University of Kentucky
UKnowledge

Theses and Dissertations--Statistics

Statistics

2016

Statistical Methods for Handling Intentional Inaccurate Responders

Kristen J. McQuerry
University of Kentucky, kristen.mcquerry@uky.edu
Digital Object Identifier: http://dx.doi.org/10.13023/ETD.2016.280

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Recommended Citation

McQuerry, Kristen J., "Statistical Methods for Handling Intentional Inaccurate Responders" (2016). *Theses and Dissertations--Statistics*. 17.
https://uknowledge.uky.edu/statistics_etds/17

This Doctoral Dissertation is brought to you for free and open access by the Statistics at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Statistics by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Kristen J. McQuerry, Student

Dr. Arnold Stromberg, Major Professor

Dr. Constance Wood, Director of Graduate Studies

STATISTICAL METHODS FOR HANDLING INTENTIONAL INACCURATE RESPONDERS

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in the
College of Arts & Sciences
at the University of Kentucky

By

Kristen McQuerry

Lexington, Kentucky

Director: Dr. Arnold Stromberg, Professor of Statistics

Co-Director: Dr. Heather Bush, Professor of Biostatistics

Lexington, Kentucky

Copyright © Kristen McQuerry 2016

ABSTRACT OF DISSERTATION

STATISTICAL METHODS FOR HANDLING INTENTIONAL INACCURATE RESPONDERS

In self-report data, participants who provide incorrect responses are known as intentional inaccurate responders. This dissertation provides statistical analyses for address intentional inaccurate responses in the data.

Previous work with adolescent self-report, labeled survey participants who intentionally provide inaccurate answers as mischievous responders. This phenomenon also occurs in clinical research. For example, pregnant women who smoke may report that they are nonsmokers. Our advantage is that we do not solely have self-report answers and can verify responses with lab values. Currently, there is no clear method for handling these intentional inaccurate respondents when it comes to making statistical inferences.

We propose a using an EM algorithm to account for the intentional behavior while maintaining all responses in the data. The performance of this model is evaluated using simulated data and real data. The strengths and weaknesses of the EM algorithm approach will be demonstrated.

KEYWORDS: intentional inaccurate response, self-report data, EM algorithm

KRISTEN MCQUERRY

Student's Signature

JUNE 24, 2016

Date

STATISTICAL METHODS FOR HANDLING INTENTIONAL INACCURATE RESPONDERS

By

Kristen McQuerry

ARNOLD STROMBERG
Director of Dissertation

HEATHER BUSH
Co-Director of Dissertation

CONSTANCE WOOD
Director of Graduate Studies

JUNE 24, 2016
Date

To Ryan and Kilian McQuerry.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor Dr. Heather Bush for her support, mentorship, and guidance throughout the process of writing this dissertation. I would like to thank Dr. Constance Wood for her ideas and advice that contributed to the methods presented in this work. Also, I would like to thank Dr. Kristin Ashford for her generosity in letting me use her data to test my methods. I extend my thanks to the remaining members of my dissertation committee, Dr. Arnold Stromberg, Dr. William Griffith, and Dr. David Fardo, for their encouragement and feedback; as well as, the Department of Statistics for all the support I have received during my time in the Statistics program. Most importantly, I have received an unmeasurable amount of moral support from Ryan McQuerry, Keith and Marilyn Eichhorn, and Mindy Eichhorn.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Self-report data are often collected in epidemiology, psychology, pediatrics, and social and behavioral science settings. Siggeirsdottir et al. (2007) Bernard et al. (1984) Sherry et al. (2007) Fan et al. (2006) Self-report data are define as data collected from surveys containing questions that ask respondents to report something about themselves. Chan (2009) Unfortunately, these data have the potential to contain inaccurate responses due to "carelessness, confusion, lack of efforts, or intentional mischief" Fan et al. (2006) Sherry et al. (2007) Siggeirsdottir et al. (2007) Fan et al. (2006) Robinson-Cimpian (2014). Sometimes Quality of Life and Patient Reported Outcomes can only be collected through self-report; therefore, members of the health and medical fields frequently use self-report data collection methods in order to analyze their research. In particular, surveys have become widely accepted among professional and policy-making organizations Fan et al. (2006) Cornell et al. (2012).

Although data quality has been a long time topic of surveys, response validity has been given greater attention in survey literature. Bernard et al. (1984) Fan et al. (2006) Cornell et al. (2012) Robinson-Cimpian (2014) Bernard et al Bernard et al. (1984) looked at respondent accuracy in the areas of child care behavior, health seeking behavior, and communication and social interactions. They concluded that in all the studies they examined, on average, about half of what a participant answers is incorrect in some form. Respondents who answer inaccurately have been labeled as "jokesters" Fan et al. (2006) or "mischievous responders" Robinson-Cimpian (2014). Here the inaccurate responders are purposely giving answers that are opposite of the social norm. While it is possible to use data responses to identify intentional inaccurate responders Robinson-Cimpian (2014), there is no clear method for how to handle these respondents and the impact on statistical inferences.

The motivating clinical example of this problem of intentional inaccurate responses occurs in data from a sample of pregnant women collected to determine if smoking status effected progesterone levels. Pregnant women self-reported their smoking status as nonsmoker or smoker. Due to the sensitive nature of asking about smoking status during pregnancy and that smoking while pregnant is highly discourage by health care providers, participants may consciously minimize their actual cigarette use. Klesges et al. (1995) As the adverse affects of smoking while pregnant are widely known Lumley et al. (2009), higher levels of nonsmoking is expected. Here the inaccurate responders are purposely giving answers that are aligned with the social norm to not smoke while pregnant. In this case, the intentional inaccurate responses are not given at random. There is a pattern to the inaccuracy because these women are reporting answers that they think they should be giving, instead of the truth. This, in turn, will bias the data towards nonsmoking and could lead to inaccurate conclusions.

Previous strategies for accounting for intentional inaccurate responses include ignoring the inaccuracy and analyzing all self-report data or removing all inaccuracies from the dataset. Cornell et al. (2012) Other methods involve using a weighted Probability-Based Index Robinson-Cimpian (2014) as a covariate in regression. However, ignoring data imposes biased caused from incomplete data. We propose extending incomplete data methodology to appropriately account for intentional inaccuracies. These methods includes Heckman's model and the EM algorithm for finite mixtures.

To compare strategies for handling intentional inaccurate responders, a simulation study was conducted based on the motivating pregnant women example, which contains a continuous outcome and a dichotomous exposure group. Simulations varied samples sizes, probabilities of an intentional inaccurate responder, and coefficients of variation.

This dissertation focuses on a particular problem that involves self-report survey values and suggests recommendations for accounting for the bias that occurs when these self-report values do not represent the truth.

Chapter 2 provides a literature review of previous work. First the focus is on the current topic of inaccurate responders and how they are handled when making statistical inferences. Then the focus turns to common incomplete data statistical methods.

Chapter 3 provides a novel application of incomplete data methods to account for intentional inaccurate responses. Assumptions of incomplete data will be revised and assumptions of inaccurate data will be created. Recommendations are made for analyzing data with non-ignorable intentional inaccurate responses using Heckman's model and the EM algorithm.

Chapter 4 is a simulation study that compares current methods to the recommended methods from Chapter 3.

Chapter 5 applies the current strategies and methods based on incomplete data applications to real data that has identified intentional inaccurate responses. Data were collected from pregnant women, in which, there are self-report and lab values indicating smoking status. This information is used to show how the effect of smoking status on progesterone levels change when different methods are considered.

Chapter 6 concludes the dissertation and discusses future work in this area.

Chapter 2

Literature Review

This chapter presents an overview of previous work analyzing self-report data, and methods of handling non-ignorable incomplete data.

2.0.1 Mischievous Responders & Probability-Based Index

An issue that emerges from self-report data collection is the presence of respondents that do not answer truthfully. These respondents have been previously named "mischeivous responders." Robinson-Cimpian Robinson-Cimpian (2014) shows examples of the dangers these participants play in causing incorrect conclusions from the data. Robinson-Cimpian's paper focuses on adolescent self-report. His mischievous responders are identified by selecting questions that are unrelated, but maybe correlated for mischievous answers. For instance, researchers would not expect to find multiple adolescents who are blind and deaf and in a gang and identify as LGBQ and parenting multiple children. In his example, it would not be expected that adolescents identifying as LGBQ would report they are blind more often than their heterosexual peers. This was the case in the self-report data that Robinson-Cimpian was analyzing. For LGBQ-identified youths, 13.9% said they were blind; whereas, 2.9% of heterosexual-identified answered this way.

The answer of yes to being blind and the answer of identifying as lesbian/gay are considered low-frequency responses. These are responses that one would not expect the majority of adolescents to select. In Robinson-Cimpian's paper, a weighted Probability-Based Index, P_i, is calculated based on the number of low-frequency responses. For instance, let $M = 2$. One individual gave a response that 10% of individuals provided for Item 1, and a response for Item 2 that 5% of individuals provided, their value of $P_i = 0.005$. On the other hand, another individual gave responses for Item 1 and Item 2 that 90% and 95% of adolescents provided would have $P_i = 0.855$. Mischievous responders are identified by having low P_i values. These P_i values were included in statistical models and used as a covariate in the regression equations in order to adjusted for the mischievous responses.

2.0.2 Weighting

Sometimes, samples from self-report data do not represent that of the population. Calculating weights to account for these biases in self-report data has been used by researchers in the past Skinner et al. (1989).

When statisticians encounter survey data, they must figure out if sampling weights for the point estimates should be used and how to estimate the variance of the weighted estimates. If weighting is related to the estimate of interest and calculated correctly, it can reduce bias.

The sampling weight w_i is the reciprocal of the probability of selection. The sum of the sample weights equals the size of the population. Many statisticians disagree with the use of sampling weights in regression analysis. Brewer and Mellor Brewer and Mellor (1973) looked into the issue and divided statisticians into two groups, model-based and designed-based.

Model-Based Approach.— Let $\mathbf{z}_1, ..., \mathbf{z}_N$ be a vector of auxiliary variables. Also, denote $p(S)$ as a simple random sampling design. In the model-based approach, inference is based with respect to the sampling distribution of statistics over $\mathbf{y}_1, ..., \mathbf{y}_N$ generated by the model, ξ Skinner et al. (1989). Thus, for a single explanatory variable, the linear regression model is

$$Y_i | x_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Here Y_i is a random variable for the response and x_i is an explanatory variable. β_0 and β_1 are unknown parameters. The ε_i's are $\sim iid\ N(0, \sigma^2)$. The ordinary least squares (OLS) estimates of the parameter β_0 and β_1 are obtained by solving normal equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n}$$

For the model-based approach, bias, known as $\xi - bias$, is defined as the expectation of the difference of $\hat{\beta}$ and β. The $\xi - bias$ of $\hat{\beta}$ is

$$bias_{\xi}(\hat{\beta}) = E_{\xi}(\hat{\beta} - \beta)$$

In this instance, the finite population parameters β are random variables with respect to ξ, which is why bias$_{\xi}(\hat{\beta})$ in the above equation contains β within the parentheses. The estimator β is $\xi - unbiased$ if bias$_{\xi}(\hat{\beta}) = 0$ Skinner et al. (1989). Consistent is defined as an estimator which converges in probability, as the sample size increases, to the parameter of which it is an estimator Brewer and Mellor (1973).

Designed-Based Approach.— The Design-based approach estimates quantities from a finite population. Here the inference is based with respect to the sampling distribution of statistics over repeated samples, S, generated by the sampling design $p(S)$. The values $\mathbf{y}_1, ..., \mathbf{y}_N, \mathbf{z}_1, ..., \mathbf{z}_N$ are held fixed. Skinner et al. (1989). The finite population quantities of interest for regression are the least squares coefficients for the population, B_0 amd B_1, that minimize $\sum_{i=1}^{N}(y_i - B_0 - B_1 x_i)^2$ over the entire finite population. We can estimate B_0 and B_1 using weights, $w_1, ..., w_n$,

$$\hat{B}_1 = \frac{\sum_{i=1}^{n} w_i x_i y_i - \frac{(\sum_{i=1}^{n} w_i x_i)(\sum_{i=1}^{n} w_i y_i)}{\sum_{i=1}^{n} w_i}}{\sum_{i=1}^{n} w_i x_i^2 - \frac{(\sum_{i=1}^{n} w_i x_i)^2}{\sum_{i=1}^{n} w_i}}$$

$$\hat{B}_0 = \frac{\sum_{i=1}^{n} w_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

For the designed-based approach, bias, known as $p - bias$, is defined in terms of the expectation over all possible samples. The $p - bias$ of a point estimator $\hat{\beta}$ of β is

$$bias_p(\hat{\beta}) = E_p(\hat{\beta}) - \beta$$

The estimator β is $p - unbiased$ if bias$_p(\hat{\beta}) = 0$ Skinner et al. (1989). Consistent is defined as an estimate that becomes exactly equal to the population value when n=N; in other words, when the sample consists of the whole population Brewer and Mellor (1973).

Model-Based or Design-Based.— The Model-Based and Design-Based approaches align when the probability of selection is part of the estimate. For example, using sampling weights from a stratified sampling frame to estimate a mean. However, if the sampling weight is the inverse of the probability of selection, but the estimate of interest is unrelated to the sampling weight, weighting will give a biased estimate.

2.0.3 Self-Report data: Bias

Incomplete Data.— In this dissertation, self-report data that includes intentional inaccurate responders is considered to be incomplete. Part of the data does not represent the truth. Incomplete data is usually used synonymously with missing data. With missing data, part of the data does not exist. The section looks at current incomplete data methodology used for missing data purposes.

Nonresponse is defined as the failure to measure some of the units in the selected sample Cochran (2007). Nonresponse leads to sample selection bias. Sample selection bias arises because samples measured are not randomly selected. The largest obstacle in accounting for selection bias is the lack of sufficient information for statistical inference. Selection bias is controllable when the factors that affect selection are measured on all study subjects. These factors can be controlled similar to confounding variables or if one knows the joint distribution of these factors in the population. In most studies, one can usually only control as appropriate and hope that no other factors have influenced selection Greenland et al. (1998).
Other techniques, can account for the incomplete data. Methods to handle the selection bias caused by an incomplete data factor in how the incompleteness depends on the response Y.

Intentional Inaccurate Response.— Intentional inaccurate responses are reponses where the participant consciously answers incorrectly. Responses exist; however, the answers are not truthful. Especially for smaller samples, these incorrect responses bias the data and could lead to incorrect conclusions. Once identified, intentional inaccurate responders are usually removed from the data. Cornell et al. (2012) Unfortunately removing participants can lead to bias from nonresponse since the data from these participants will not exist. Ideally one would identify the intentional inaccurate responders and then account for the false responses within the statistical analysis to give an unbiased estimate. This would allow one to control for the bias from untruthful responses without acquiring nonresponse bias. To achieve this scenario, modified incomplete data tactics will be used.

2.0.4 Intentional Inaccurate Response: Incomplete Data

Incomplete Data Overview.— Let $\mathbf{y} = (y_{obs}, y_{inc})$ be a data matrix partitioned into observed y_{obs} values and incomplete y_{inc} values. We can denote the incomplete-data indicator as:
If $\mathbf{y} = (y_{ij})$ an $n \times K$ matrix of n observations measured for K variables, then

$$M_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ incomplete} \\ 0, & \text{if } y_{ij} \text{ observed} \end{cases}$$

The incomplete-data mechanism is characterized by the conditional distribution of M given $\mathbf{y}$, $f(M|\mathbf{y}, \theta)$, where θ denotes unknown parameters Little and Rubin (2014). Table 2.1 gives an explanation of each data mechanism.

Table 2.1: Incomplete-Data Mechanisms

Incomplete-Data Mechanism	**Explanation**
Incomplete Completely At Random (MCAR)	If incompleteness does not depend on the values of the data $\mathbf{y}$, incomplete or observed: $f(M\|\mathbf{y}, \theta) = f(M\|\theta)$ for all $\mathbf{y}, \theta$
Incomplete At Random (MAR)	If incompleteness depends only on the components y_{obs} of $\mathbf{y}$ that are observed, and not on the components that are incomplete: $f(M\|\mathbf{y}, \theta) = f(M\|y_{obs}, \theta)$ for all y_{inc}, θ
Not Incomplete At Random (NMAR)	If the distribution of M depends on the incomplete values in the data matrix $\mathbf{y}$.

Non-Ignorable Incomplete Data.— When the non-response of a participant has no relationship to the incomplete values of the variables, this type of incomplete value is ignorable. However, some data have incomplete values with a relationship to the non-response. In this case, the incomplete data are non-ignorable. There is a pattern to the incomplete data that must be modeled. With non-ignorable incomplete data models, the maximum likelihood estimation requires a model for the incomplete-data mechanism and maximization of the full likelihood. Little and Rubin (2014) Below are examples of handling non-ignorable incomplete data.

Pattern-Mixture and Selection Models for Univariate Nonresponse.— Suppose that incomplete values are confined to a single variable.
Let $y_i = (y_{i1}, y_{i2})$, where y_{i1} is fully observed and random variable y_{i2} is observed for $i = 1, ..., r$ and incomplete for $i = r + 1, ..., n$.
Let $M_i = M_{i2} = 1$ if y_{i2} is incomplete and $M_{i2} = 0$ if y_{i2} is observed. Then the density of y_{obs} and M is:

$$F(y_{obs}, M|\theta) = \prod_{i=1}^{r} f(y_{i1}, y_{i2}|M_{i2} = 0, \theta)P(M_{i2} = 0|\theta)$$
$$\prod_{i=r+1}^{n} f(y_{i1}|M_{i2} = 1, \theta)P(M_{i2} = 1|\theta)$$

The basic difficulty is that there is no data to estimate the distribution $f(y_{i2}|y_{i1}, M_{i2} = 1, \theta)$ since all observations with $M_{i2} = 1$ have y_{i2} incomplete. To make any headway, the distribution of $f(y_{i2}|y_{i1}, M_{i2} = 1, \theta)$ for nonrespondents must be related to the corresponding distribution $f(y_{i2}|y_{i1}, M_{i2} = 0, \theta)$ for respondents. Little and Rubin (2014)

Grouped Normal Data with Covariates.— Suppose hypothetical complete data are an independent random sample $(y_1, ..., y_n)$ from the normal distribution with a linear regression on fully observed covariates $x_1, x_2, ..., x_p$. That is, y_i is normal with mean $\beta_0 + \sum_{k=1}^{p} \beta_k x_{ik}$ and constant variance σ^2. Here y_i is observed for $i = 1, ..., r < n$. The remaining $n - r$ cases are grouped into J categories such that the j^{th} category contains values of y_i known to lie between a_j and b_j. Observed data for these $n - r$ cases are counts m_j of observations in the j^{th} category for $j = 1, ..., J$, $\sum_{j=1}^{J} m_j = n - r$. This formation includes censored data, where $a_j > 0$ and $b_j = \infty$, as well as situations where $r = 0$.
Let the incomplete-data indicator be

$$M_i = \begin{cases} 1, & \text{if } y_i \text{ falls in the } j^{th} \text{ nonresponse category } (a_j < y_i < b_j)\ j = 1, ..., J \\ 0, & \text{ if } y_i \text{ is observed} \end{cases}$$

The complete-data sufficient statistics are $\sum y_i, \sum y_i x_{ik} (k = 1, ..., p)$, and $\sum y_i^2$. Using the EM algorithm, the E step computes

$$E(\sum_{i=1}^{r} y_i|Y_{obs}, M, \theta = \theta^{(t)}) = \sum_{i=1}^{r} y_i + \sum_{i=r+1}^{n} \hat{y}_i^{(t)},$$
$$E(\sum_{i=1}^{r} y_i x_{ik}|Y_{obs}, M, \theta = \theta^{(t)}) = \sum_{i=1}^{r} y_i x_{ik} + \sum_{i=r+1}^{n} \hat{y}_i^{(t)} x_{ik},\ k = 1, ..., p,$$
$$E(\sum_{i=1}^{r} y_i^2|Y_{obs}, M, \theta = \theta^{(t)}) = \sum_{i=1}^{r} y_i^2 + \sum_{i=r+1}^{n} (\hat{y}_i^{(t)2} + \hat{s}_i^{(t)2}),$$

where $\theta = (\beta_0, \beta_1, ..., \beta_p, \sigma^2), \theta^{(t)} = (\beta_0^{(t)},\ \beta_1^{(t)}, ..., \beta_p^{(t)}, \sigma^{(t)2})$ is the current estimate of $\theta, \hat{y}_i^{(t)} = \mu_i^{(t)} + \sigma^{(t)}\delta_i^{(t)}, \hat{s}_i^{(t)2} = \sigma^{(t)2}(1 - \gamma_i^{(t)}), \mu_i^{(t)} = \beta_0^{(t)} + \sum_{k=1}^{p} \beta_k^{(t)} x_{ik}$, and $\delta_i^{(t)}$ and $\gamma_i^{(t)}$ are corrections for the nonignorable nonresponse, which take the form

$$\delta_i^{(t)} = -\frac{\phi(d_i^{(t)}) - \phi(c_i^{(t)})}{\Phi(d_i^{(t)}) - \Phi(c_i^{(t)})},$$

$$\gamma_i^{(t)} = \frac{\delta_i^{(t)2} + d_i^{(t)}\phi(d_i^{(t)}) - c_i^{(t)}\phi(c_i^{(t)})}{\Phi(d_i^{(t)}) - \Phi(c_i^{(t)})},$$

where ϕ and Φ are the standard normal density and cumulative distribution function, and for units i in the j^{th} category ($M_i = j$ or equivalently, $a_j < y_i < b_j$), $c_i^{(t)} = \frac{a_j - \mu_i^{(t)}}{\sigma^{(t)}}$ and $d_i^{(t)} = \frac{b_j - \mu_i^{(t)}}{\sigma^{(t)}}$. The M step calculates the regression of y_i on $x_{i1}, ..., x_{ip}$ using the expected values of the complete-data sufficient statistics found in the E step. Little and Rubin (2014)

