sufficiently large sample size so that we may assume the sample proportion $\hat{\theta}(X) \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$ and conduct a test with test statistic $Z_p = \frac{\hat{\theta}-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$, the sole quantity determining the p-value. In this case, the power $\pi(\cdot)$ at a given value of prevalence θ' is

$$\pi(\theta') = 1 - \Phi\left(\sqrt{\frac{\theta_0(1-\theta_0)}{\theta'(1-\theta')}}\left[z_{1-\alpha} + \frac{\theta_0 - \theta'}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}\right]\right)$$

Calculating the observed power at $\hat{\theta}$ yields

$$\pi(\hat{\theta}) = 1 - \Phi\left(\sqrt{\frac{\theta_0(1-\theta_0)}{\hat{\theta}(1-\hat{\theta})}}\left[z_{1-\alpha} - Z_p\right]\right)$$

Therefore, observed power depends on $\hat{\theta}$ not just through the p-value and the claimed relationship of Hoenig and Heisey [2001] does not hold. If instead the Wald test is used, i.e. the test statistic is $\frac{\hat{\theta}-\theta_0}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}}$, the 1:1 relationship between the p-value and observed power is maintained.

Next, we consider a test of the same proportion using an exact binomial test, supposing $\theta_0 = 0.5$ with p-value threshold $\alpha = 0.05$ without loss of generality. The p-value after observing x out of n successes, defined below as a function of x and n, reduces to

$$\begin{aligned} p(x, n) &= \mathbb{P}[X \geq x | \theta = 0.5] \\ &= 0.5^n \sum_{k=x}^{n} \binom{n}{k} \end{aligned}$$

Two types of exact binomial tests are considered. In the first, the null hypothesis is rejected

whenever the p-value is below the threshold α, a.k.a. a *standard exact binomial test*, yielding power at a given θ' equal to

$$\begin{aligned} \pi(\theta') &= \mathbb{P}[X \geq c_n; \theta'] \\ &= \sum_{k=c_n}^{n} \binom{n}{k} (\theta')^k (1-\theta')^{n-k}, \text{ where} \end{aligned} \tag{B.1}$$

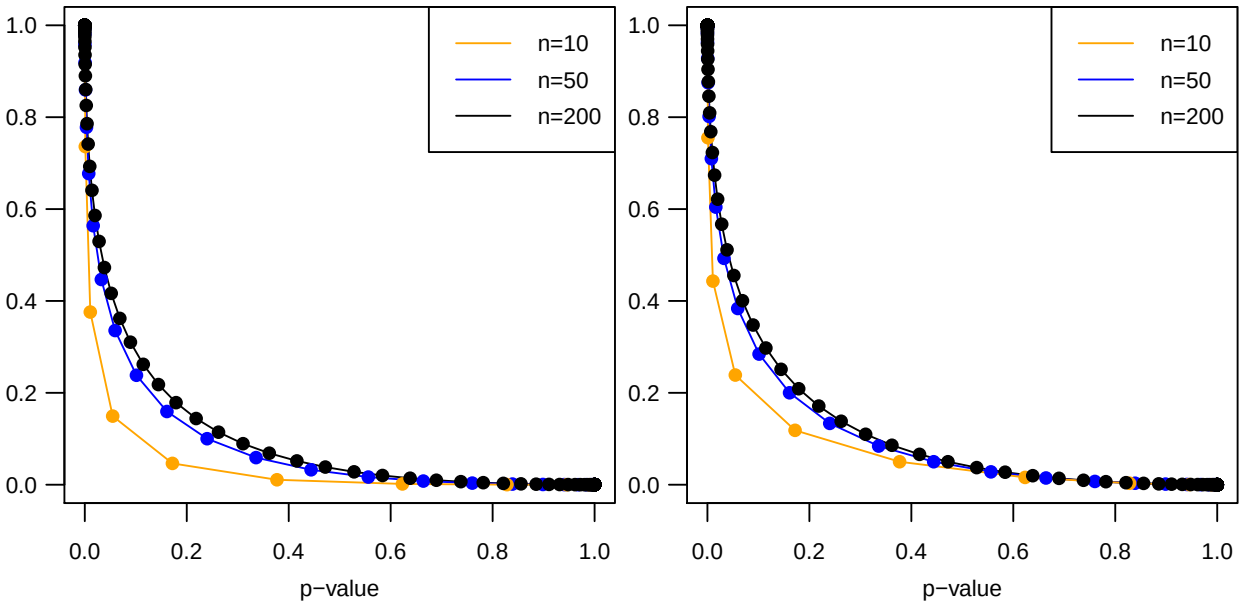
$$\text{Critical value } c_n = min\{x : p(x,n) \leq \alpha\}. \tag{B.2}$$