Censored Normal Data with Covariates - Tobit Model.— A special case of the grouped normal data with covariates occurs when positive values of y_i are fully recorded but negative values are censored. These censored values can lie anywhere in the interval $(-\infty, 0)$. In the notation in the above example, all observed y_i are positive, $J = 1, a_1 = -\infty$, and $b_1 = 0$. We have for censored cases $c_i^{(t)} = -\infty, d_i^{(t)} = -\frac{\mu_i^{(t)}}{\sigma^{(t)}}, \delta_i^{(t)} = -\frac{\phi(d_i^{(t)})}{\Phi(d_i^{(t)})}, \gamma_i^{(t)} = \delta_i^{(t)}(\delta_i^{(t)} + \frac{\mu_i^{(t)}}{\sigma^{(t)}})$. Thus

$$\hat{y}_i^{(t)} = E(y_i|\theta^{(t)}, x_i, y_i \leq 0) = \mu_i^{(t)} - \sigma^{(t)}\lambda(-\frac{\mu_i^{(t)}}{\sigma^{(t)}}),$$

where $\lambda(z) = \frac{\phi(z)}{\Phi(z)}$ (the inverse of the Mills Ratio), and $-\sigma^{(t)}\lambda(-\frac{\mu_i^{(t)}}{\sigma^{(t)}})$ is the correction for censoring. Substituting ML estimates of the parameters yields the predicted values

$$\hat{y}_i^{(t)} = E(y_i|\hat{\theta}, x_i, y_i \leq 0) = \hat{\mu_i} - \hat{\sigma}\lambda(-\frac{\hat{\mu_i}}{\hat{\sigma}}),$$

for censored cases, where $\hat{\mu_i} = \hat{\beta_0} + \sum_{k=1}^{p} \hat{\beta}_k x_{ik}$. This model is sometimes called the Tobit model. An interesting extension of this model contains an incompletely observed variable (y_1) that has a linear regression on covariates and is observed if and only if the value of another completely unobserved variable (y_2) exceeds a threshold, such as zero. Little and Rubin (2014)

Heckman's Model.— Heckman's Sample Selection Model was created to handle censored or truncated dependent variables. Often, Heckman's model is used to handle non-ignorable incomplete data to control for sample selection biases that could arise from the existence of unobservable variables. These unobservable variables determine the association of sample selection bias to a particular response variable.
Consider the following equations:

$$\begin{aligned} y_{i2} &= x_{i2}\beta_2 + \epsilon_{i2}, \\ y_{i1} &= x_{i1}\beta_1 + \epsilon_{i1} \text{ if } y_{i2} > 0, \text{and} \\ y_{i1} &= \text{not observed if } y_{i2} \leq 0, \end{aligned}$$

where y_{i1} is the random variable of interest. However, it is not observed under all conditions and these conditions are specified by the dependent variable y_{i2}. y_{i1} is observed only when the corresponding value of y_{i2} is greater than 0. Without loss of generality, we can assume that out of a total of N observations, the first $i = 1, ..., r < N$ are complete. Here the y_{i1} values are known for them. The first equation is referred to as the selection equation since it determines whether a certain response is present in the survey or not.
The expectation for y_{i1} given the independent variables x_{i1} and that $y_{i2} > 0$.
$E[y_{i1}|x_{i1}, y_{i2} > 0] = x_{i1}\beta_1 + E[\epsilon_{i1}|\epsilon_{i2} > -x_{i2}\beta_2]$
The dependent variable y_{i1} is incompletely observed, and y_{i2} is never observed. The independent variable x_i is completely observed. Heckman (1976) The conditional density $f(y_i|x_i, \theta)$ is given by

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} x_{i1}\beta_1 \\ x_{i2}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right].$$

$x_{10} \equiv 1$ and $x_{20} \equiv 1$ are the constant terms. Also, ϵ_{i1} and ϵ_{i2} are distributed as

$$\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right].$$

Further $f(M_i|x_i, y_i, \theta)$ is given by

$$M_{i1} = \begin{cases} 1, & \text{if } y_{i2} \leq 0, \\ 0, & \text{if } y_{i2} > 0, \end{cases}$$
$$\text{and } M_{i2} \equiv 1.$$

The second stage of Heckman's model consists of calculating:

$$E[y_{i1}|M_{i1} = 0, x_{i1}] = x_{i1}\beta_1 + E[\epsilon_{i1}|M_{i1} = 0] = x_{i1}\beta_1 + E[\epsilon_{i1}|\epsilon_{i2} > -x_{i2}\beta_2]$$

We can express the expected value of ϵ_{i1} conditioned on ϵ_{i2} as

$$E[\epsilon_{i1}|\epsilon_{i2} > -x_{i2}\beta_2] = \rho\sigma_{\epsilon_{i1}}\sigma_{\epsilon_{i2}} \left[\frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)} \right]$$

where $\sigma_{\epsilon_{i1}}$ and $\sigma_{\epsilon_{i2}}$ are the variances from the OLS and probit models. Since $\sigma_{\epsilon_{i2}}$ is unidentified, it is set to 1. $\frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)}$ is the inverse Mills ratio, defined by the ratio of the density function of the normal distribution, ϕ, to its cumulative distribution function, Φ. When incorporated in the second step estimation of the response variable, this ratio serves as a control for potential biases arising from intentional inaccurate responders. Let the inverse Mills ratio be λ, then we have

$$E[y_{i1}|M_{i1} = 0, x_{i1}] = x_{i1}\beta_1 + \rho\sigma_{\epsilon_{i1}}\lambda_i$$

where ρ gives the covariance estimate of the unobserved effects on the indicator variable and the response variable. If significant, this estimate indicates that sample selection is present.

Bivariate Normal Stochastic Censoring Model.— Suppose y_{i1} is incompletely observed, y_{i2} is never observed, p covariates x_i are fully observed, and for case i, $f(y_i|x_i, \theta)$ is specified by

$$\begin{pmatrix} y_{i1} \\ y_{i2} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} x_{i1}\beta_1 \\ x_{i2}\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right]$$

where the constant term ($x_{i0} \equiv 1$) and predictors (x_{i1}, x_{i2}) for case i, β_1 and β_2 are $(p+1) \times 1$ vectors of regression coefficient. Further, let $M_i = (M_{i1}, M_{i2})$, where M_{ij} is the incomplete-data indicator for y_{ij}; $f(M_i|x_i, y_i, \theta)$ is specified by

$$M_{i1} = \begin{cases} 1, & \text{if } y_{i2} \leq 0, \\ 0, & \text{if } y_{i2} > 0, \end{cases}$$
$$\text{and } M_{i2} \equiv 1.$$

Since y_{i2} is always incomplete, we can integrate it out of the model and omit M_{i2}. From the above equations, the distribution of M_{i1} given y_{i1} and x_i is Bernoulli with probability of nonresponse

$$\begin{aligned} P(M_{i1} = 1|y_{i1}, x_i) &= P(y_{i2} \leq 0|y_{i1}, x_i) \\ &= 1 - \Phi \left[\frac{\mu_{i2} + \rho\sigma_1^{-1}(y_{i1} - \mu_{i1})}{\sqrt{1-\rho^2}} \right], \end{aligned}$$

where $\mu_{i1} = x_{i1}\beta_1, \mu_{i2} = x_{i2}\beta_2$. When $\rho \neq 0$ (that is y_1 and y_2 are correlated), this probability is a monotonic function of the values y_{i1}, which are sometimes incomplete, so the incomplete-data mechanism is non-igonorable. Little and Rubin (2014) This model was introduced by Heckman Heckman (1976) to describe selection of women into the labor force. The Tobit model of section 2.0.4 is obtained when $y_1 = \sigma_2 y_2$. Two estimation procedures have been proposed for this model, ML and the two-step method of Heckman. Heckman (1976) ML estimation by the EM algorithm for the case where no constraints are placed on the coefficients β_1, β_2, with hypothetical complete data defined as cases with both y_1 and y_2 completely observed. The complete-data sufficient statistics are then $\{\sum_i y_{i1}x_{ij}, \sum_i y_{i2}x_{ij}, \sum_i y_{i1}y_{i2}, \sum_i y_{i1}^2, \sum_i y_{i2}^2\}$ for $j = 0, 1, ..., p$. Since $\{x_{ij}\}$ are fully observed, the E step consists of replacing incomplete values of $y_{i1}, y_{i2}, y_{i1}y_{i2}, y_{i1}^2$, and y_{i2}^2 by their expectations given the parameters and the observed data. Properties of the bivariate normal distribution yield, for cases with y_{i1} incomplete:

$$\begin{aligned}
&E[y_{i2}|y_{i2} \leq 0] = \mu_{i2} - \lambda(-\mu_{i2}),\\
&E[y_{i1}|y_{i2} \leq 0] = \mu_{i1} - \rho\sigma_1\lambda(-\mu_{i2}),\\
&E[y_{i2}^2|y_{i2} \leq 0] = 1 + \mu_{i2}^2 - \mu_{i2}\lambda(-\mu_{i2}),\\
&E[y_{i1}^2|y_{i2} \leq 0] = \mu_{i1}^2 + \sigma_1^2 - \rho\sigma_1\lambda(-\mu_{i2})(2\mu_{i1} - \rho\sigma_1\mu_{i2}),\\
&E[y_{i1}y_{i2}|y_{i2} \leq 0] = \mu_{i1}[\mu_{i2} - \lambda(-\mu_{i2})] + \rho\sigma_1,
\end{aligned}$$

and for cases with y_{i1} observed:

$$\begin{aligned}
&E[y_{i2}|y_{i1}, y_{i2} > 0] = \mu_{i2\cdot 1} + \sqrt{1-\rho^2}\lambda(\frac{\mu_{i2\cdot 1}}{\sqrt{1-\rho^2}}),\\
&E[y_{i2}^2|y_{i1}, y_{i2} > 0] = 1 - \rho^2 + \mu_{i2\cdot 1}^2 + \mu_{i2\cdot 1}\sqrt{1-\rho^2}\lambda(\frac{\mu_{i2\cdot 1}}{\sqrt{1-\rho^2}}),
\end{aligned}$$

where conditioning on x_i and the parameters is implicit in these expressions, $\lambda(\cdot)$ is the inverse Mills ratio, as defined earlier, and $\mu_{i2\cdot 1} = \mu_i + \rho\sigma_1^{-1}(y_{i1} - \mu_{i1})$. Current values of the parameters are substituted to yield estimates for the E step.

The M step consists of the following computations, performed with complete-data sufficient statistics replaced by estimates for the E step:

1. Regress y_2 on x, yielding coefficients $\hat{\beta}_2$ of the response equation.
2. Regress y_1 on y_2 and x, yielding coefficients $\hat{\delta}$ for y_2 and $\hat{\beta}_1^*$ for x, and residual variance $\hat{\sigma}_{1\cdot 2}^2$.
3. Set $\hat{\beta}_1 = \hat{\beta}_1^* + \hat{\delta}\hat{\beta}_2, \hat{\sigma}_1^2 = \hat{\sigma}_{1\cdot 2}^2 + \hat{\delta}^2,$ and $\hat{\rho} = \frac{\hat{\delta}}{\hat{\sigma}_1}$.

Maximum Likelihood from Incomplete Data via the EM Algorithm for Finite Mixtures.— Likelihood theory based on the full likelihood for θ with non-ignorable data parallels that of ignorable nonresponse. Little and Rubin (2014) However, if a large portion of the data is incomplete, convergence to a maximum may be slow.

Suppose that an observable y is represented as n observations $y = (y_1, y_2, ..., y_n)$. Suppose further that there exists a finite set of R states, and that each y_i is associated with an unobserved state. Thus, there exists an unobserved vector $z = (z_1, z_2, ..., z_n)$, where z_i is the indicator vector of length R whose components are all zero except for one equal to 1 indicating the unobserved state associated with y_i. Defining the complete data to be $x = (y, z)$, Dempster et al Dempster et al. (1977) show the theory for EM applies for $f(x|\theta)$.

One way to understand mixture components is to think first of the marginal distribution of the indicators $\mathbf{z}$, and then to specify the distribution of $\mathbf{y}$ given $\mathbf{z}$. Assume that $(z_1, z_2, ..., z_n)$ are independently and identically drawn from a density $v(...|\theta)$. Further assume there is a set of R densities $u(...|\mathbf{r}, \theta)$ for $\mathbf{r} = (1, 0, ..., 0), (0, 1, 0, ..., 0), (0, ..., 0, 1)$ such that the y_i given z_i are conditionally independent with densities $u(...|z_i, \theta)$. Finally denoting

$$\mathbf{U}(\mathbf{y}|\theta) = (\log u(\mathbf{y_i}|(1, 0, ..., 0), \theta), \log u(\mathbf{y_i}|(0, 1, ..., 0), \theta), \log u(\mathbf{y_i}, (0, 0, ..., 1)|\theta))$$

and

$$\mathbf{V}(\theta) = (\log v((1,0,...,0)|\theta), \log v((0,1,...,0)|\theta), \log v((0,0,...,1)|\theta)),$$

we can express the complete-data log-likelihood as

$$\log f(\mathbf{x}|\theta) = \sum_{i=1}^{n} z_i^T \mathbf{U}(\mathbf{y_i}|\theta) + z_i^T \mathbf{V}(\theta).$$

Since The complete-data likelihood is linear in the components of each z_i, the E-step of the EM algorithm requires us to estimate the components of z_i given the observed y and the current fitted parameters. These estimated components of z_i are the current conditional probabilities that y_i belongs to each of the R states. In many examples, the θ parameters of $u(...|\theta)$ and $v(...|\theta)$ are unrelated, so the first and second terms may be maximized separately. The M-step is then equivalent to the complete-data maximization except that each observation y_i contributes to the log-likelihood associated with each of the R states, with weights given by the R estimated components of z_i, and the counts in the R states sum of the estimated components of the z_i. Dempster et al. (1977)

SEM Algorithm - Obtaining Asymptotic Variance-Covariance Matrices Meng & Rubin Meng and Rubin (1991) supplemented the EM algorithm with asymptotic variance-covariance matrix for estimates. The variance-covariance matrix obtained by SEM is based on the second derivatives of the observed-data log-likelihood and is asymptotically valid for inference. Let $y = (y_{obs}, y_{inc})$ be a data matrix partitioned into observed (y_{obs}) values and incomplete (y_{inc}) values. Suppose a model for the complete data y, with associated density $f(y|\theta)$, where $\theta = (\theta_1, ..., \theta_d)$. The EM algorithm defines a mapping $\theta \rightarrow M(\theta)$ from parameter space of θ, $\mathbf{\Theta}$, to itself such that

$$\theta^{(t+1)} = M(\theta^{(t)}), \text{for } t = 0, 1, ...$$

By Taylor series expansion in the neighborhood of θ^*,

$$\theta^{(t+1)} - \theta^* \approx (\theta^{(t)} - \theta^*)DM,$$

where

$$DM = \left(\frac{\partial M_j(\theta)}{\partial \theta_i} \right) \Big|_{\theta=\theta^*}$$

is the $d \times d$ Jacobian matrix for $M(\theta) = (M_1(\theta), ..., M_d(\theta))$ evaluated at $\theta = \theta^*$. Once the complete-data observed information matrix,

$$I_O(\theta|\mathbf{y}) = -\frac{\partial^2 \log f(\mathbf{y}|\theta)}{\partial\theta \cdot \partial\theta}$$

is obtained the complete-data variance-covariance matrix can be found by taking the inverse of $I_O(\theta|\mathbf{y})$. Compute the expectation over the conditional distribution $f(y_{inc}|y_{obs}, \theta)$ evaluated at $\theta = \theta^*$:

$$I_{OC} = E[I_O(\theta|\mathbf{y})|y_{obs}, \theta]\Big|_{\theta=\theta^*}$$

The matrix I_{OC}^{-1} is important because the observed variance-covariance matrix, V, can be written as a function of I_{OC}^{-1} and DM, the matrix rate of convergence of EM. Thus,

$$V = I_{OC}^{-1} + \Delta V,$$

where

$$\Delta V = I_{OC}^{-1} DM (I - DM)^{-1}$$

is the increase in variance due to incomplete information, and I is the $d \times d$ identity matrix. Meng and Rubin (1991)

2.0.5 Summary

Currently for statistical inferences with intentional inaccurate responses, inaccurate responses are ignored, removed, or a probability-based index is calculated and added to the model to adjust for the bias. In order to provide unbiased estimates, inaccurate responses cannot be removed, but the inaccuracy must be considered. In order to consider this inaccuracy, the incorrect responses could be handled similarly to incomplete data. This dissertation will take incomplete data techniques, in order to account for the non-ignorable inaccurate responses.

Chapter 3

Analytical Approaches

The Probability-Based Index and each of the incomplete data techniques from Chapter 2 will be discussed and applied to an intentional inaccurate response situation.
Consider a scenario of self-report data that includes a sensitive question that identifies a participant as a smoker or nonsmoker. Let those participants who answered ”smoker” answered truthfully and those participants who answered ”nonsmoker” were either truthful or intentionally inaccurate. Assume each participant had their blood analyzed to determine their progesterone level. The inaccurate responders in the data have been identified.

3.0.6 Probability-Based Index

Applying this method to the intentional inaccurate responders example, a probability-based index would be calculated for each participant based on a set of covariates. This probability-based index would be included in the regression model along with the self-reported smoking status variable to determine the effect of smoking on progesterone.

3.0.7 Non-ignorable Intentional Inaccurate Response Overview

Consider different levels of intentional inaccuracy. Let $y = (y_{hon}, y_{ina})$ be a data matrix partitioned into honest (y_{hon}) values and inaccurate (y_{ina}) values. We can denote the inaccurate response data indicator as:
If $y = y_{ij}$ an $n \times K$ matrix of n observations measured for K variables, then

$$I_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ inaccurate} \\ 0, & \text{if } y_{ij} \text{ honest} \end{cases}$$

Looking at the complete data y and the inaccurate response data indicator matrix $I = I_{ij}$, the inaccurate response data mechanism is defined by the conditional distribution of I given y, $f(I|y, \theta)$, where θ denotes the unknown parameters. Table 3.1 gives an explanation of each data mechanism.

Table 3.1: Inaccurate-Response Mechanisms

Inaccurate-Response Mechanism	**Explanation**
Inaccurate Completely At Random (ICAR)	If inaccuracy does not depend on the values of the data y, inaccurate or honest: $f(I|y, \theta) = f(I|\theta)$ for all y, θ
Inaccurate At Random (IAR)	If inaccuracy depends only on the components y_{hon} of y that are honest, and not on the components that are inaccurate: $f(I|y, \theta) = f(I|y_{hon}, \theta)$ for all y_{ina}, θ
Not Inaccurate At Random (NIAR)	If the distribution of I depends on the inaccurate values in the data matrix y.

3.0.8 Non-Ignorable Inaccurate Data

With some datasets, the incorrect responses have a relationship to intentional behavior, similar to incomplete values having a relationship to the non-response. In this case, the inaccurate responses are non-ignorable. There is a pattern to the inaccurate responses that must be modeled. Since the inaccurate responders are not inaccurate at random, the likelihood estimation needs a model for the inaccurate-response mechanism and maximize the full likelihood. Below are examples of handling non-ignorable inaccurate responses.

3.0.9 Pattern-Mixture and Selection Models for Univariate Inaccurate Response

For an inaccurate response example, let $J = 3$ and y_i be a smoking score assigned to each participant. For $i = 1, ..., r < n$, y_i are values from participants who were truthful about their smoking status (those who claimed to be non-smokers were actually non-smokers and those who claimed to be smokers were actually smokers). Let the inaccurate-response indicator

$$I_i = \begin{cases} 0, & \text{if } y_i \text{ is from a truthful respondent} \\ 1, & \text{if } y_i \text{ falls in the } j^{th} \text{ inaccurate category } (a_j < y_i < b_j) \; j = 1, ..., J \end{cases}$$

In this illustration, $I_i = 1$ if y_i falls into any of the three intervals: $(-18, -16), (-16, -14), (-14, -12)$. Presently, when $J = 1$, $a_1 = -18$ and $b_1 = -16$.

3.0.10 Censored Normal Data with Covariates - Tobit Model

Consider a partially honest variable y_1 that has a linear regression on covariates and is honest if and only if the value of another completely inaccurate variable y_2 exceeds a threshold, such as -12. With inaccurate responses, the above example can be used when values of y are fully honest but values below -12 are censored. For instance, censored values will lie anywhere in the interval $(-\infty, -12)$. All truthful y_i are greater than -12, $J = 1, a_1 = - = \infty$, and $b_1 = -12$.

3.0.11 Heckman's Model

To compare Heckman's model with inaccurate response, let y_{i1} be progesterone levels and y_{i2} be smoking scores. The inaccurate-response indicator is

$$I_{i1} = \begin{cases} 1, & \text{if } y_{i2} \leq 0 \\ 0, & \text{if } y_{i2} > 0 \end{cases}$$
$$I_{i2} \equiv 1.$$

Heckman's model is widely used by social scientists and the convenience of Heckman's model is that a procedure in SAS already exists, proc qlim, and can be calculated easily. However, the selection score must be continuous in order to use Heckman's model. Also, if the selection equation does not capture the inaccurate responses, parameter estimates for the OLS equation will be biased.

3.0.12 Bivariate Normal Stochastic Censoring Model

For the intentional inaccurate response example, use the example from Heckman's model and solve using the EM algorithm.

3.0.13 Maximum Likelihood from Partial Inaccurate Data via the EM Algorithm

Assume intentional inaccurate responders have a different distribution than the participants who told the truth. The expectation of the smokers and nonsmokers are maximized iteratively until the estimate becomes stable at $1e - 7$. To accomplish this, the intentional inaccurate responses are substituted by estimated values. Then mean and variance of the response variable are estimated for the smokers and the nonsmokers, followed by an re-estimation of the intentional inaccurate responses assuming the new mean and variance estimates are correct. The re-estimating continues until convergence is met. The EM algorithm is not filling in values for the intentional inaccurate responses, but using the functions of the intentional inaccurate responses in the complete-data log-likelihood. Little and Rubin (2014)

3.0.14 Summary

A variety of methods for handling inaccurate responders have been proposed. Below, Table 3.2 summarizes when to use each method.

Table 3.2: Handling Inaccurate-Response Summary

Inaccurate-Response Method	**Proposed When to Use**
Pattern-Mixture and Selection Models for Univariate Nonresponse	Use when inaccurate responders have a different distribution than truthful responders for a single dependent variable.
Grouped Normal Data with Covariates	Use when inaccurate responses fall into specific interval values for a dependent variable.
Censored Normal Data with Covariates	Use when inaccurate responses fall above or below a certain value, so that, the inaccurate responses can be censored for a dependent variable.
Heckman's Model	Use when inaccurate responses for a dependent variable fall above or below a certain value, and use this selection criteria to choose responses used to analyze another dependent variable. Parameters are estimated using maximum likelihood.
Bivariate Normal Stochastic Censoring Model	Use when inaccurate responses for a dependent variable fall above or below a certain value, and use this selection criteria to choose responses used to analyze another dependent variable. Parameter are estimated using EM algorithm.

Chapter 4

Simulation Study

Self-report data are often collected in epidemiology, psychology, pediatrics, and social and behavioral science settings. Siggeirsdottir et al. (2007) Bernard et al. (1984) Sherry et al. (2007) Fan et al. (2006) Self-report data are define as data collected from surveys containing questions that ask respondents to report something about themselves. Chan (2009) Unfortunately, these data have the potential to contain inaccurate responses due to 'carelessness, confusion, lack of efforts, or intentional mischief' Fan et al. (2006). This issue has been seen in multiple discipliness Sherry et al. (2007) Siggeirsdottir et al. (2007) Fan et al. (2006) Robinson-Cimpian (2014). Sometimes Quality of Life and Patient Reported Outcomes can only be collected through surveys; therefore, members of the health and medical fields frequently use self-report data collection methods in order to analyze their research. In particular, surveys have become widely accepted among professional and policy-making organizations Fan et al. (2006) Cornell et al. (2012).

Although data quality has been a long time topic of surveys, response validity has been giver greater attention in survey literature. Bernard et al. (1984) Fan et al. (2006) Cornell et al. (2012) Robinson-Cimpian (2014) Bernard et al Bernard et al. (1984) looked at respondent accuracy in the areas of child care behavior, health seeking behavior, and communication and social interactions. They concluded that in all the studies they examined, on average, about half of what a participant answers is incorrect in some form.

Respondents who answer inaccurately have been labeled as 'joksters' Fan et al. (2006) or 'mischievous responders' Robinson-Cimpian (2014). While it is possible to use data responses to identify intentional inaccurate responders Robinson-Cimpian (2014), there is no clear method for how to handle these respondents and the impact on statistical inferences.

The motivating clinical example of intentional inaccurate responses occurs in data using a sample of pregnant women to determine if smoking status effects progesterone levels. Pregnant women self-report their smoking status as nonsmoker or smoker. Due to the sensitive nature of asking about smoking status during pregnancy and that smoking while pregnant is highly discourage by health care providers, participants may consciously minimize their actual cigarette use. Klesges et al. (1995) As the adverse affects of smoking while pregnant are widely known Lumley et al. (2009), higher levels of nonsmoking is expected. In this case, the intentional inaccurate responses are not given at random. There is a pattern to the inaccuracy. This, in turn, will bias the data towards nonsmoking and could lead to inaccurate conclusions.

Previous startegies for accounting for intentional inaccurate responses included ignoring the inaccuracy and analyzing all self-report data or removing all inaccuracies from the dataset. Cornell et al. (2012) Other methods involve using a Probability-Based Index Robinson-Cimpian (2014) as a covariate in regression. We propose extending incomplete data methodology to appropriately account for intentional inaccuracies. These methods includes Heckman's model and the EM algorithm for finite mixtures.

To compare strategies for handling intentional inaccurate responders, a simulation study was conducted based the motivating pregnant women example. The simulations were ran with varying samples sizes, probabilities of an intentional inaccurate responder, and coefficients of variation.

This chapter is structured as follows. Section 2 describes current and proposed methods for handling intentional inaccurate responders in statistical analyses. Section 3 reports the results from the simulation study. Section 4 provides a conclusion and general discussion of appropriate usage of the EM algorithm.

4.0.15 Motivating Example

Determining Smoking Status.— Data were collected to better understand prenatal passive smoke exposure and birth outcomes. Participants self-reported their smoking status by answering the question 'Do you currently smoke cigarettes or use smokeless tobacco (loose leaf, dip, chew, snuff) even just once in a while?'. Based on their self-report response to this questions, participants were labeled as smokers or nonsmokers. Besides their self-reported answers from the survey, participants agreed to supply a urine sample. Smoking was confirmed by urine cotinine using $NicAlert^{TM}$ strips. According to $NicAlert^{TM}$, lab values ranged from 0 to 6. Level 3 or higher (level 4, 5, or 6) indicated use of tobacco products, level 0 indicated no detectable level of cotinine or tobacco product use, and levels 1 and 2 indicated no use of tobacco products. Based on these classifications, a participant whose $NicAlert^{TM}$ value was 0-2 was considered a non-smoker and a participant whose $NicAlert^{TM}$ value was 3-6 was considered a smoker. From these lab values, participants were labels as smokers or nonsmokers.

Comparing the self-report values to the lab classification suggested there were intentional inaccurate responders in the data, Table 4.1. Of the women who claimed to be nonsmokers, 89(89.9%) were nonsmokers, but 10(10.1%) of the women were actually smokers. Because the self-report and lab values did not result in classification agreement, these 10 women were considered as intentional inaccurate responders.

Table 4.1: Accuracy of Smoking Status

Self-Report	Lab Value		
Frequency	Non-Smoker	Smoker	Total
Non-Smoker	89	10	99
Smoker	0	9	9
Total	89	19	108

Question of Interest.— Leaving the intentional inaccurate responses in the data, removing them, or adjusting for them in some way may affect how the self-report smoking status predicts progesterone level. This chapter examines the impact of intentional inaccurate responders on the relationship between smoking status and progesterone level. To answer this question, a variety of methods will be engaged. First, traditional linear regressions are run, followed by proposed methods to account for intentional inaccurate responders, such as, probability-based index, Heckman's model, and the EM algorithm. All analyses were complete using R 3.2.2 (R Core Team, Vienna, Austria). Significance is defined as $p < 0.05$.

4.0.16 Design of Simulation Study

Data were simulated to represent the true smokers, the true nonsmokers, and the intentional inaccurate nonsmokers. In the smoking exposure and birth outcomes data, there were close to 110 participants and of the 99 self-reported nonsmokers, about 10% were inaccurate. Using sampling with replacement, 1000 simulations of sample sizes of 55, 110, 550, and 1100 each with the probability of an inaccurate responder of .10, .20, and .40, and coefficients of variation of 0.3 and 23.3. 1000 simulations was found to be sufficient in determining the effect of intentional inaccurate responders. Burton et al. (2006) Sample size of 110 was based on the motivating example. Sample sizes were decrease and increase to compare sample size effect on the varying models. Probabilities were increased from 0.10 to 0.40 to see the methods performance under increased probability of inaccurate response. In order to determine if models were sensitive to variability, small and large coefficients of variation were simulated. Independent datasets were created in each simulation. Results from methods using only self-report values were compared with the lab values to examine the bias and variability of each statistical method.

Progesterone levels were assumed to come from a normal distribution. Since the inaccurate nonsmokers and true nonsmokers appeared to come from 2 different normal distributions, the self-report nonsmoker values were generated using a pattern mixture of normal distributions. The smokers were simulated so that their progesterone values came from a $N(40, 4)$ and $N(40, 325)$. The inaccurate nonsmokers had their progesterone levels simulated from $N(50, 5)$ and $N(50, 350)$, and honest

Table 4.2: Simulation Design - *Small CV is 0.3 and large CV is 23.3

Sample Size	Probability of Intentional Inaccurate Response	CV*
55	0.10	Small
		Large
	0.20	Small
		Large
	0.40	Small
		Large
110	0.10	Small
		Large
	0.20	Small
		Large
	0.40	Small
		Large
550	0.10	Small
		Large
	0.20	Small
		Large
	0.40	Small
		Large
1100	0.10	Small
		Large
	0.20	Small
		Large
	0.40	Small
		Large

nonsmokers from $N(60, 6)$ and $N(60, 375)$. The first set of normal distributions in each grouping represent a small CV and the second set represents a large CV. Table 4.2 shows the simulation design.

4.0.17 Intentional Inaccurate Response - Current Methods

A linear regression was run using the lab values to determine if smoking status predicted progesterone level. This model represents the truth and is used as the basis of comparison for all other models, current and proposed.

First current method, inaccurate responses are ignored and all the self-report values are analyzed. Here, a linear regression using all the self-report smoking status values to predict progesterone level was employed. The second current method consists of identifying and removing inaccurate responses from the data before analysis Cornell et al. (2012). Self-report values with the inaccurate responses removed were used as an independent variable to predict progesterone level.

4.0.18 Intentional Inaccurate Response - Proposed Methods

Intentional inaccurat responses can be accounted for similarly as incomplete data. These inaccurate responses are not due to a data mistake or carelessness, which might be considered random inaccuracy. Here participants are purposefully selecting the incorrect response. Since intentional inaccurate responses have a relationship to the self-report smoking status, this is comparable to incomplete values having a relationship to the non-response. Non-ignorable incomplete-data models are models where data are not incomplete at random and the incomplete data are modeled. Likewise, the intentional inaccurate data must be modeled since the data are not inaccurate at random. Three alternative methods are proposed for accounting for intentional inaccurate responses:

Probability-Based Index, Heckman's model, and the EM algorithm.

Probability-Based Index.— The probability-based index was introduced by Robinson-Cimpian in 2014 Robinson-Cimpian (2014). In his paper, he used the screener index as a way to measure the amount of low-frequency responses on screener items. This measurement was similar to a weight that was included as a covariate in his model to account for intentional inaccurate responders. With the simulated data, the probability-based index was used to measure the amount of inconsistency with the self-report smoking status. This method was run two different ways. The first uses a probability-based index that does not capture the inaccurate responses well. The second index was created so that those participants who told the truth were given more weight than those who were inaccurate. Probability-Based Index values were calculated for each participant and were used as a covariate in the regression equation predicting progesterone level in order to adjusted for the inaccurate responses.

Heckman's Model.— Heckman's Sample Selection Model was created to handle censored or truncated dependent variables and is often used to handle non-ignorable incomplete data. With the simulated data, Heckman's model is used to control for biases that could appear from the existence of inaccurate responses. Heckman's Model is a two-step method. Consider the following equations.

$$\begin{aligned} y_{i1} &= x_{i1}\beta_1 + \epsilon_{i1}, \\ y_{i2} &= x_{i2}\beta_2 + \epsilon_{i2}. \end{aligned}$$

In the first step, which involves the selection equation, a dichotomous variable is defined to indicate which group the participant is categorized:

$$\begin{aligned} y_{i2} &= x_{i2}\beta_2 + \epsilon_{i2} \\ y_{i2}* &= \begin{cases} 1, \text{if } y_{i2} \geq 0 \\ 0, \text{if } y_{i2} < 0, \end{cases} \end{aligned}$$

where y_{i2} is a smoking score, $y_{i2}*$ is an indicator for truth status, the x_{i2} are the explanatory variables of the smoking score, β_2 is a vector of parameter estimates, and ϵ_{i2} is an error term having a standard normal distribution. The first stage estimates β_2 using the probit maximum likelihood method. The second stage involves estimating an OLS regression of the response variable, y_{i1}, conditional on $y_{i2}*$ and x_{i1}. Vance and Buchheim (2005) Therefore,

$$E[y_{i1}|y_{i2}* = 1, x_{i1}] = x_{i1}\beta_1 + E[\epsilon_{i1}|y_{i2}* = 1] = x_{i1}\beta_1 + E[\epsilon_{i1}|\epsilon_{i2} \geq -x_{i2}\beta_2]$$

where y_{i1} is the progesterone levels and x_{i1} is the self-report smoking status variable of y_{i1}, β_1 is a vector of parameter estimates, and ϵ_{i1} is an error term having a standard normal distribution. Let ϵ_{i1} and ϵ_{i2} be correlated by ρ. Now,

$$E[y_{i1}|y_{i2}* = 1, x_{i1}] = x_{i1}\beta_1 + \rho\sigma_1\lambda_i$$

where σ_1 are the error variances of the OLS model, λ_i is the inverse Mills ratio such that

$$\lambda_i = \frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)}$$

where $\frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)}$ is defined by the ratio of the density function of the normal distribution, ϕ, to its cumulative distribution function, Φ.

Participants were identified as truthful or inaccurate and a smoking score was created. This smoking score, y_{i2}, was a continuous variable and was the sum of each participants cotinine level and intentional inaccurate responder status. Heckman's model was run predicting the smoking score two ways. The first y_{i2} model did not predict smoking status well, but the second did. The cutoff value for $y_{i2}*$ was chosen so that the majority of predicted y_{i2} values were appropriately identifying participants as truthful or intentionally inaccurate. Once x_{i2} were found, these variables were used in the second step of Heckman's model to estimate λ_i.

The EM Algorithm.— Dempster et al Dempster et al. (1977) introduced the EM algorithm for finite mixtures. Little & Rubin Little and Rubin (2014) used the EM algorithm to account for non-ignorable incomplete data. Non-ignorable incomplete data models are models where data are not incomplete at random. This means that the maximum likelihood estimation requires a model for the incomplete data.

The smoking status example, the inaccurate responses are non-ignorable. There is a pattern to the inaccurate responses that must be modeled. Since the intentional inaccurate responders are not inaccurate at random, the likelihood estimation needs to model the intentional inaccurate responses. The simulated data are instances when intentional inaccurate responders' progesterone levels seem to have a different distribution than truthful responders; therefore, the EM algorithm for finite mixtures was used. Consider two states for the self-reported smoking status. For example,

$$\begin{aligned} z_1 &= (1,0) \rightarrow \text{intentionally inaccurate} \\ z_2 &= (0,1) \rightarrow \text{honest} \end{aligned}$$

Assume that (z_1, z_2) are independently and identically drawn from a density $Bern(r|\pi)$ for $r = (1,0),(0,1)$ where there is a set of R densities $u(y_i|(r,\theta))$. Let y be continuous measurements and n be the number of participants to self-report as nonsmokers. Using the EM methods from Dempster et al Dempster et al. (1977),

$$\begin{aligned} z_i &\sim Bern(r|\pi) \\ y_i|z_i &\sim u(y_i|r,\theta) \\ U(y|\theta) &= (\log u(y_i|(1,0),\theta), \log u(y_i|(0,1),\theta)) \\ V(\pi) &= (\log Bern((1,0)|\pi), \log Bern((0,1)|\pi)) \end{aligned}$$

Now, the log-likelihood for self-reported nonsmokers is

$$\log f_{NS}(x|\theta,\pi) = \sum_{i=1}^{n} z_i^T U(y_i|\theta) + z_i^T V(\pi)$$

Therefore, let m be the number of self-reported smokers and y be progesterone levels. Assume the y_i are independently and identically drawn from a density $w(...|\theta)$. Let

$$W(y|\theta) = \log w(y|\theta).$$

Now, the log-likelihood for the self-reported smokers is

$$\log f_S(x|\theta) = \sum_{j=1}^{m} W(y_i|\theta).$$

Thus, the complete data log-likelihood is

$$\log f(x|\theta, \pi) = \sum_{j=1}^{m} W(y_i|\theta) + \sum_{i=1}^{n} z_i^T U(y_i|\theta) + z_i^T V(\pi).$$

Using this complete data log-likelihood, the EM algorithm was used to calculate an estimate of the true mean for smokers (μ_0), the true mean for nonsmokers (μ_1), and the probability of being a intentional inaccurate responder (π).
Meng & Rubin Meng and Rubin (1991) supplemented the EM algorithm with asymptotic variance-covariance matrix for estimates. All standard errors from the EM algorithm were calculated using the SEM.
In this approach, $z_i \sim Bern(r|\pi)$; however, it can be extended to a logistic distribution that estimates more parameters.
The expectation of the smokers and nonsmokers were maximized iteratively until the estimate became stable at $1e-7$.

4.0.19 Results

Results from the simulations can be seen in Table 4.3, Table 4.4, Table 4.5, and Table 4.6. The average estimated difference in progesterone level difference between smokers and nonsmokers for lab values, and the 95% confidence intervals represent the truth.

4.0.20 Current Methods

Ignoring the intentional inaccurate responses and using all the self-report values consistently showed a larger progesterone difference among smokers and nonsmokers. This is probably due to the fact the average progesterone levels for smokers seemed to be very low, and average progesterone for non-smokers was a little lower than it should be, resulting in a larger difference. Of the 24 simulations, only 9 contained the lab value estimate in the confidence interval. However, 6 of those confidence intervals, when sample size were 55 and 110 with large CV, were quite wide so there was no surprise that the lab value estimates were contained in them. Only when the CV was large did the confidence intervals for lab values and self-report values have any overlap.
When removing the inaccurate self-report values, the estimated progesterone difference was also always larger than the lab values estimate. This is not surprising, since the smokers average progesterone was very low and the non-smokers average progesterone was very high. As a result, the average differnce was bigger. In 9 simulations did the lab value estimate appear in the inaccurate response removed confidence intervals, and only 10 times did the confidence intervals show any overlap. Once again, when sample sizes were 55 and 110 with large CV, there were 6 large confidence intervals, see Table 4.3.
As far as determining statistical significance, self-report values were consistent in finding statistical significance, except for when the sample size was 55 or 110 with large CV. For removing inaccurate response, this analysis underreported significance similarly to self-report values; however, removing inaccurate responses did over-report with the sample size was 110, probability of intentional inaccurate response was 0.10, and large CV. This can be seen in Table 4.7.

4.0.21 Probability-Based Index

For the first version of the probability-based index method, the estimated progesterone difference between smokers and nonsmokers was positive with large confidence intervals when the probability of an inaccurate response was high. In fact, for version 1, the confidence intervals contained positive estimates whenever the CV was large and the sample size was small, 55 and 110. Since version 1 was specifically created to poorly identify inaccurate responders, the result gave higher average progesterone levels to smokers.

Table 4.3: Simulation Results for Estimated $\mu_S - \mu_{NS}$ (95% Confidence Intervals) - Current Methods

Sample Size	π	CV	Lab Values	All Self Report Values	Intentional Inaccurate Responses Removed
55	0.10	Small	-15.20 (-17.96,-12.44)	-18.98 (-21.11,-16.84)	-19.99 (-21.97,-18.00)
		Large	-15.00 (-28.89,-1.11)	-19.12 (-36.83,-1.42)	-20.09 (-37.94,-2.24)
	0.20	Small	-13.47 (-15.40,-11.53)	-17.99 (-20.24,-15.73)	-20.00 (-22.01,-17.99)
		Large	-13.54 (-25.20,-1.88)	-18.15 (-35.22,-1.07)	-20.17 (-37.54,-2.81)
	0.40	Small	-12.05 (-13.42,-10.67)	-16.02 (-18.40,-13.64)	-19.98 (-21.98,-17.99)
		Large	-11.79 (-22.03,-1.55)	-15.42 (-32.34,1.50)	-19.38 (-36.69,-2.07)
110	0.10	Small	-15.12 (-16.96,-13.27)	-18.99 (-20.47,-17.52)	-20.00 (-21.37,-18.63)
		Large	-15.13 (-24.43,-5.83)	-18.92 (-31.15,-6.69)	-19.92 (-32.20,-7.63)
	0.20	Small	-13.37 (-14.75,-11.99)	-17.99 (-19.55,-16.44)	-19.99 (-21.34,-18.64)
		Large	-13.53 (-21.79,-5.26)	-17.94 (-29.61,-6.27)	-19.99 (-31.82,-8.15)
	0.40	Small	-12.03 (-13.03,-11.02)	-16.03 (-17.75,-14.32)	-20.02 (-21.48,-18.57)
		Large	-12.25 (-19.58,-4.92)	-15.80 (-28.08,-3.52)	-19.92 (-32.52,-7.32)
550	0.10	Small	-15.00 (-15.81,-14.18)	-18.99 (-19.66,-18.32)	-19.99 (-20.60,-19.38)
		Large	-15.10 (-19.15,-11.05)	-19.06 (-24.25,-13.87)	-20.07 (-25.33,-14.81)
	0.20	Small	-13.36 (-13.92,-12.79)	-18.00 (-18.67,-17.34)	-20.00 (-20.61,-19.40)
		Large	-13.33 (-16.80,-9.86)	-17.95 (-23.32,-12.59)	-19.96 (-25.31,-14.61)
	0.40	Small	-12.00 (-12.42,-11.57)	-16.00 (-16.75,-15.24)	-19.99 (-20.63,-19.36)
		Small	-11.93 (-15.11,-8.76)	-15.82 (-21.10,-10.54)	-19.82 (-25.24,-14.40)
1100	0.10	Small	-15.00 (-15.57,-14.43)	-19.00 (-19.46,-18.53)	-20.00 (-20.43,-19.56)
		Large	-15.07 (-18.03,-12.11)	-19.04 (-22.91,-15.18)	-20.05 (-23.93,-16.18)
	0.20	Small	-13.33 (-13.73,-12.94)	-17.99 (-18.49,-17.50)	-20.00 (-20.43,-19.57)
		Large	-13.31 (-15.89,-10.72)	-17.98 (-21.63,-14.33)	-19.97 (-23.72,-16.22)
	0.40	Small	-11.99 (-12.29,-11.70)	-16.00 (-16.52,-15.48)	-20.00 (-20.44,-19.56)
		Large	-11.99 (-14.28,-9.71)	-15.98 (-19.75,-12.21)	-19.98 (-23.85,-16.11)

Table 4.4: Simulation Results for Estimated $\mu_S - \mu_{NS}$ (95% Confidence Intervals) - Probability Based Index

Sample Size	π	CV	Lab Values	Prob Based Index v1	Prob Based Index v2
55	0.10	Small	-15.20 (-17.96,-12.44)	-9.96 (-13.39,-6.52)	-19.99 (-21.97,-18.00)
		Large	-15.00 (-28.89,-1.11)	-10.46 (-35.50,14.57)	-20.09 (-37.94,-2.24)
	0.20	Small	-13.47 (-15.40,-11.53)	-8.04 (-11.85,-4.23)	-20.00 (-22.01,-17.99)
		Large	-13.54 (-25.20,-1.88)	-8.15 (-29.69,13.40)	-20.17 (-37.54,-2.81)
	0.40	Small	-12.05 (-13.42,-10.67)	4.49 (-57.11,66.08)	-19.98 (-21.98,-17.99)
		Large	-11.79 (-22.03,-1.55)	3.69 (-70.02,77.40)	-19.38 (-36.69,-2.07)
110	0.10	Small	-15.12 (-16.96,-13.27)	-9.94 (-12.08,-7.80)	-20.00 (-21.37,-18.63)
		Large	-15.13 (-24.43,-5.83)	-9.79 (-26.52,6.93)	-19.92 (-32.20,-7.63)
	0.20	Small	-13.37 (-14.75,-11.99)	-8.22 (-10.79,-5.65)	-19.99 (-21.34,-18.64)
		Large	-13.53 (-21.79,-5.26)	-7.91 (-22.96,7.13)	-19.99 (-31.82,-8.15)
	0.40	Small	-12.03 (-13.03,-11.02)	9.61 (-54.35,73.57)	-20.02 (-21.48,-18.57)
		Large	-12.25 (-19.58,-4.92)	11.99 (-69.57,93.55)	-19.92 (-32.52,-7.32)
550	0.10	Small	-15.00 (-15.81,-14.18)	-10.00 (-10.93,-9.07)	-19.99 (-20.60,-19.38)
		Large	-15.10 (-19.15,-11.05)	-9.96 (-16.85,-3.07)	-20.07 (-25.33,-14.81)
	0.20	Small	-13.36 (-13.92,-12.79)	-8.31 (-9.34,-7.28)	-20.00 (-20.61,-19.40)
		Large	-13.33 (-16.80,-9.86)	-8.26 (-15.25,-1.27)	-19.96 (-25.31,-14.61)
	0.40	Small	-12.00 (-12.42,-11.57)	6.07 (-4.93,17.07)	-19.99 (-20.63,-19.36)
		Large	-11.93 (-15.11,-8.76)	6.72 (-10.16,23.60)	-19.82 (-25.24,-14.40)
1100	0.10	Small	-15.00 (-15.57,-14.43)	-10.00 (-10.62,-9.37)	-20.00 (-20.43,-19.56)
		Large	-15.07 (-18.03,-12.11)	-9.97 (-15.15,-4.79)	-20.05 (-23.93,-16.18)
	0.20	Small	-13.33 (-13.73,-12.94)	-8.31 (-9.11,-7.52)	-20.00 (-20.43,-19.57)
		Large	-13.31 (-15.89,-10.72)	-8.35 (-12.79,-3.92)	-19.97 (-23.72,-16.22)
	0.40	Small	-11.99 (-12.29,-11.70)	5.58 (-1.12,12.29)	-20.00 (-20.44,-19.56)
		Large	-11.99 (-14.28,-9.71)	5.54 (-4.08,15.16)	-19.98 (-23.85,-16.11)

For version 2, typically greater differences were found compared to lab values. Here, inaccurate responders had very low Probability-Based Indices, giving more weight to the truthful responders and resulting in larger differences. There were 7 simulations where the lab values estimate were contained in the probability-based index confidence intervals, all were large CVs and primarily with smaller sample sizes. Of these 7, 6 of the confidence intervals were very wide. Confidence intervals overlapped with lab value confidence intervals in 10 of the simulations.