Due to the empty intersection of binomial distribution values across sample sizes, the exact Type I error rate will differ across tests with different sample sizes; furthermore, it is not possible to obtain the same p-value from tests with two different sample sizes. To alleviate the former source of discrepancy between tests of different sizes, for the second type of test, decisions are randomized for observed counts of successes one less than the critical value so that the Type I error rate is exactly α for all sample sizes, a.k.a. an *exact binomial test with randomized decision*. In this case, at a given θ', the power of the test is

$$\begin{aligned} \pi(\theta') &= \mathbb{P}[X \geq c_n; \theta'] + \gamma_n \mathbb{P}[X = c_n - 1; \theta'], \text{ where} \\ \gamma_n &= \frac{\alpha - \mathbb{P}[X \geq c_n; \theta = 0.5]}{\mathbb{P}[X = c_n; \theta = 0.5]}, \end{aligned} \tag{B.3}$$

with c_n defined as in Equation (B.2). Uniqueness of distribution values to each sample size will continue to make p-values unique to tests with each sample size.

Figure B.4 displays observed power for all possible observable p-values (using points, connected by solid lines) for sample sizes of 10, 50, and 200 assuming use of the two types of exact binomial tests above. In this case, observed power curves for smaller sample sizes sit below those of larger sample sizes, albeit by a small amount. The difference in observed power curves diminishes slightly when the Type I error rate is controlled. While determination of observed power is largely dictated by the p-value, sample size does influence the calculation, particularly for smaller p-values. In particular, note the discrepancy can be sizeable for



(a) Exact Binomial Test, Equation B.1

(b) Randomized Decision, Equation B.3

Figure B.4: Observed power for a binomial exact test of proportion θ. Observed power calculations only possible at plotted round points, with lines plotted to connect them for visual aid.

just-non-significant results.

Lack of applicability of the results of the 1:1 relationship between p-value and observed power by Hoenig and Heisey [2001] to exact binomial tests may be argued due to impossibility of the same p-value across tests with different testing parameters (i.e. sample size). For those who suppose approximate equivalence of observed power with similar p-values across tests with different sample sizes, and given increasing temptation to calculate observed power with just-non-significant results, investigators should consider extension of the results of Hoenig and Heisey [2001] to exact binomial tests with caution.

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

C.1 Shrinkage and Loss Function for Select Estimators

Estimator	$\mathbf{s(Z^2)}$	$\mathbf{f(s)}$
Z&P Mean	$s_{\mathrm{E}}(Z^2) : Z = s_{\mathrm{E}}(Z^2)Z + \frac{\phi\left(s_{\mathrm{E}}(Z^2)Z-c\right)-\phi\left(-s_{\mathrm{E}}(Z^2)Z-c\right)}{\Phi(s_{\mathrm{E}}(Z^2)Z-c)+\Phi(-s_{\mathrm{E}}(Z^2)Z-c)}$	Numerical approximation using Equation (3.12)
Z&P MSE	$\frac{1+Z^2 s_{\mathrm{E}}(Z^2)(1-s_{\mathrm{E}}(Z^2))^2}{1+Z^2(1-s_{\mathrm{E}}(Z^2))^2}$	Numerical approximation using Equation (3.12)
Thompson(k), $k > 0$	$\frac{Z^2}{k+Z^2}$	$\frac{1}{(1-s)^{1+2k}} \exp\left\{ \frac{\frac{2(k-1)\sqrt{k}\,\mathrm{ArcTan}\left[\frac{\sqrt{k}(1+2s)}{\sqrt{4-k}}\right]}{\sqrt{4-k}} - (1+k)[1+ks(1+s)]}{1+2k} \right\}$
Stein(a), $a \geq 0$	$\max(0, 1 - a/Z^2)$	$\frac{1}{1-s}\left(\frac{1-s}{1+a(1+s)}\right)^{\frac{2+2a}{1+2a}}$
Step($s = 0, k \geq 0$)	$1_{Z^2>k^2}$	$\frac{1-s}{1+k^2(1-s^2)}$

Table C.1: Shrinkage functions $s(Z^2)$ for select data-adaptive shrinkage functions, along with functions $f(s)$ defining the loss in Equation 3.8 yielding $s(Z^2)$ as a Bayes rule. All functions $f(s)$ may multiplied by a constant; note that the constant multiplier for Thompson(k) may need to be complex to yield real-valued function $f(s)$.