Also, the probability based index verion 1 consistently underreported significance compare to the lab values; however, version 2 only underreported 4 times when the CV was large and over-reported when the sample size was 110, probability of intentional inaccurate response was 0.10, and large CV. See Table 4.4 and Table 4.7, respectively.

4.0.22 Heckman's Model

The results from Heckman's model can be seen in Table 4.5. Using both versions of Heckman's model to analyze the self-report data, the estimated progesterone differences among self-reported smokers and nonsmokers were consistently larger than the lab value estimates. Once again, version 1 was specifically created to poorly identify inaccurate responses. As a result, version 1 estimated a difference too large. Version 2 did have larger differences; however, the estimates were not as biased as the previous methods. In both versions, when the CV was large and the sample size was 55, the confidence intervals contained positive estimates and were wide. Of the 24 simulations, version 1 contained the lab value estimate in the confidence interval 9 times; whereas, version 2 contained the lab estimate 16 times.

In Table 4.8, compared to the lab values, both versions of Heckman's model underreported statistical significance when the CV is large and the sample size is either 55 or 110.

4.0.23 The EM Algorithm

The estimate for average difference in progesterone levels using lab values appeared in every EM algorithm confidence interval when sample sizes were 55 and 110. Lab value estimates also appeared with confidence intervals when the sample size was 550 and the probability of an intentional inaccurate response was 0.40. All lab value confidence intervals overlapped with the EM algorithm confidence intervals, Table 4.6. The EM algorithm consistently underreported statistical significance when compared to the lab values, especially when the CVs were large and sample sizes were small, see Table 4.8.

The EM algorithm is know to have convergence issues, see Table 4.9. Clearly, when the CV is large and the probability of inaccurate response is high, the EM algorithm struggles.

The EM algorithm also estimates the probability of an intentional inaccurate response. When the CV is small, the EM algorithm consistently gives estimates close to that of the true probability. However, when the CV is large, the EM algorithm is not as strong, especially with the probability of an intentional inaccurate response is low, see Table 4.10.

4.0.24 Simulation Graphs

Figures 4.1 - 4.24 show the distribution of the $\mu_S - \mu_{NS}$ estimates for all of the methods in each simulation. The bold red line represents the lab value estimate. The EM algorithm distribution is consistently within the lab value confidence interval.

Table 4.5: Simulation Results for Estimated $\mu_S - \mu_{NS}$ (95% Confidence Intervals) - Heckman's Model

Sample Size	π	CV	Lab Values	Heckman's v1	Heckman's v2
55	0.10	Small	-15.20 (-17.96,-12.44)	-19.79 (-22.23,-17.34)	-17.66 (-20.61,-14.71)
		Large	-15.00 (-28.89,-1.11)	-20.24 (-40.69,0.22)	-18.05 (-37.46,1.36)
	0.20	Small	-13.47 (-15.40,-11.53)	-19.25 (-21.81,-16.69)	-15.96 (-19.03,-12.89)
		Large	-13.54 (-25.20,-1.88)	-19.29 (-39.37,0.79)	-16.02 (-34.37,2.32)
	0.40	Small	-12.05 (-13.42,-10.67)	-17.53 (-20.34,-14.73)	-13.38 (-16.80,-9.96)
		Large	-11.79 (-22.03,-1.55)	-16.88 (-35.41,1.65)	-12.81 (-30.89,5.27)
110	0.10	Small	-15.12 (-16.96,-13.27)	-19.73 (-21.34,-18.12)	-17.72 (-19.71,-15.74)
		Large	-15.13 (-24.43,-5.83)	-19.63 (-34.24,-5.01)	-17.74 (-30.70,-4.77)
	0.20	Small	-13.37 (-14.75,-11.99)	-19.15 (-20.92,-17.38)	-16.01 (-18.06,-13.97)
		Large	-13.53 (-21.79,-5.26)	-19.05 (-32.49,-5.62)	-15.91 (-28.41,-3.41)
	0.40	Small	-12.03 (-13.03,-11.02)	-17.47 (-19.44,-15.49)	-13.56 (-15.98,-11.13)
		Large	-12.25 (-19.58,-4.92)	-17.22 (-30.57,-3.87)	-13.30 (-26.13,-0.47)
550	0.10	Small	-15.00 (-15.81,-14.18)	-19.63 (-20.40,-18.87)	-17.73 (-18.58,-16.88)
		Large	-15.10 (-19.15,-11.05)	-19.65 (-25.87,-13.44)	-17.79 (-23.36,-12.21)
	0.20	Small	-13.36 (-13.92,-12.79)	-19.05 (-19.78,-18.31)	-16.04 (-16.89,-15.20)
		Large	-13.33 (-16.80,-9.86)	-19.01 (-25.05,-12.96)	-16.02 (-21.75,-10.29)
	0.40	Small	-12.00 (-12.42,-11.57)	-17.39 (-18.24,-16.54)	-13.66 (-14.50,-12.82)
		Large	-11.93 (-15.11,-8.76)	-17.24 (-22.91,-11.57)	-13.49 (-18.93,-8.06)
1100	0.10	Small	-15.00 (-15.57,-14.43)	-19.62 (-20.15,-19.09)	-17.74 (-18.34,-17.14)
		Large	-15.07 (-18.03,-12.11)	-19.76 (-24.17,-15.35)	-17.76 (-21.86,-13.66)
	0.20	Small	-13.33 (-13.73,-12.94)	-19.01 (-19.54,-18.47)	-16.04 (-16.67,-15.41)
		Large	-13.31 (-15.89,-10.72)	-18.96 (-23.10,-14.82)	-16.08 (-19.87,-12.29)
	0.40	Small	-11.99 (-12.29,-11.70)	-17.38 (-17.99,-16.78)	-13.67 (-14.26,-13.09)
		Large	-11.99 (-14.28,-9.71)	-17.34 (-21.38,-13.29)	-13.67 (-17.54,-9.80)

Table 4.6: Simulation Results for Estimated $\mu_S - \mu_{NS}$ (95% Confidence Intervals) - EM Algorithm

Sample Size	π	CV	Lab Values	EM Algorithm
55	0.10	Small	-15.20 (-17.96,-12.44)	-14.19 (-16.43,-11.95)
		Large	-15.00 (-28.89,-1.11)	-10.58 (-20.63,-0.53)
	0.20	Small	-13.47 (-15.40,-11.53)	-12.91 (-14.64,-11.17)
		Large	-13.54 (-25.20,-1.88)	-10.77 (-18.25,-3.29)
	0.40	Small	-12.05 (-13.42,-10.67)	-11.73 (-13.04,-10.42)
		Large	-11.79 (-22.03,-1.55)	-10.61 (-17.46,-3.77)
110	0.10	Small	-15.12 (-16.96,-13.27)	-14.16 (-15.70,-12.62)
		Large	-15.13 (-24.43,-5.83)	-10.61 (-17.47,-3.75)
	0.20	Small	-13.37 (-14.75,-11.99)	-12.81 (-14.05,-11.57)
		Large	-13.53 (-21.79,-5.26)	-10.67 (-15.90,-5.45)
	0.40	Small	-12.03 (-13.03,-11.02)	-11.71 (-12.67,-10.75)
		Large	-12.25 (-19.58,-4.92)	-10.51 (-14.47,-6.56)
550	0.10	Small	-15.00 (-15.81,-14.18)	-14.07 (-14.76,-13.38)
		Large	-15.10 (-19.15,-11.05)	-10.56 (-13.55,-7.57)
	0.20	Small	-13.36 (-13.92,-12.79)	-12.79 (-13.30,-12.27)
		Large	-13.33 (-16.80,-9.86)	-10.50 (-12.78,-8.21)
	0.40	Small	-12.00 (-12.42,-11.57)	-11.68 (-12.10,-11.26)
		Large	-11.93 (-15.11,-8.76)	-10.43 (-12.09,-8.77)
1100	0.10	Small	-15.00 (-15.57,-14.43)	-14.06 (-14.56,-13.57)
		Large	-15.07 (-18.03,-12.11)	-10.56 (-12.78,-8.33)
	0.20	Small	-13.33 (-13.73,-12.94)	-12.77 (-13.13,-12.41)
		Large	-13.31 (-15.89,-10.72)	-10.52 (-12.14,-8.90)
	0.40	Small	-11.99 (-12.29,-11.70)	-11.68 (-11.96,-11.39)
		Large	-11.99 (-14.28,-9.71)	-10.44 (-11.63,-9.25)

Table 4.7: Number of times and percent smoking status was significant ($p < 0.05$) Current Methods and Probability Based Screener Index

Sample Size	π	CV	Lab Values	Self Report	Inaccurate Resp Removed	Prob Based Index v1	Prob Based Index v2
55	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	585 (59%)	544 (54%)	586 (59%)	133 (13%)	589 (59%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	990 (99%)	1000 (100%)
		Large	600 (60%)	495 (50%)	592 (59%)	93 (9%)	588 (59%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	717 (72%)	1000 (100%)
		Large	605 (61%)	373 (37%)	527 (53%)	140 (14%)	541 (54%)
110	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	895 (90%)	829 (83%)	870 (87%)	178 (18%)	868 (87%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	999 (100%)	1000 (100%)
		Large	892 (89%)	809 (81%)	885 (89%)	155 (16%)	885 (89%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	761 (76%)	1000 (100%)
		Large	906 (91%)	675 (68%)	856 (86%)	172 (17%)	860 (86%)
550	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	746 (75%)	1000 (100%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	635 (64%)	1000 (100%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	894 (89%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	292 (29%)	1000 (100%)
1100	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	961 (96%)	1000 (100%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	933 (93%)	1000 (100%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	970 (97%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	382 (38%)	1000 (100%)

Table 4.8: Number of times and percent smoking status was significant ($p < 0.05$) Heckman's Model and EM Algorithm

Sample Size	π	CV	Lab Values	Heckman v1	Heckman v2	EM
55	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	972 (100%)
		Large	585 (59%)	475 (48%)	426 (43%)	52 (5%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	989 (100%)
		Large	600 (60%)	455 (46%)	347 (35%)	55 (6%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	998 (100%)
		Large	605 (61%)	381 (38%)	260 (26%)	159 (19%)
110	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	974 (100%)
		Large	895 (90%)	758 (76%)	730 (73%)	115 (12%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	999 (100%)
		Large	892 (89%)	739 (74%)	632 (63%)	63 (7%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	906 (91%)	675 (68%)	508 (51%)	272 (37%)
550	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	993 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	990 (99%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	999 (100%)	998 (100%)	995 (100%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	999 (100%)	316 (63%)
1100	0.10	Small	1000 (100%)	1000 (100%)	1000 (100%)	999 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
	0.20	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
	0.40	Small	1000 (100%)	1000 (100%)	1000 (100%)	1000 (100%)
		Large	1000 (100%)	1000 (100%)	1000 (100%)	329 (77%)

Table 4.9: Number of times the EM Algorithm did not converge

Sample Size	π	CV	Count (%)
55	0.10	Small	27 (3%)
		Large	16 (2%)
	0.20	Small	11 (1%)
		Large	71 (7%)
	0.40	Small	0 (0%)
		Large	173 (17%)
110	0.10	Small	25 (3%)
		Large	7 (1%)
	0.20	Small	1 (0%)
		Large	63 (6%)
	0.40	Small	0 (0%)
		Large	255 (26%)
550	0.10	Small	7 (1%)
		Large	0 (0%)
	0.20	Small	0 (0%)
		Large	3 (0%)
	0.40	Small	0 (0%)
		Large	499 (50%)
1100	0.10	Small	1 (0%)
		Large	0 (0%)
	0.20	Small	0 (0%)
		Large	0 (0%)
	0.40	Small	0 (0%)
		Large	570 (57%)

Table 4.10: Estimated Intentional Inaccurate Probabilities (95% Confidence Intervals) from EM Algorithm

Sample Size	π	CV	$\hat{\pi}$	$1-\hat{\pi}$
55	0.10	Small	0.11 (0.03,0.19)	0.89 (0.81,0.97)
		Large	0.13 (0.10,0.15)	0.87 (0.85,0.90)
	0.20	Small	0.20 (0.08,0.33)	0.80 (0.67,0.92)
		Large	0.22 (0.16,0.27)	0.78 (0.73,0.84)
	0.40	Small	0.38 (0.17,0.59)	0.62 (0.41,0.83)
		Large	0.37 (0.12,0.61)	0.63 (0.39,0.88)
110	0.10	Small	0.11 (0.05,0.17)	0.89 (0.83,0.95)
		Large	0.13 (0.12,0.14)	0.87 (0.86,0.88)
	0.20	Small	0.21 (0.13,0.29)	0.79 (0.71,0.87)
		Large	0.22 (0.19,0.25)	0.78 (0.75,0.81)
	0.40	Small	0.40 (0.25,0.55)	0.60 (0.45,0.75)
		Large	0.40 (0.27,0.53)	0.60 (0.47,0.73)
550	0.10	Small	0.11 (0.09,0.14)	0.89 (0.86,0.91)
		Large	0.13 (0.12,0.14)	0.87 (0.86,0.88)
	0.20	Small	0.21 (0.18,0.24)	0.79 (0.76,0.82)
		Large	0.22 (0.21,0.23)	0.78 (0.77,0.79)
	0.40	Small	0.41 (0.37,0.45)	0.59 (0.55,0.63)
		Large	0.41 (0.40,0.42)	0.59 (0.58,0.60)
1100	0.10	Small	0.11 (0.09,0.13)	0.89 (0.87,0.91)
		Large	0.13 (0.13,0.13)	0.87 (0.87,0.87)
	0.20	Small	0.21 (0.19,0.24)	0.79 (0.76,0.81)
		Large	0.22 (0.21,0.23)	0.78 (0.77,0.79)
	0.40	Small	0.41 (0.38,0.44)	0.59 (0.56,0.62)
		Large	0.41 (0.40,0.42)	0.59 (0.58,0.60)

Figure 4.1: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.10, CV: Small

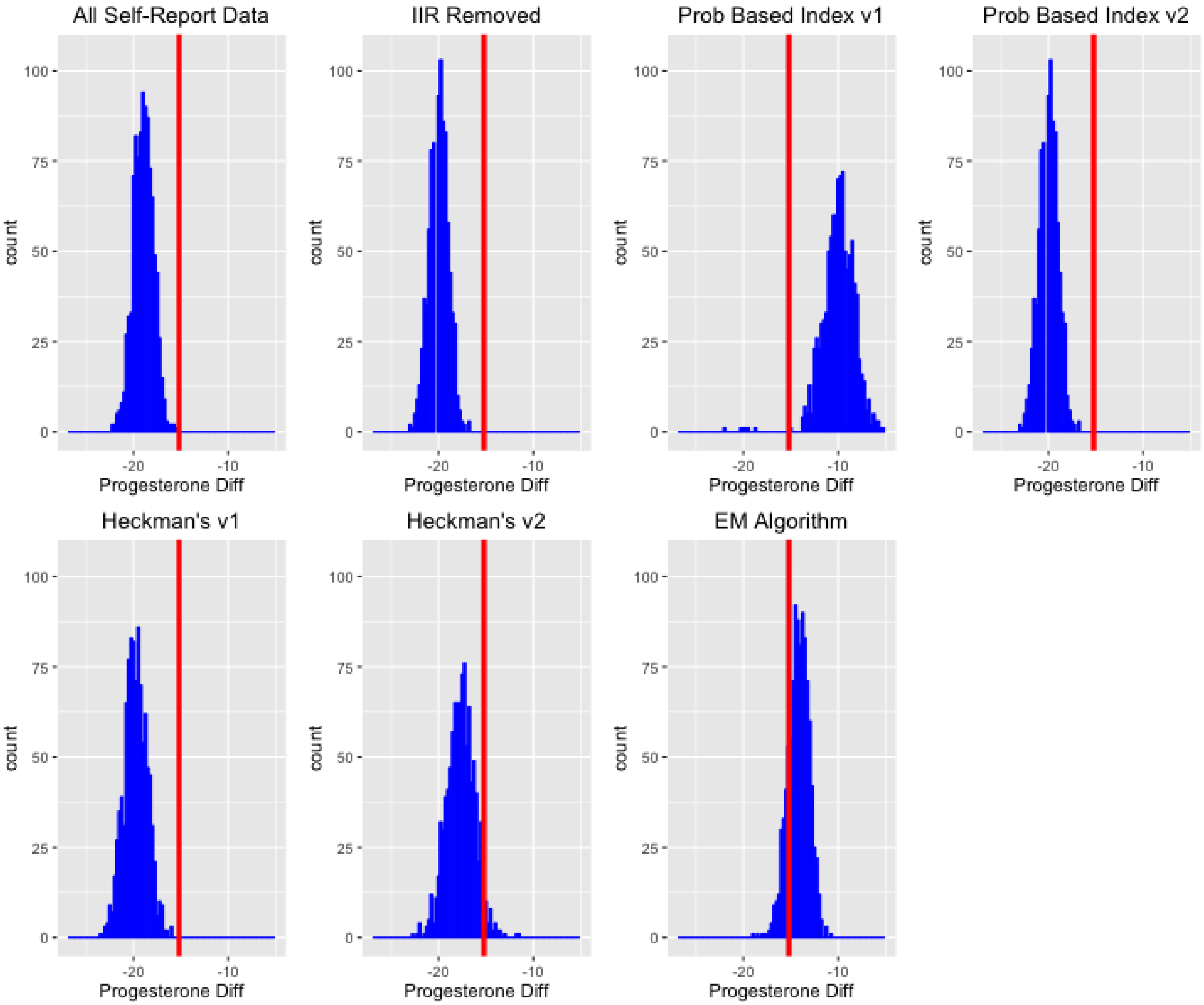
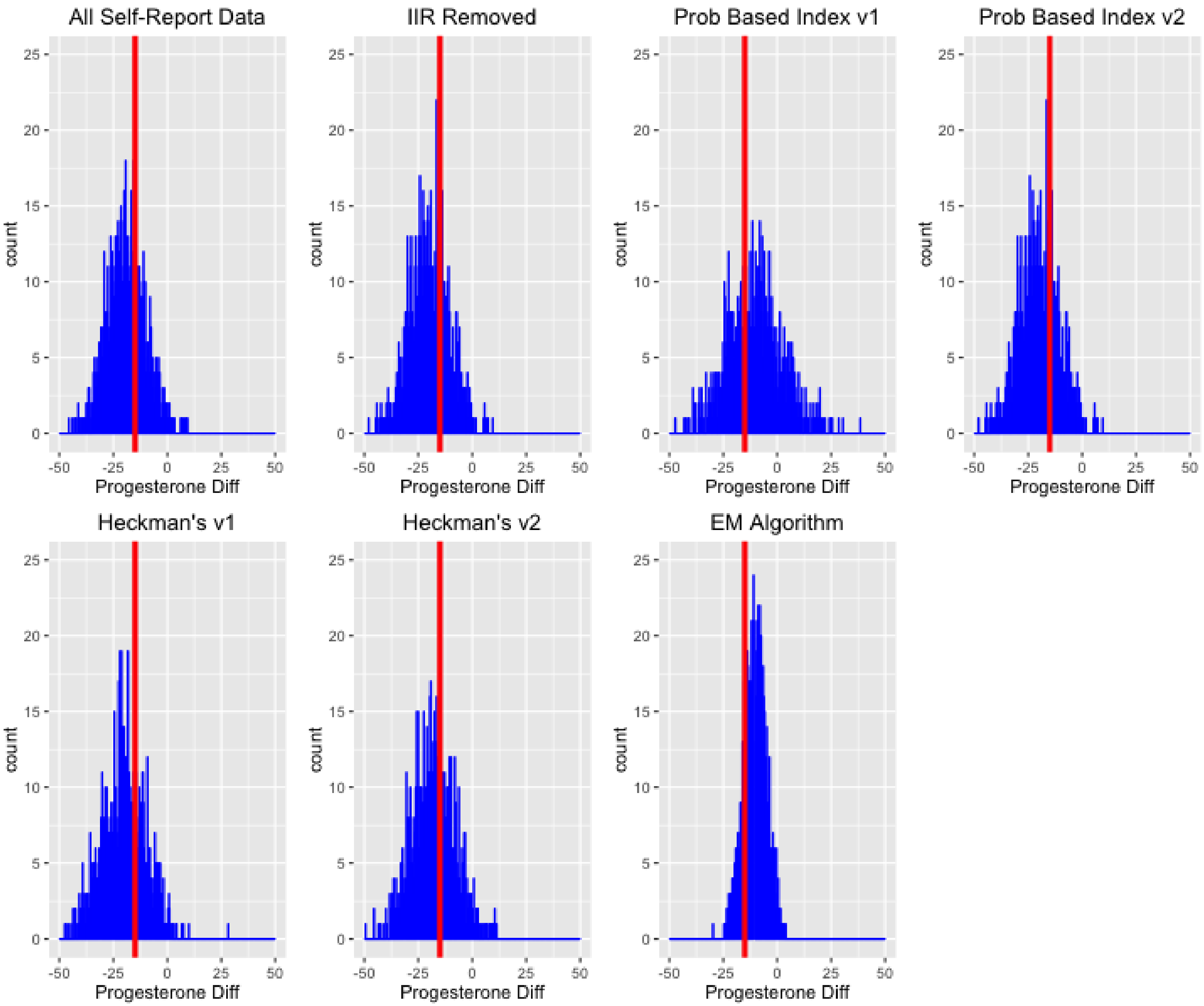


Figure 4.2: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.10, CV: Large



All Self-Report Data

IIR Removed

Prob Based Index v1

Prob Based Index v2

Heckman's v1

Heckman's v2

EM Algorithm

Figure 4.3: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.20, CV: Small

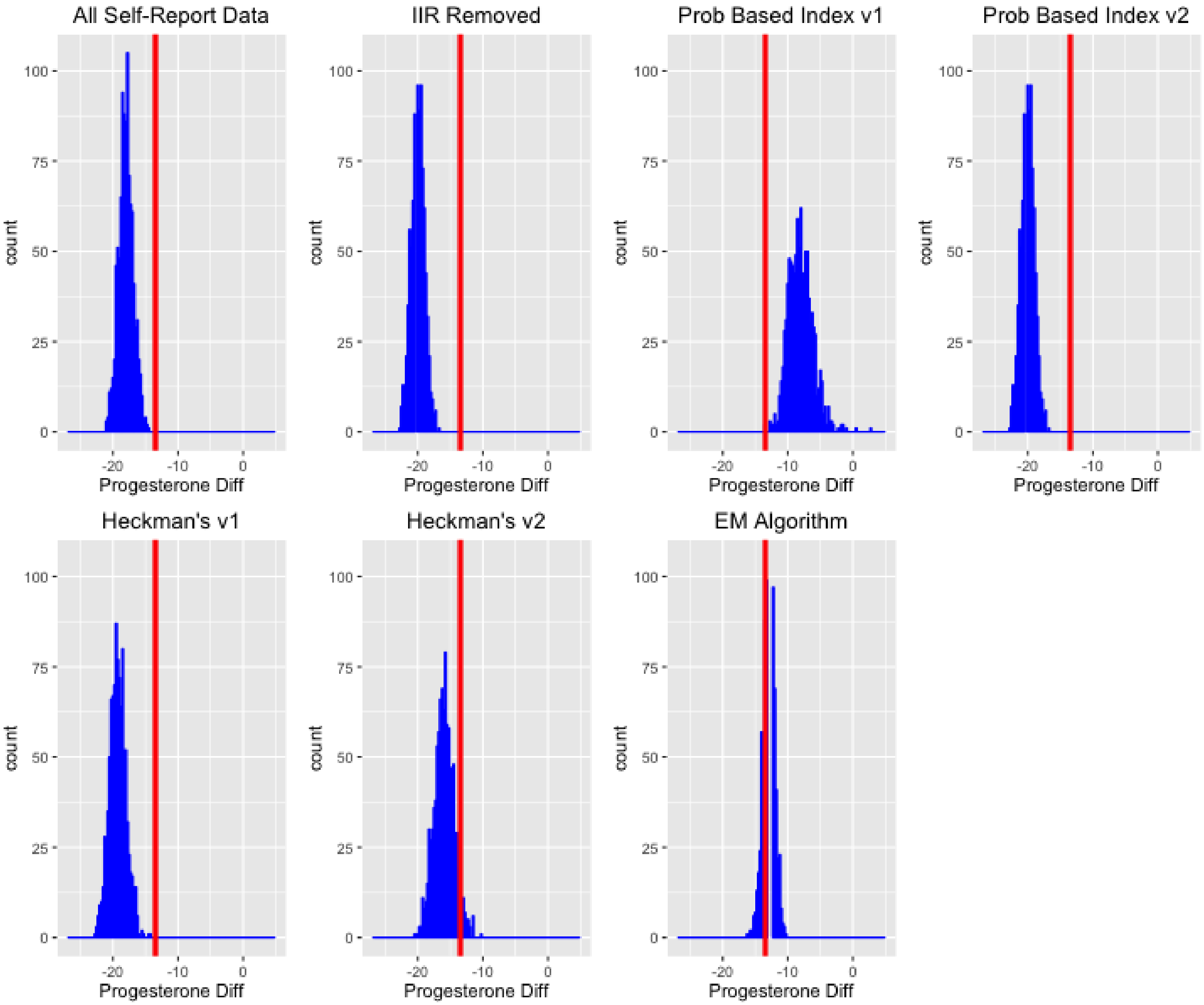


Figure 4.4: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.20, CV: Large

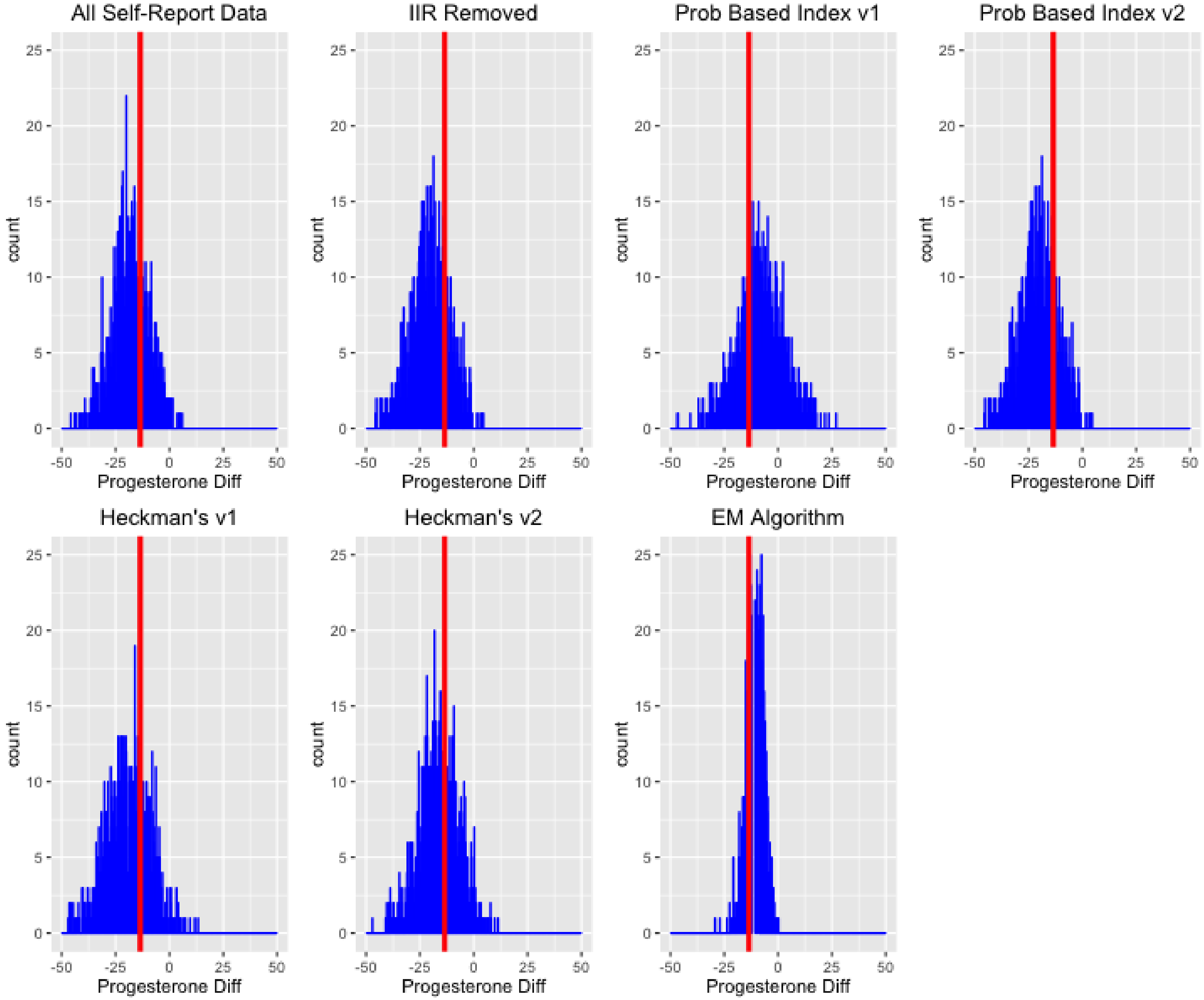


Figure 4.5: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.40, CV: Small

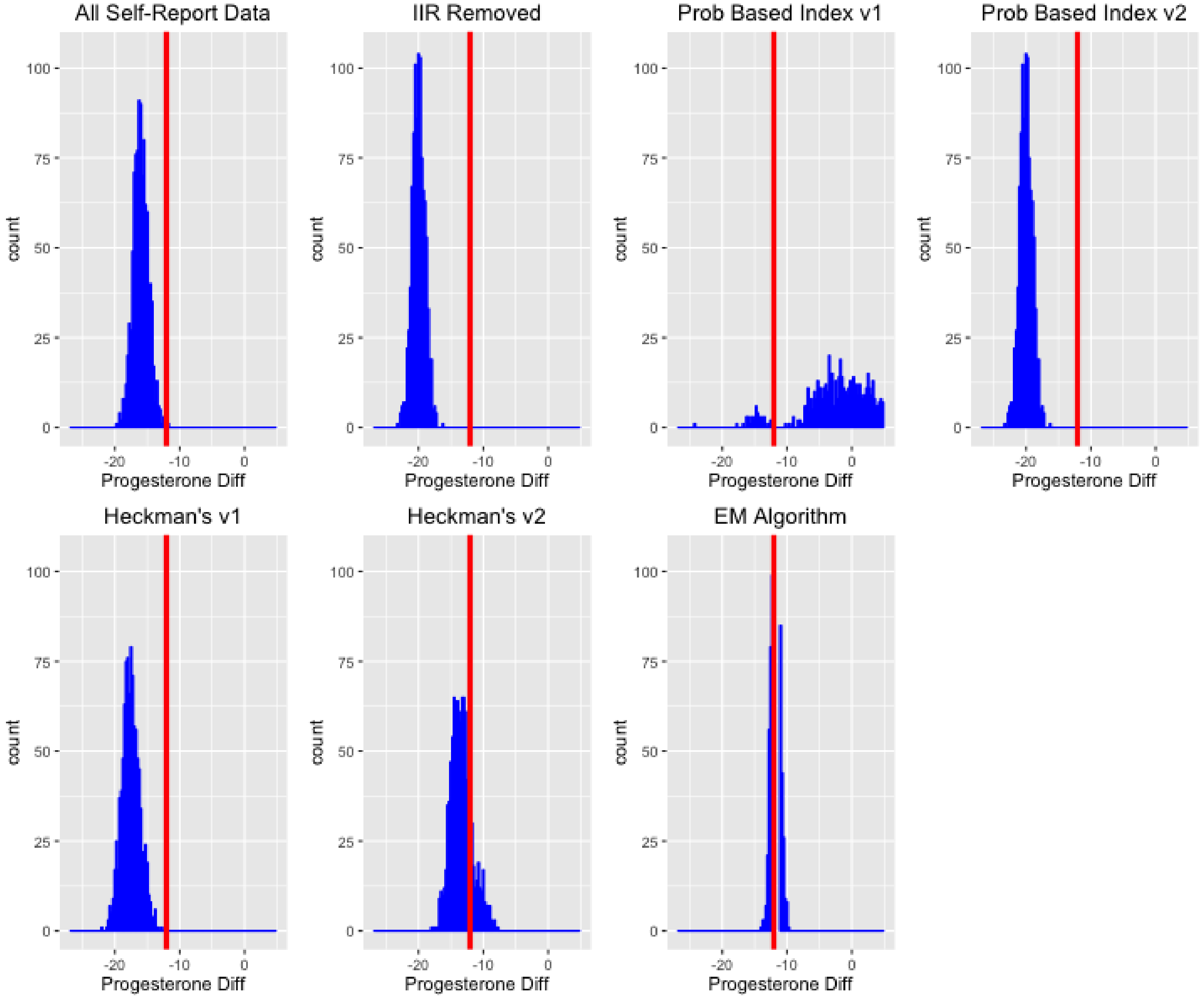


Figure 4.6: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 55, Probability of Intentional Inaccurate Response: 0.40, CV: Large

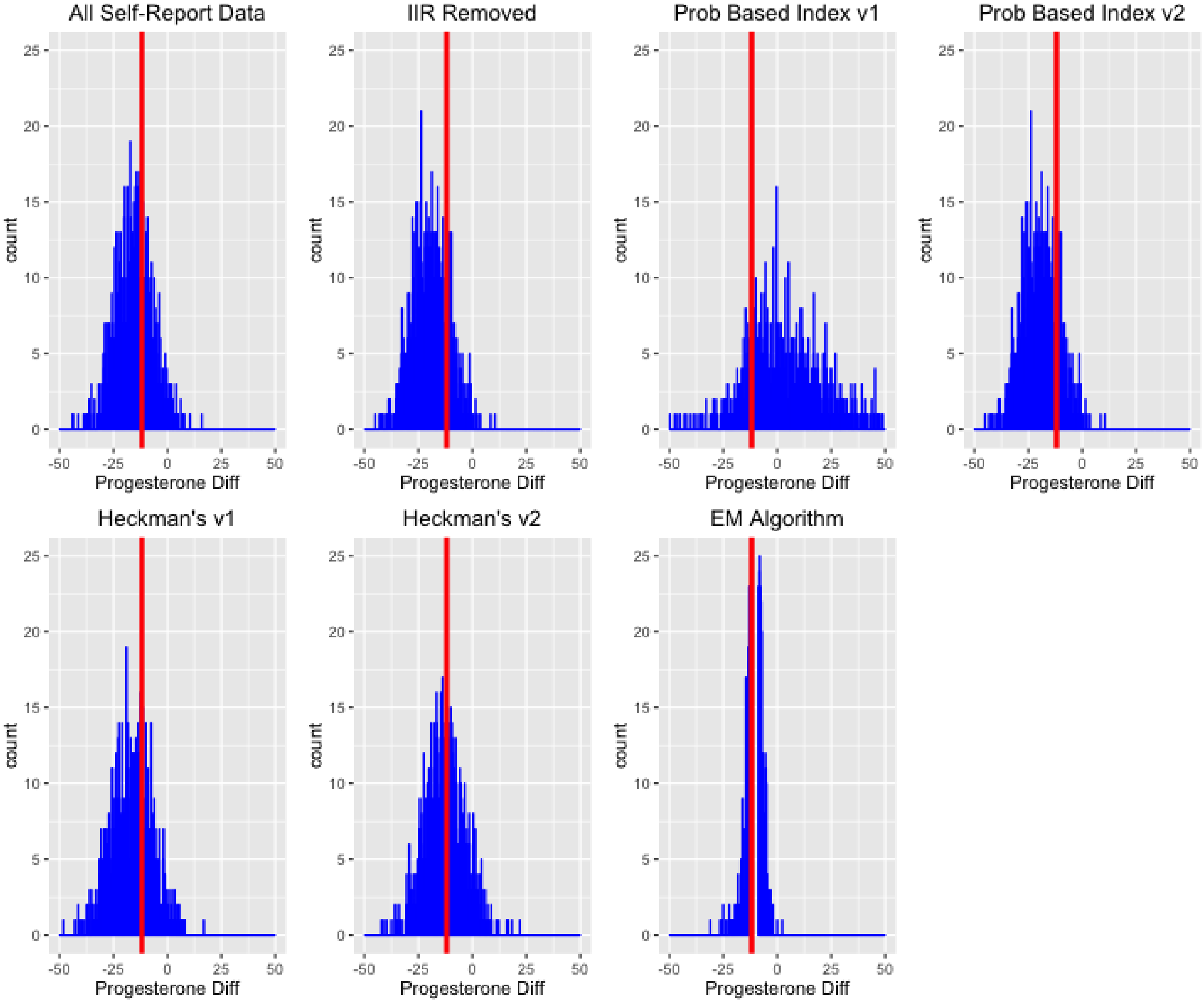
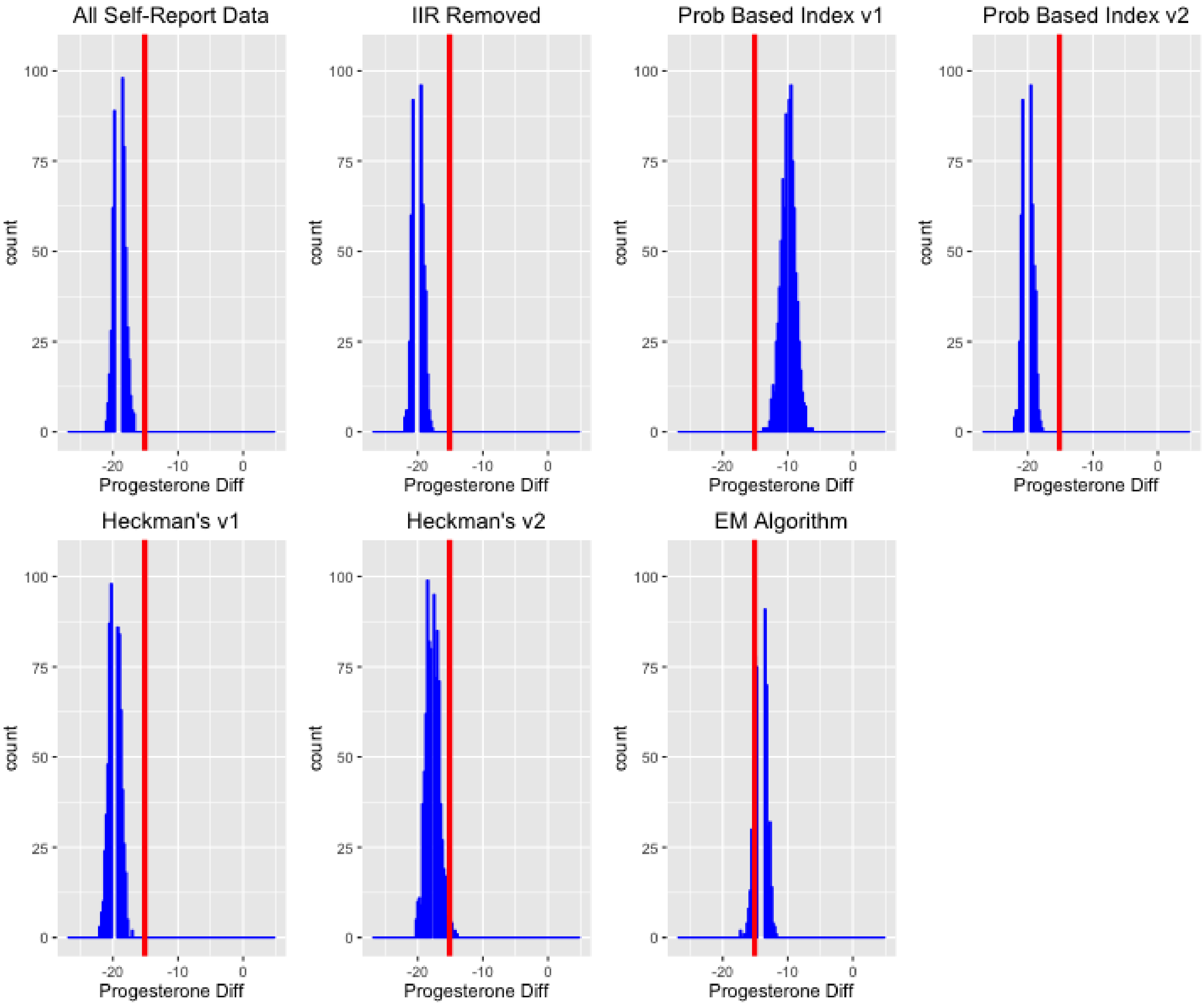


Figure 4.7: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.10, CV: Small



All Self-Report Data

IIR Removed

Prob Based Index v1

Prob Based Index v2

Heckman's v1

Heckman's v2

EM Algorithm

Figure 4.8: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.10, CV: Large

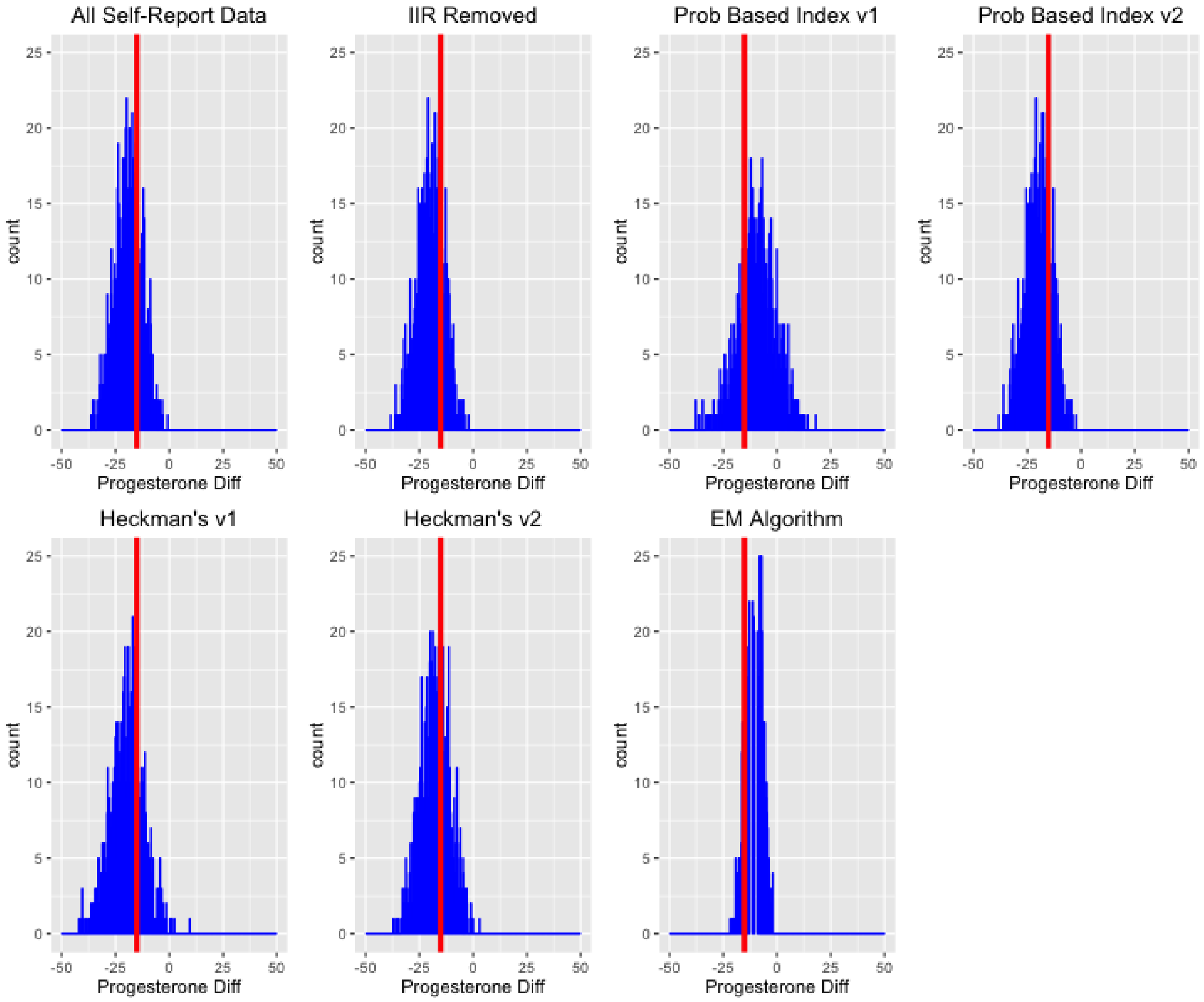


Figure 4.9: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.20, CV: Small

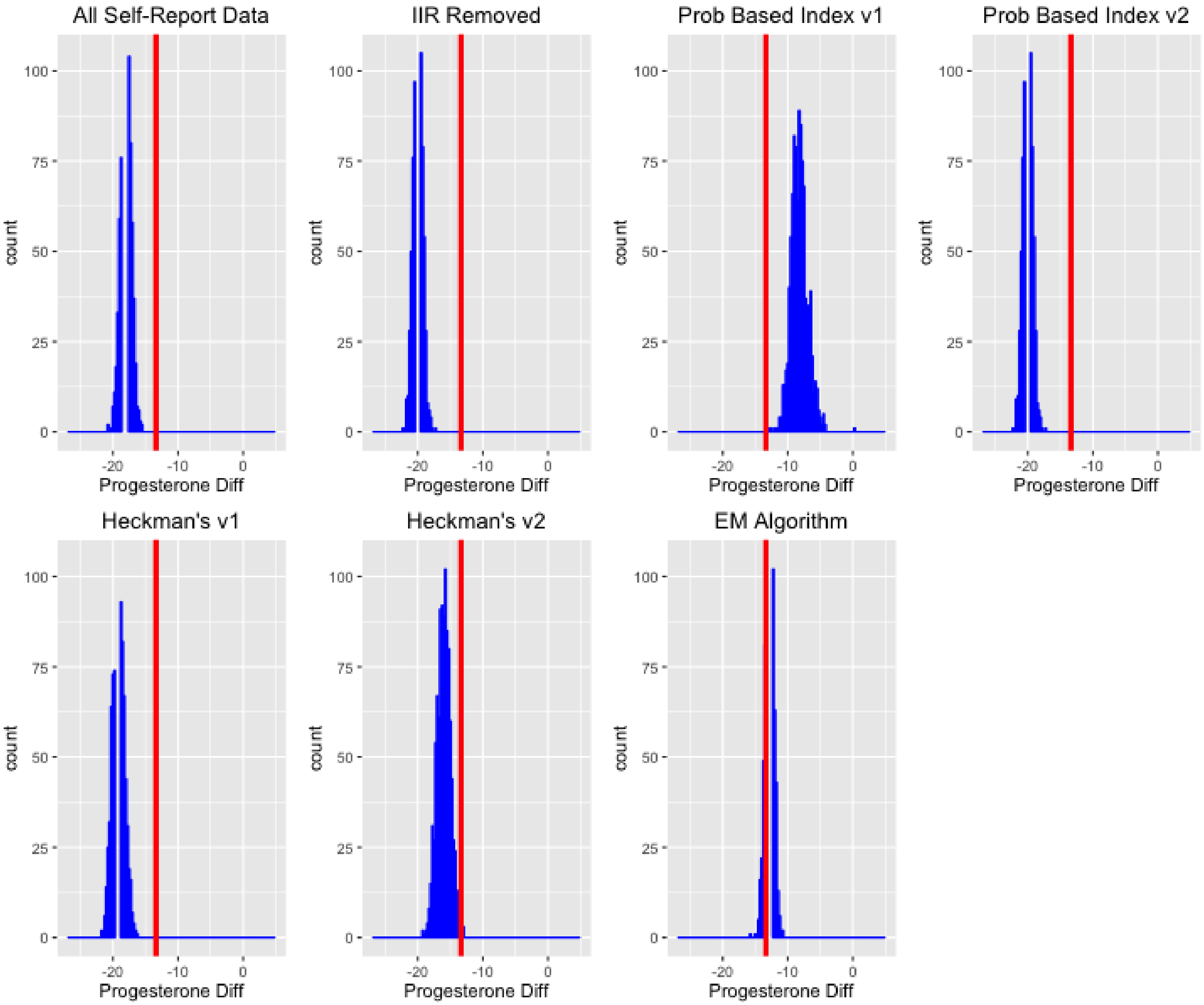


Figure 4.10: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.20, CV: Large

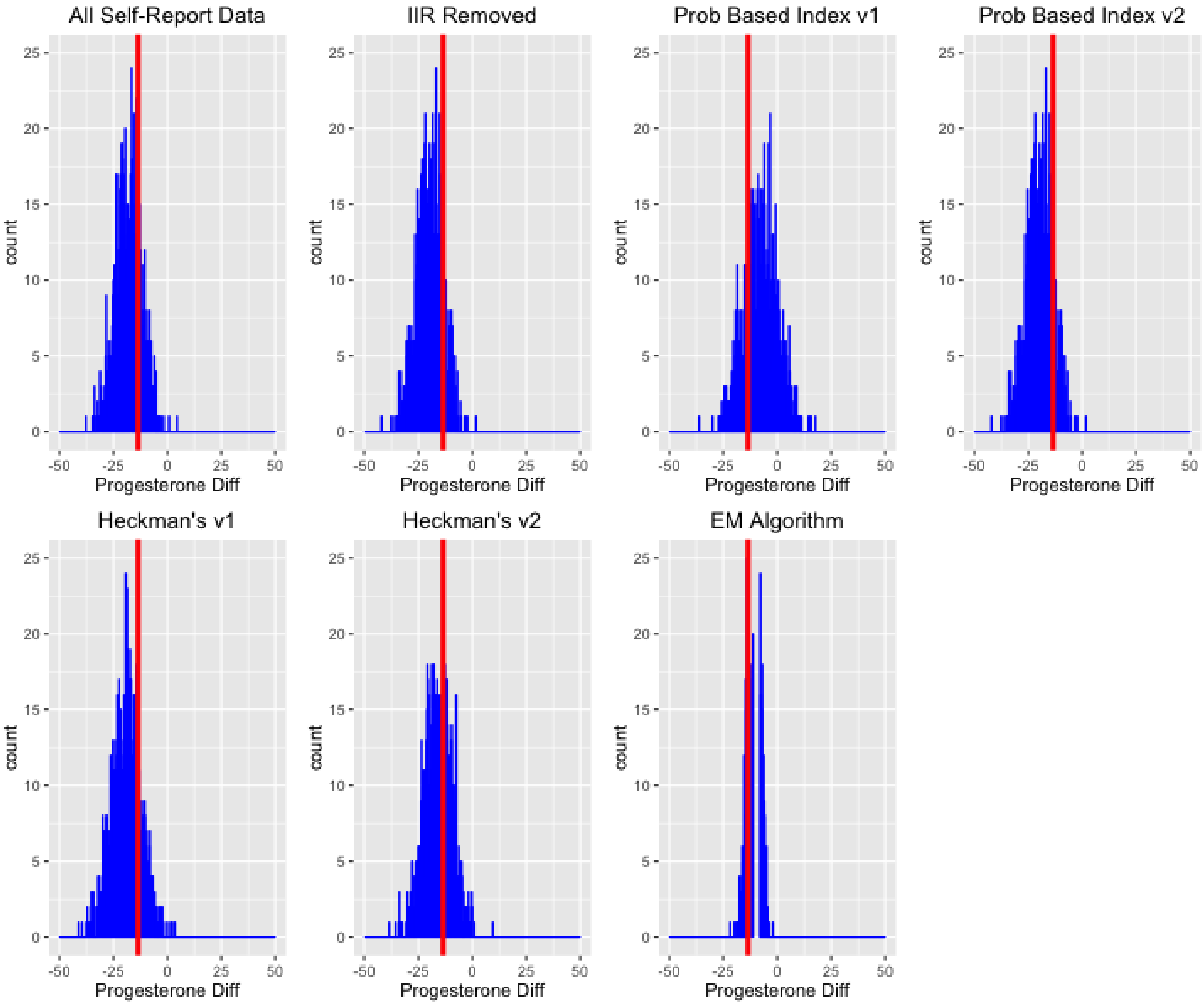


Figure 4.11: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.40, CV: Small

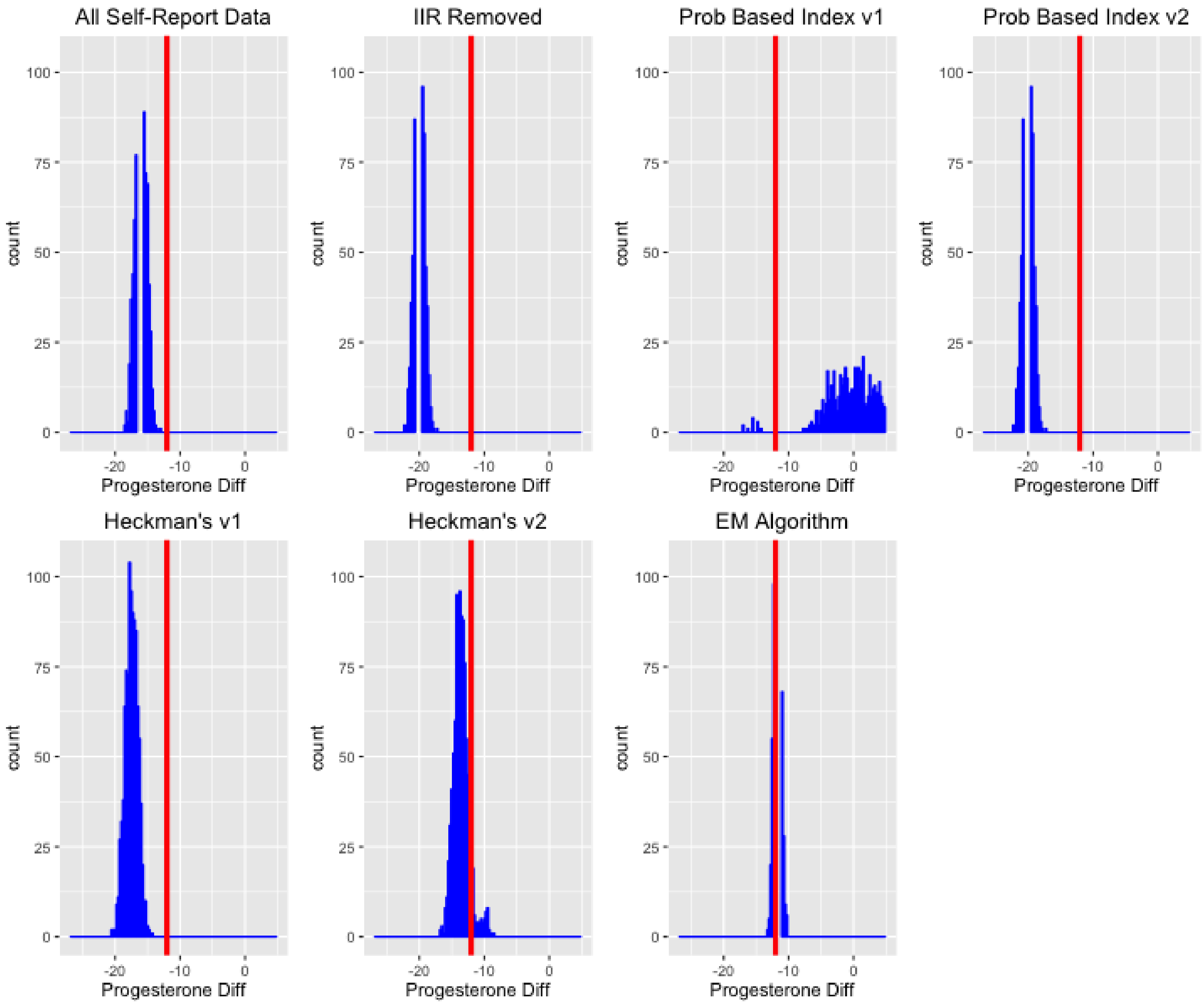


Figure 4.12: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 110, Probability of Intentional Inaccurate Response: 0.40, CV: Large

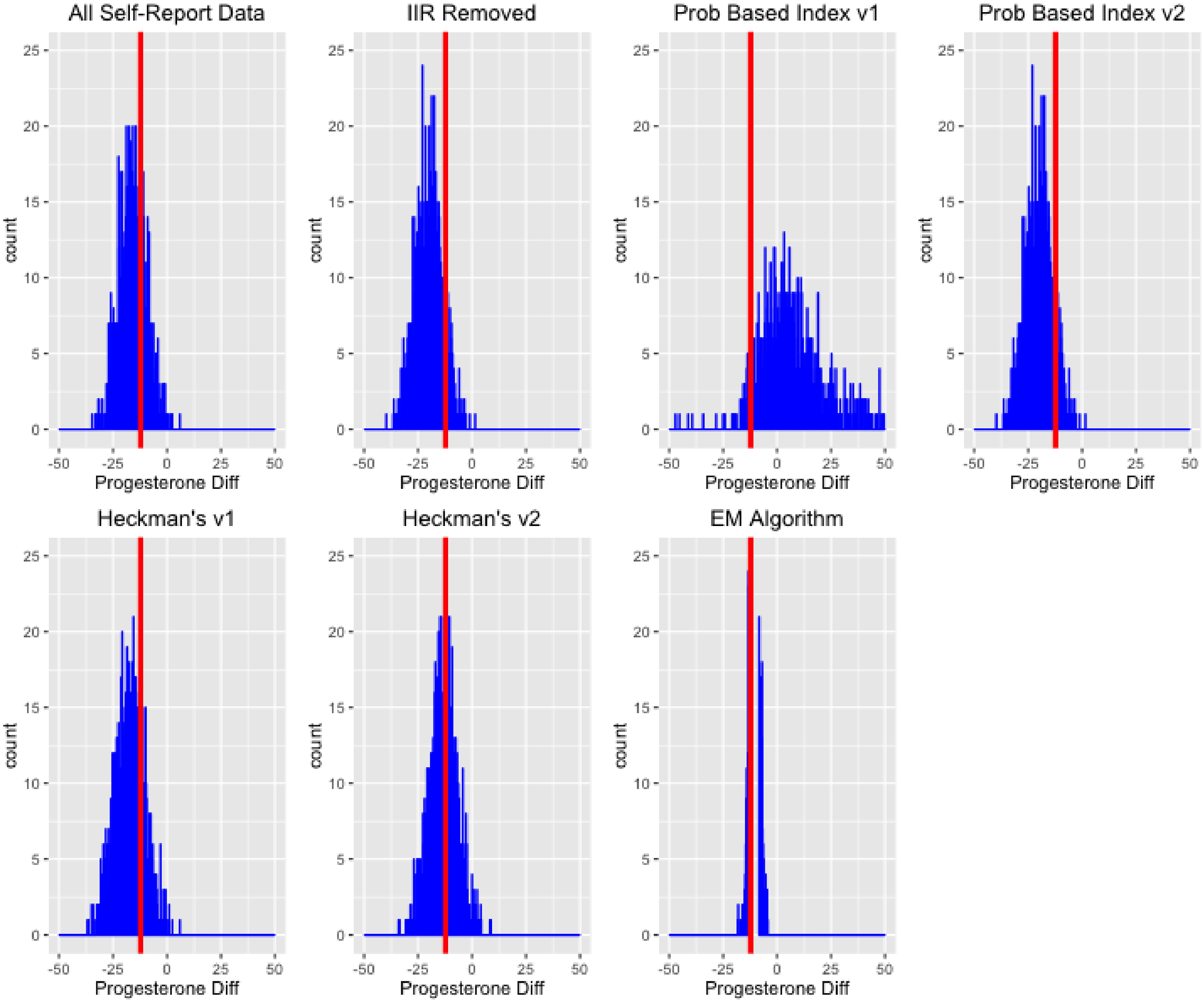


Figure 4.13: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.10, CV: Small

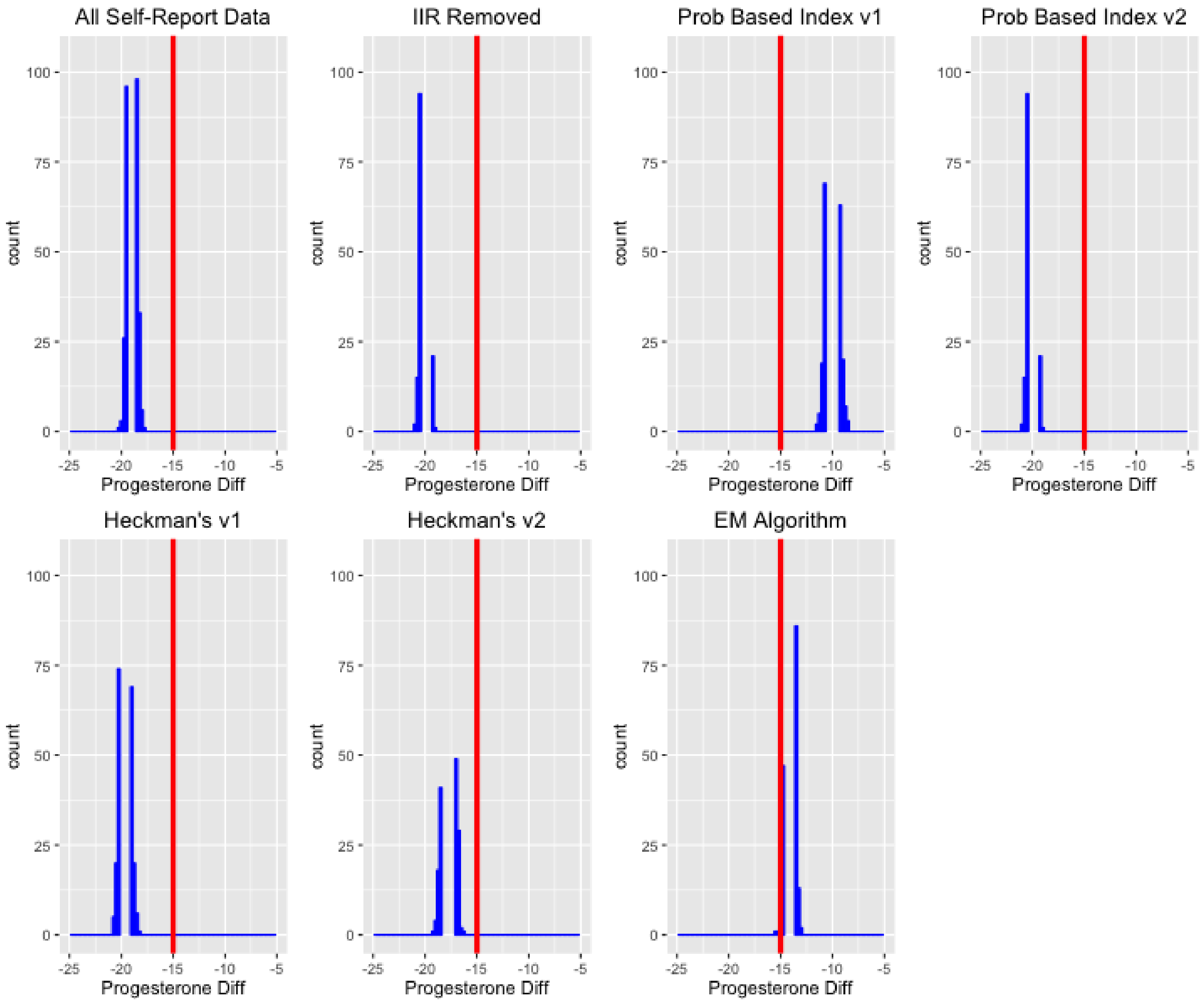


Figure 4.14: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.10, CV: Large

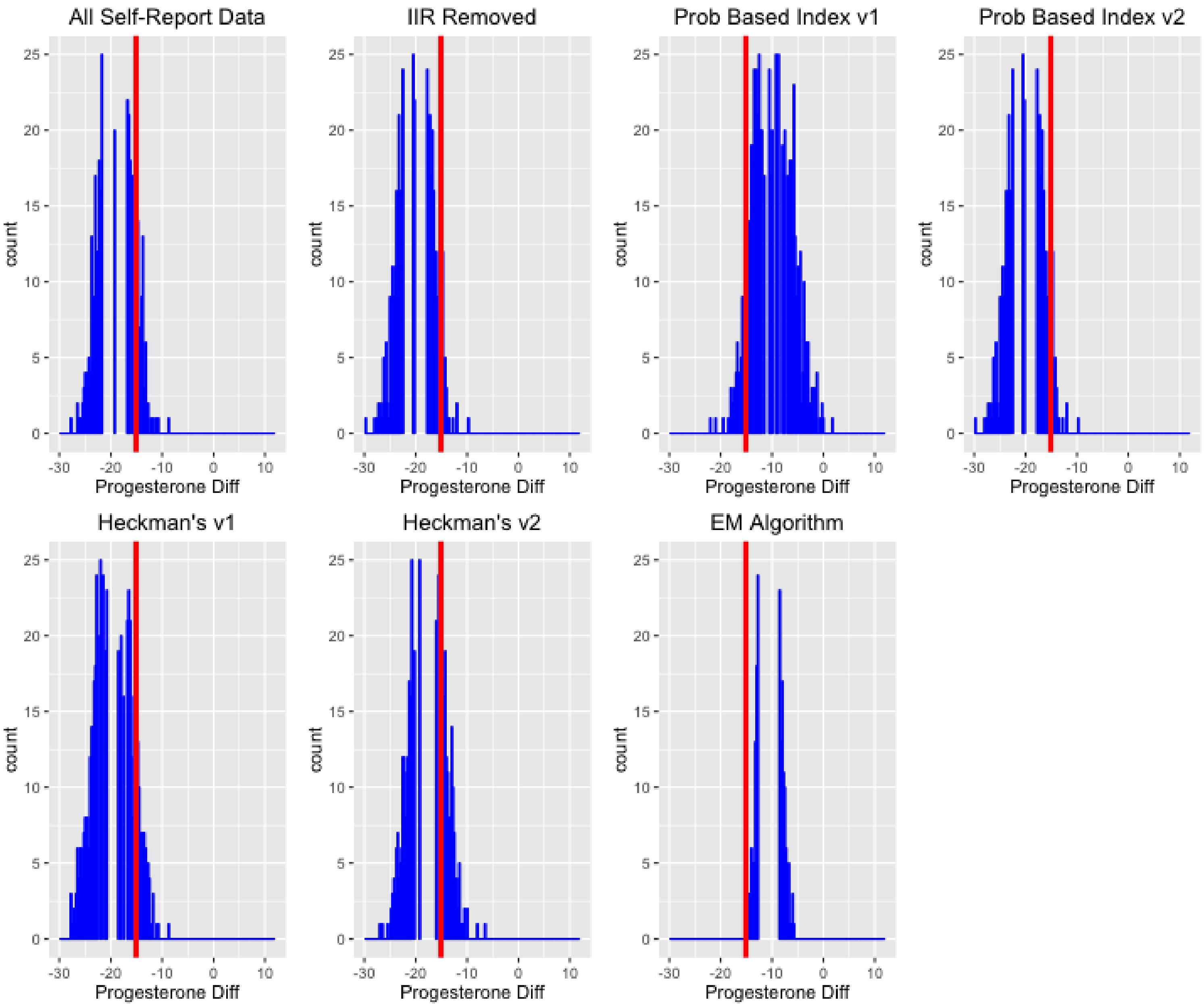
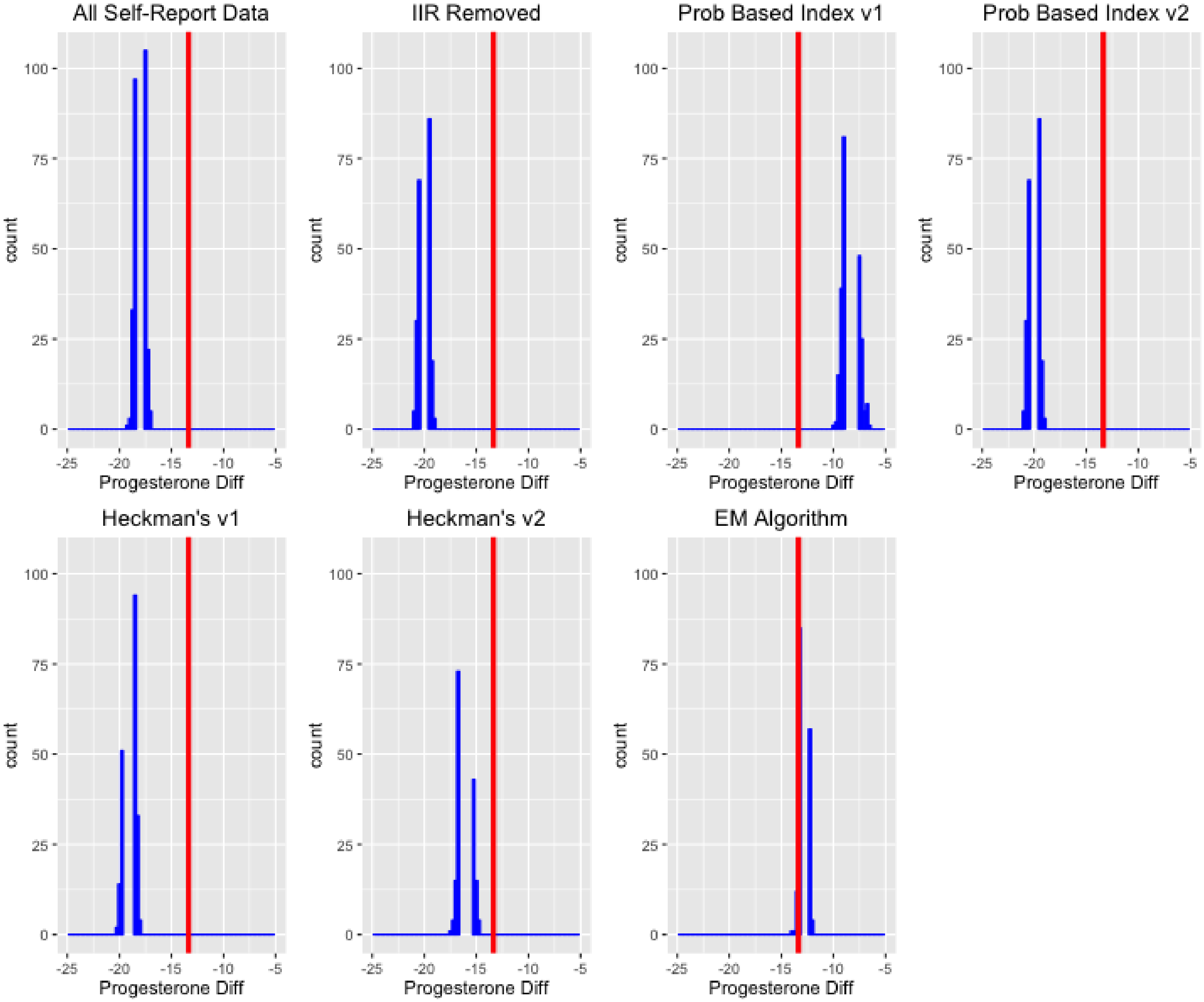


Figure 4.15: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.20, CV: Small



All Self-Report Data

IIR Removed

Prob Based Index v1

Prob Based Index v2

Heckman's v1

Heckman's v2

EM Algorithm

Figure 4.16: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.20, CV: Large

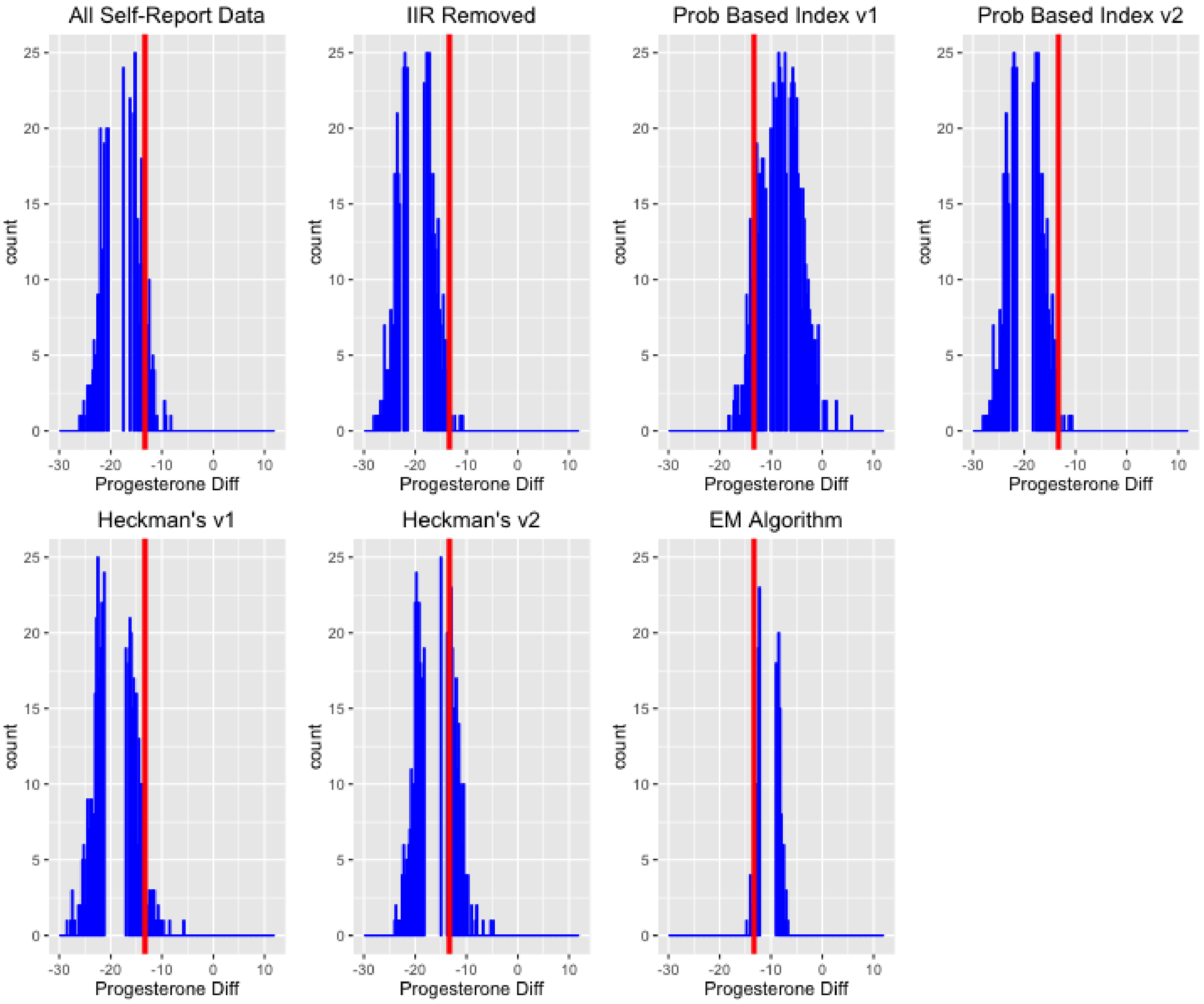


Figure 4.17: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.40, CV: Small

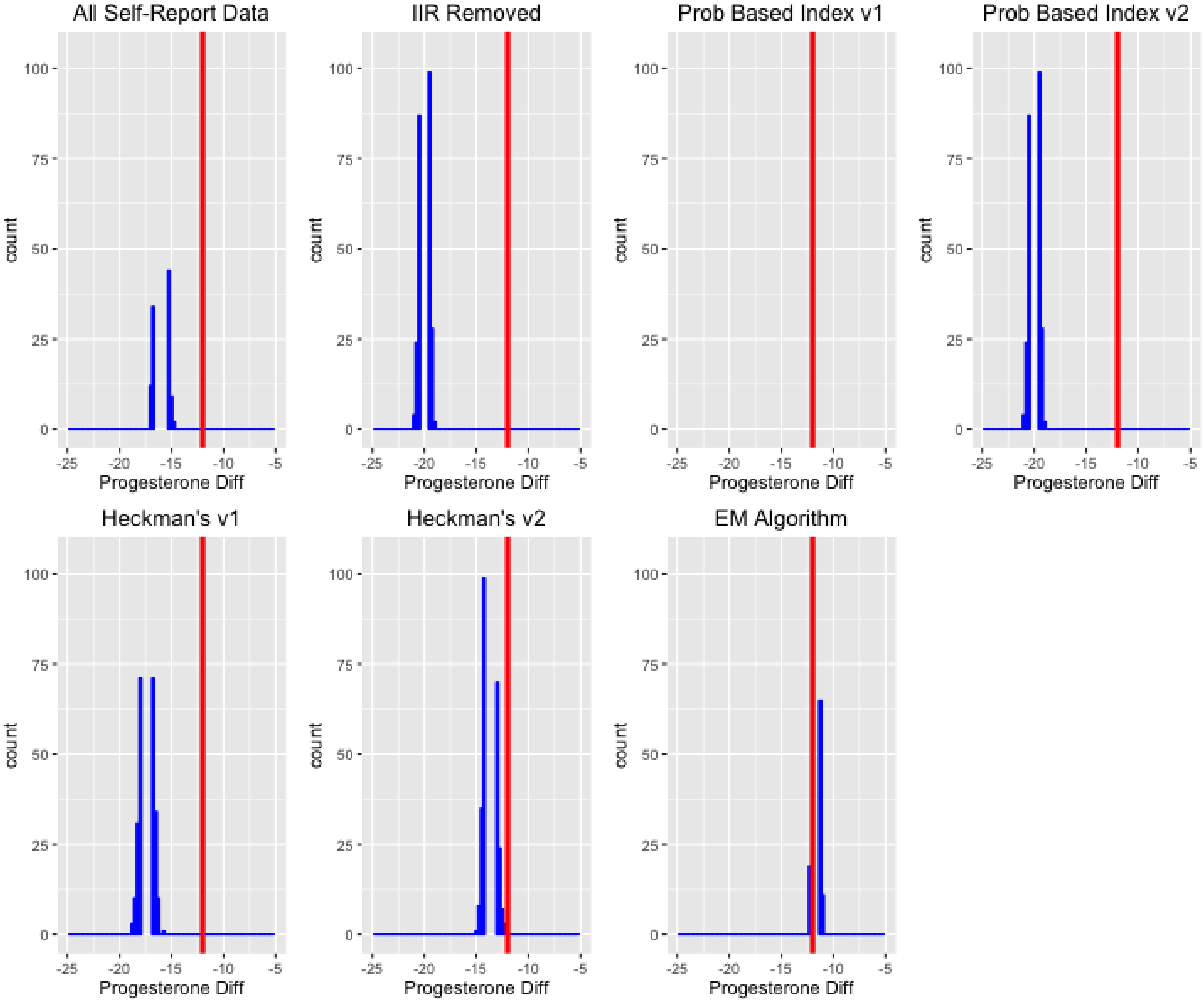


Figure 4.18: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 550, Probability of Intentional Inaccurate Response: 0.40, CV: Large

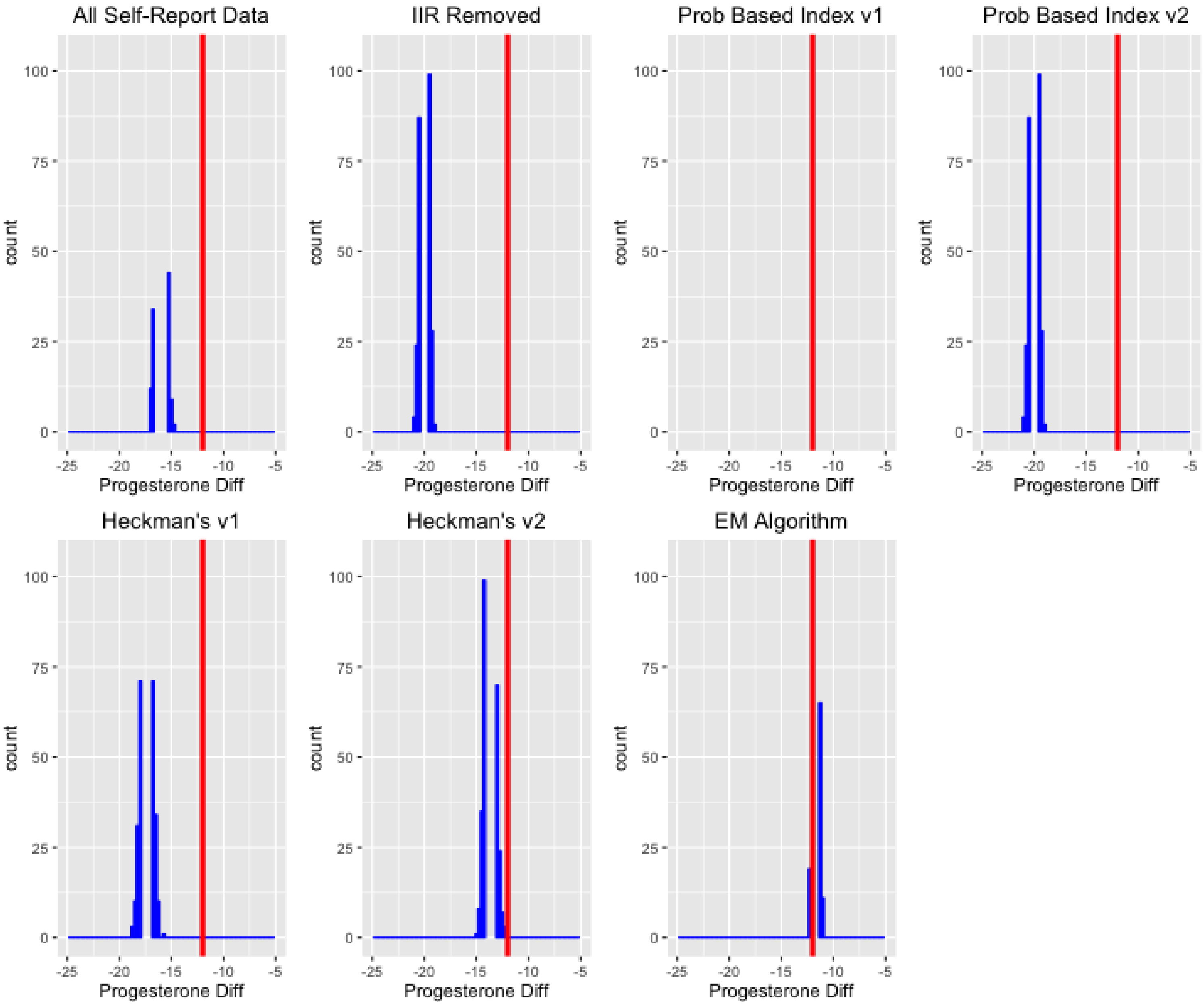


Figure 4.19: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.10, CV: Small

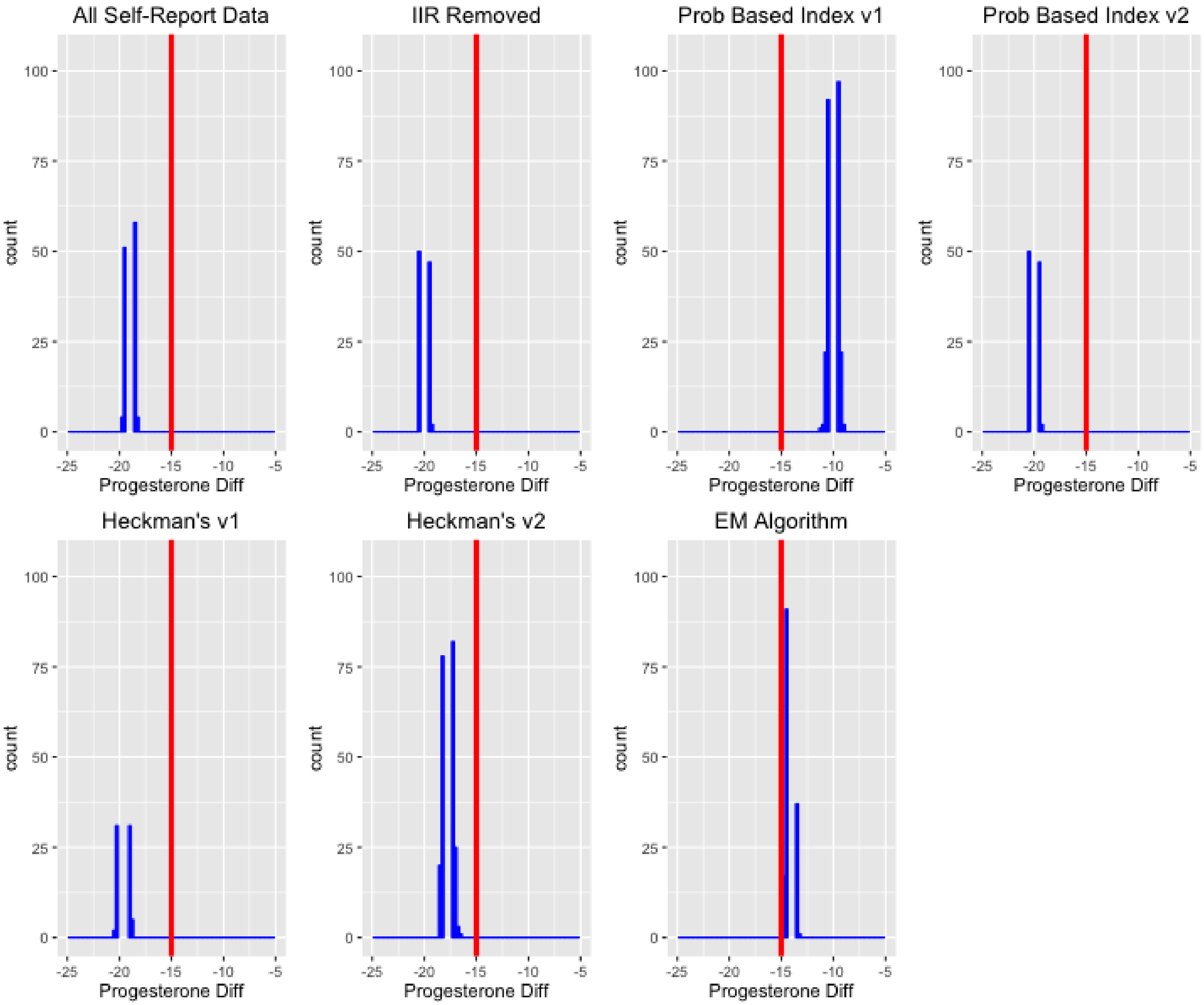


Figure 4.20: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.10, CV: Large

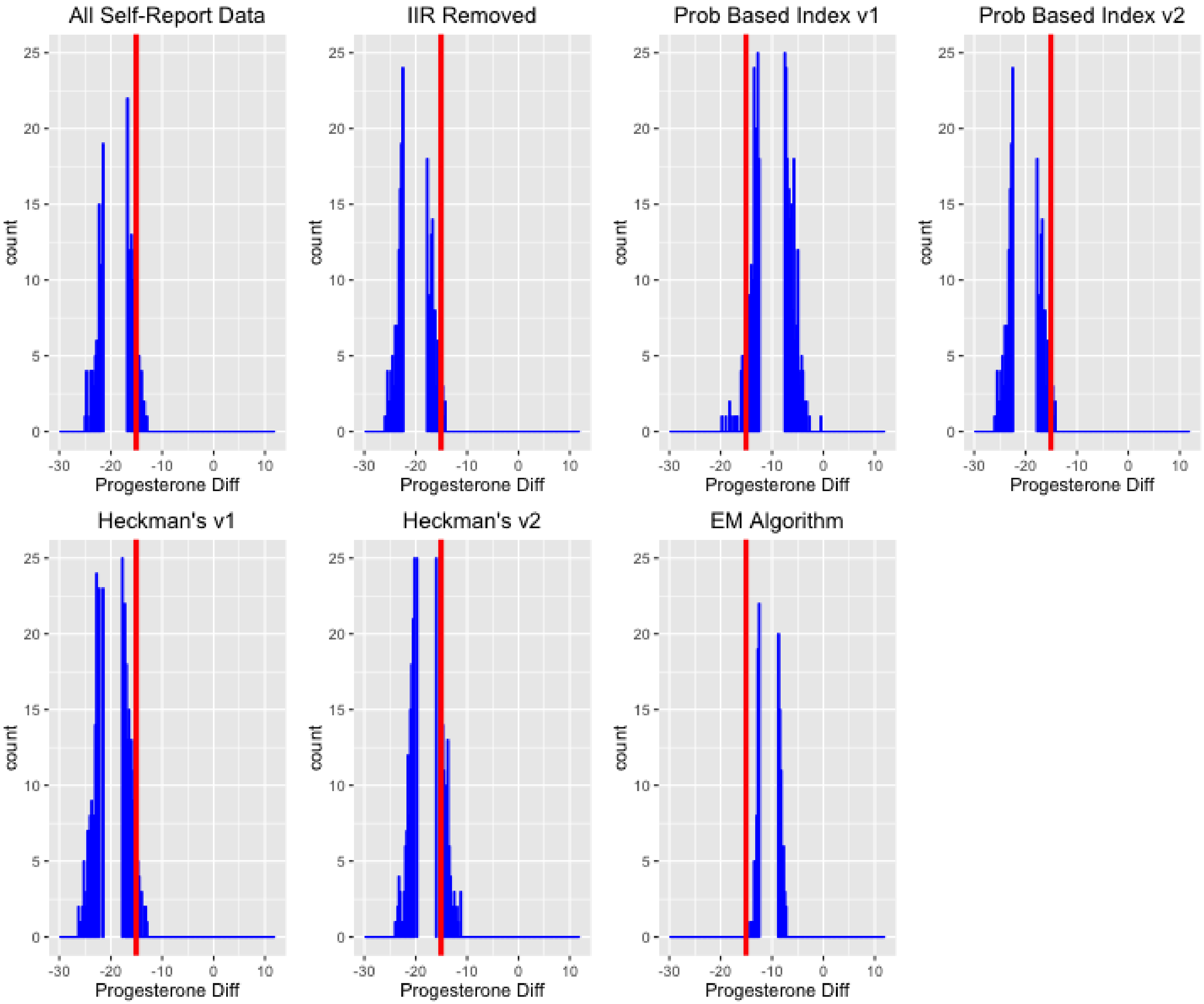


Figure 4.21: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.20, CV: Small

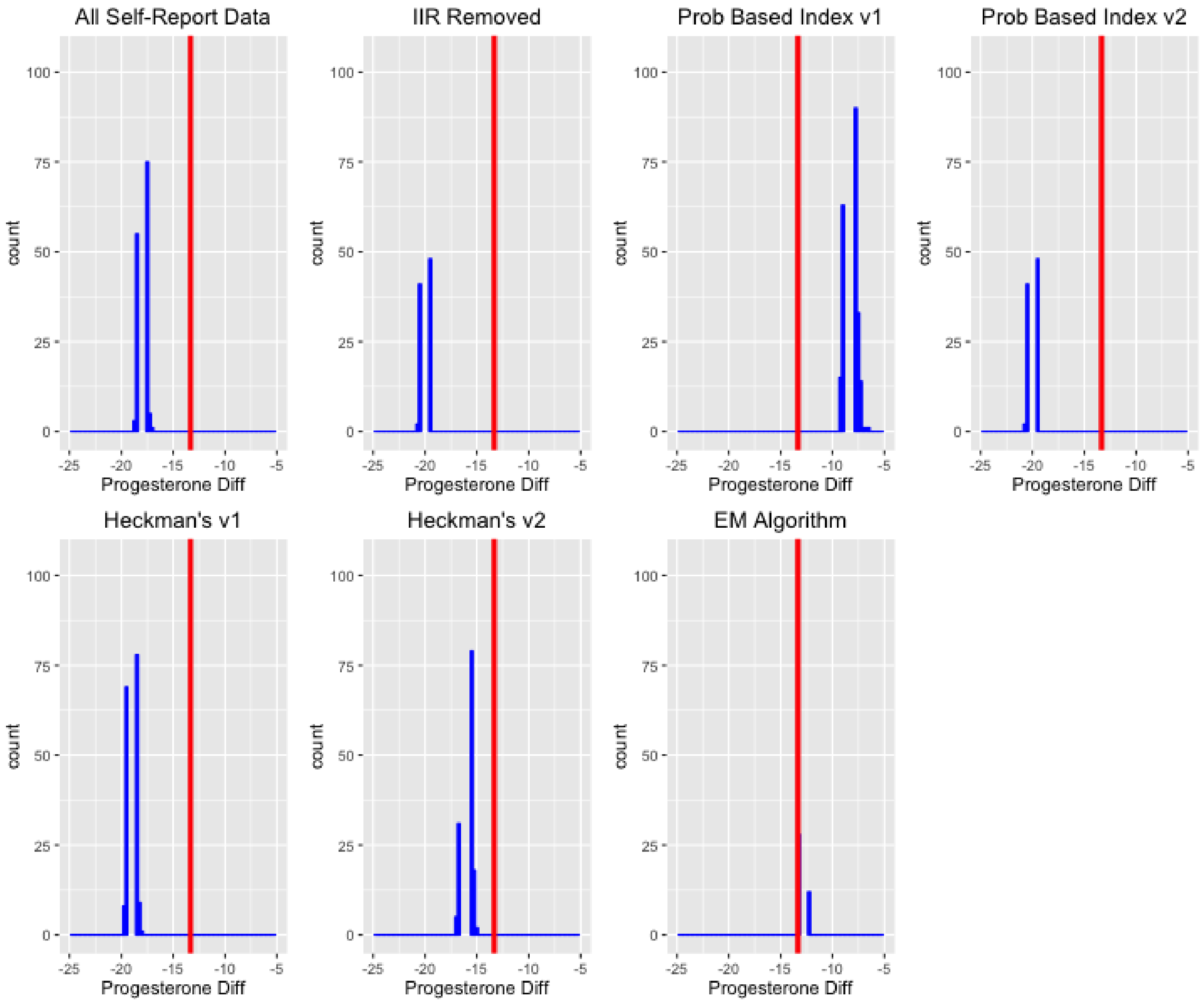


Figure 4.22: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.20, CV: Large

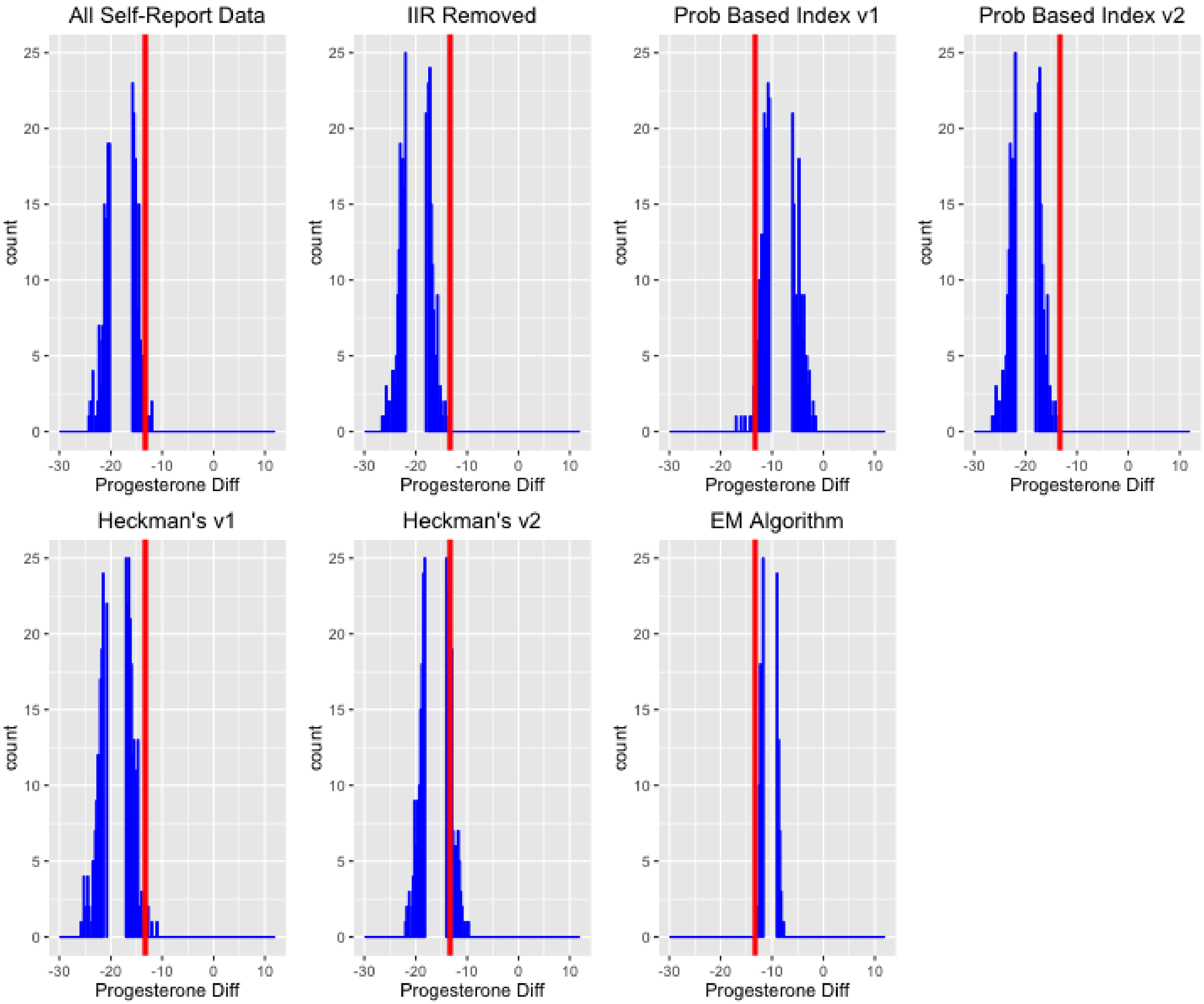
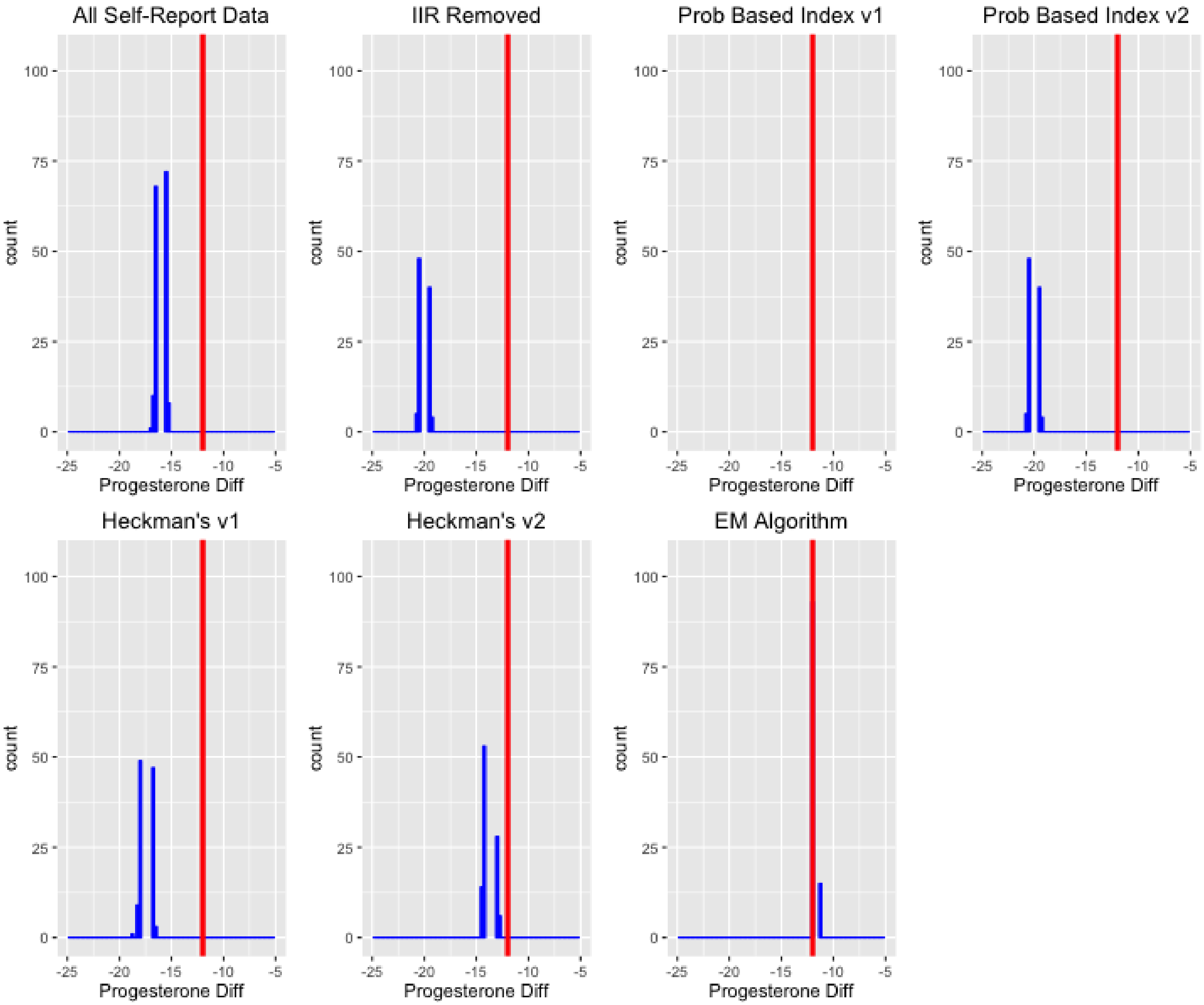


Figure 4.23: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.40, CV: Small



Figures 4.1 - 4.24 show the distribution of the $\mu_S - \mu_{NS}$ estimates for all of the methods in each simulation. The bold red line represents the lab value estimate and the surrounding red lines represent the confidence interval about the lab value estimate. The EM algorithm distribution is consistently within the lab value confidence interval.

4.0.25 Conclusion

Often, lab values are not collected and researchers must rely on information from self-report values. This simulation study shows the dangers of not treating intentional inaccurate responders carefully in self-report data analysis.

When intentional inaccurate responses are ignored and all the self-report data is analyzed, the estimated progesterone difference is not close to the truth when CVs are small, no matter the sample size or probability of inaccurate response. Using this method would cause a conclusion of a larger difference in progesterone than that of the truth.

If investigators chose to ignore the noningorable inaccurate data by removing the intentional inaccurate responses, once again, the difference of progesterone levels were overestimated. The precision of the this method became worse as the sample size increased.

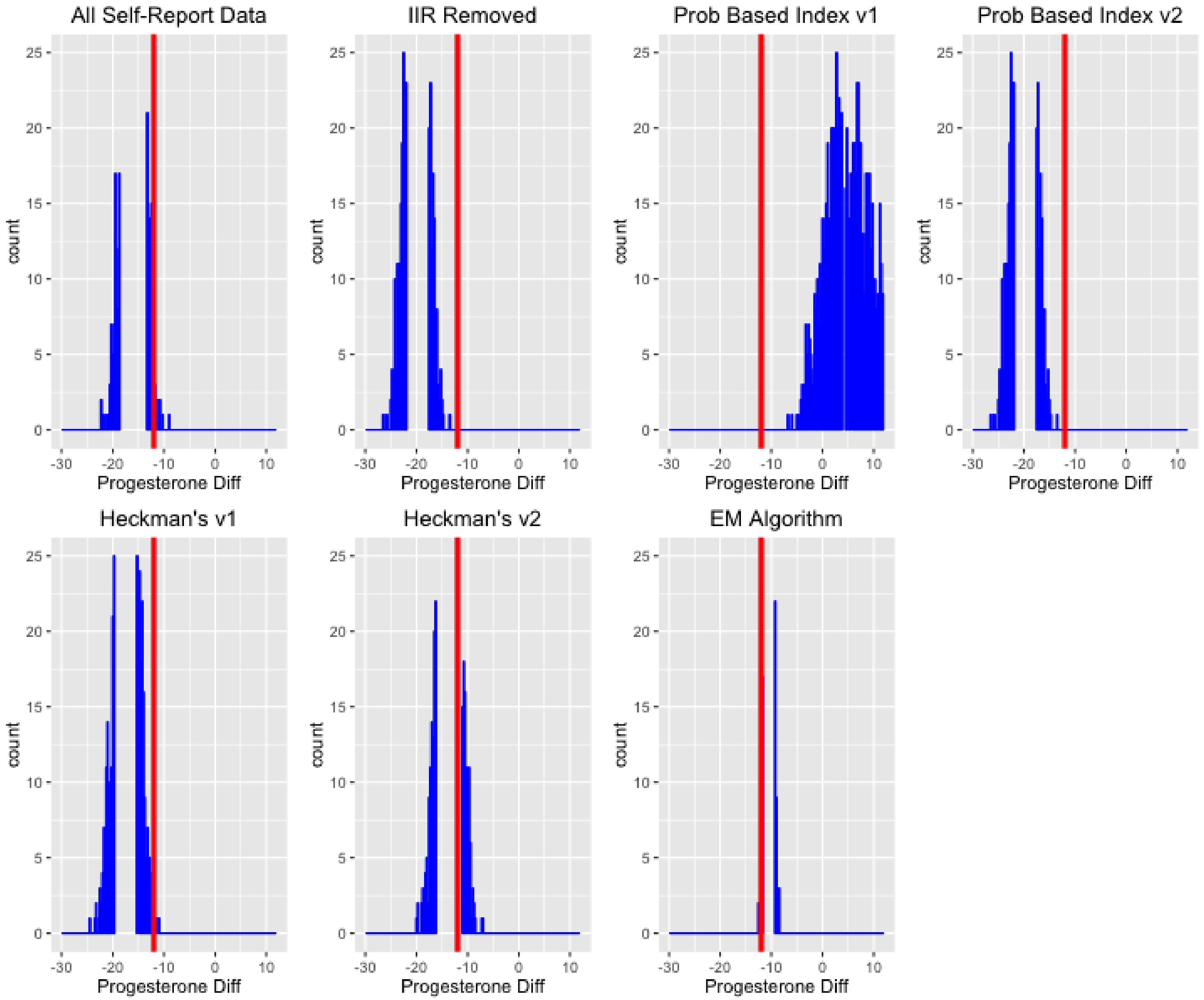
The concern for the probability-based index method is the discrepancy in estimates depending upon the calculation of the index. Version 1 represents a model using a covariate that does not capture the inaccurate responses well. Table 4.4 shows reverse conclusions when the probability of inaccurate response is high. This could lead to wildly incorrect conclusions. Version 2 represents a model using a covariate the does explain the intentional inaccuracies. Version 2 of the probability-based index ability to estimate progesterone differences is comparable to the current methods.

Heckman's model has a similar situation to the probability-index method. In version 1, the inverse mills ratio calculated from $x_{i2}\beta_2$ in the selection equation did not account very well for the intentional inaccurate responders. Every simulation in which the CV was small, version 1 confidence intervals did not contain the lab value estimate. Version 2 confidence intervals contained the lab value estimate more often. Both versions; however, had very large confidence intervals when the sample size was 55 and a large CV.

Of all the methods, the EM algorithm provided the closest estimation to the lab value. The success of the EM algorithm depends on the CV for larger sample sizes, 550 and 1100. It is documented that EM algorithm estimates are not as precise when component densities in the mixture are not well separated. Redner and Walker (1984) It is of no surprise that the EM algorithm performs better with small CVs. Even with large CVs; however, the EM algorithm and lab value confidence intervals still overlapped. This did not change with sample size or probability of intentional inaccurate response.

In conclusion, as long as the probability of intentional inaccurate response is low, the EM algorithm is an overall better method for estimating the difference in progesterone levels between smokers and nonsmokers no matter the sample size or CV.

Figure 4.24: Distribution of the $\mu_S - \mu_{NS}$ estimates. Sample Size: 1100, Probability of Intentional Inaccurate Response: 0.40, CV: Large



Chapter 5

Application

Smoking prevalence among pregnant women is usually recorded by self-report. Cnattingius (2004) Due to the sensitive nature of asking about smoking status during pregnancy, participants may consciously minimize their actual cigarette use Klesges et al. (1995). Since smoking status can now be validated using biochemical markers, such as cotinine, research shows that pregnant women diminish their amount of smoking. More nonsmokers are consistently seen than exist in the population. This, in turn, causes the data to be biased towards nonsmoking. Respondents who willingly do not answer truthfully have been named intentional inaccurate responders.

Robinson-Cimpian Robinson-Cimpian (2014) shows examples of the dangers these participants play in causing incorrect conclusions from the data. Robinson-Cimpian's work centered about adolescent self-report where these mischievous responders typically answered questions incorrectly because they thought it was funny. This idea can be extended to the clinical setting where participants do not feel comfortable answering the truth. Here, participants are not randomly answering inaccurately, but purposefully. Previously when intentional inaccurate responses were identified, they were either ignored or removed from the dataset. Since the intentional inaccurate responses are not making a random mistake, the inaccuracies needs to be accounted for in statistical modeling. Robinson-Cimpian's paper calculates a Probability-Based Index and uses it as a covariate in the regression equations in order to adjusted for the mischievous responses. Incomplete data techniques will also be used to consider the non-ignorable intentional inaccurate responses. The techniques consist of Heckman's model and the EM algorithm for finite mixtures.

The focus of this paper is on three groups of pregnant women. The first group are smokers who self-report that they are smokers. The second group are smokers who self-report they are non-smokers, and the last group are non-smokers who self-report they are non-smokers. The group of smokers who self-reported as non-smokers are considered intentionally inaccurate. A variety of methods will be looked at in order to determine the best approach for handling these intentional inaccurate responses.

5.0.26 Methods

In 2009, data were created to better understand prenatal passive smoke exposure and birth outcomes. The study recruited pregnant women from each of the following institutions and academic site: University of Kentucky (Lexington, KY), University of Virginia (Charlottsville, VA) and Norton Healthcare (Louisville, KY). Data were collected through a survey administered at each trimester and postpartum. This study received funding from the National Center for Research Resources. 307 women were recruited to the study and clinical lab values for serum cytokines were included in the data collection. At each trimester, smoking status was validated by urine cotinine levels. Intentional inaccurate responders were identified if the self-report smoking status data did not match the lab values from the cotinine. The focus of this analysis is on the 108 participants with which there exist data for the self-report smoking question, the urine cotinine, and progesterone levels.

The main question of interest is if smoking status has an effect on progesterone level. The lab values are used as a basis for comparison to methods that could be used if lab values did not exist. To answer this question, a variety of methods were engaged. First, current methods of linear regressions were run, followed by methods to account for intentional inaccurate responders, such as, a Probability-Based Index, Heckman's model, and the EM algorithm. All analyses were complete using R 3.2.2 (R Core Team, Vienna, Austria). Significance is defined as $p < 0.05$.

5.0.27 Measures of Tobacco Use

Pregnant women participating in the survey were asked: 'Do you currently smoke cigarettes or use smokeless tobacco (loose leaf, dip, chew, snuff) even just once in a while?'. Besides their self-reported answers from the survey, participants agreed to supply a urine sample. Smoking was confirmed by urine cotinine using $NicAlert^{TM}$ strips. Comparing the self-report values to the $NicAlert^{TM}$ value, it was evident that there were participants who self-reported as non-smokers; although, their $NicAlert^{TM}$ level indicated otherwise. As smoking during pregnancy is highly discouraged by health care providers, smokers are likely to answer that they were not currently smoking. Ford et al. (1997) According to $NicAlert^{TM}$, lab values ranged from 0 to 6. Level 3 or higher (level 4, 5, or 6) indicated use of tobacco products, level 0 indicated no detectable level of cotinine or tobacco product use, and levels 1 and 2 indicated no use of tobacco products. Based on these classifications, a participant whose $NicAlert^{TM}$ value was 0-2 was considered a non-smoker and a participant whose $NicAlert^{TM}$ value was 3-6 was considered a smoker. Table 5.1 shows the comparison of answers between self-report and the true $NicAlert^{TM}$ values. Of the women who claimed to be non-smokers, 89 were non-smokers, but 10 of the women were actually smokers. We are considering these 10 women as intentional inaccurate responders.

Table 5.1: Accuracy of Smoking Status

Self-Report	$NicAlert^{TM}$ Value		
Frequency	Non-Smoker	Smoker	Total
Non-Smoker	89	10	99
Smoker	0	9	9
Total	89	19	108

Unfortunately, knowing the truth is not always accessible. Failure to account for these intentional inaccurate responses in statistical analysis can result in incorrect conclusions. For instance, in this dataset, one of the cytokines of interest was progesterone. Progesterone helps nurture the fetus. After 8 to 10 weeks of pregnancy, the placenta takes over progesterone production from the ovaries and substantially increases progesterone production. This helps maintain a supportive environment for the developing fetus. Siiteri et al. (1977) Investigators wanted to know if smoking status has an effect on levels of progesterone. In Table 5.2, the self-reported smoking values show no significant effect on the levels of progesterone. However, the $NicAlert^{TM}$ values tell a different story. Here, smokers have a statistically significant lower levels of progesterone.

Table 5.2: Comparison of Self-Report Smoking to $NicAlert^{TM}$ values

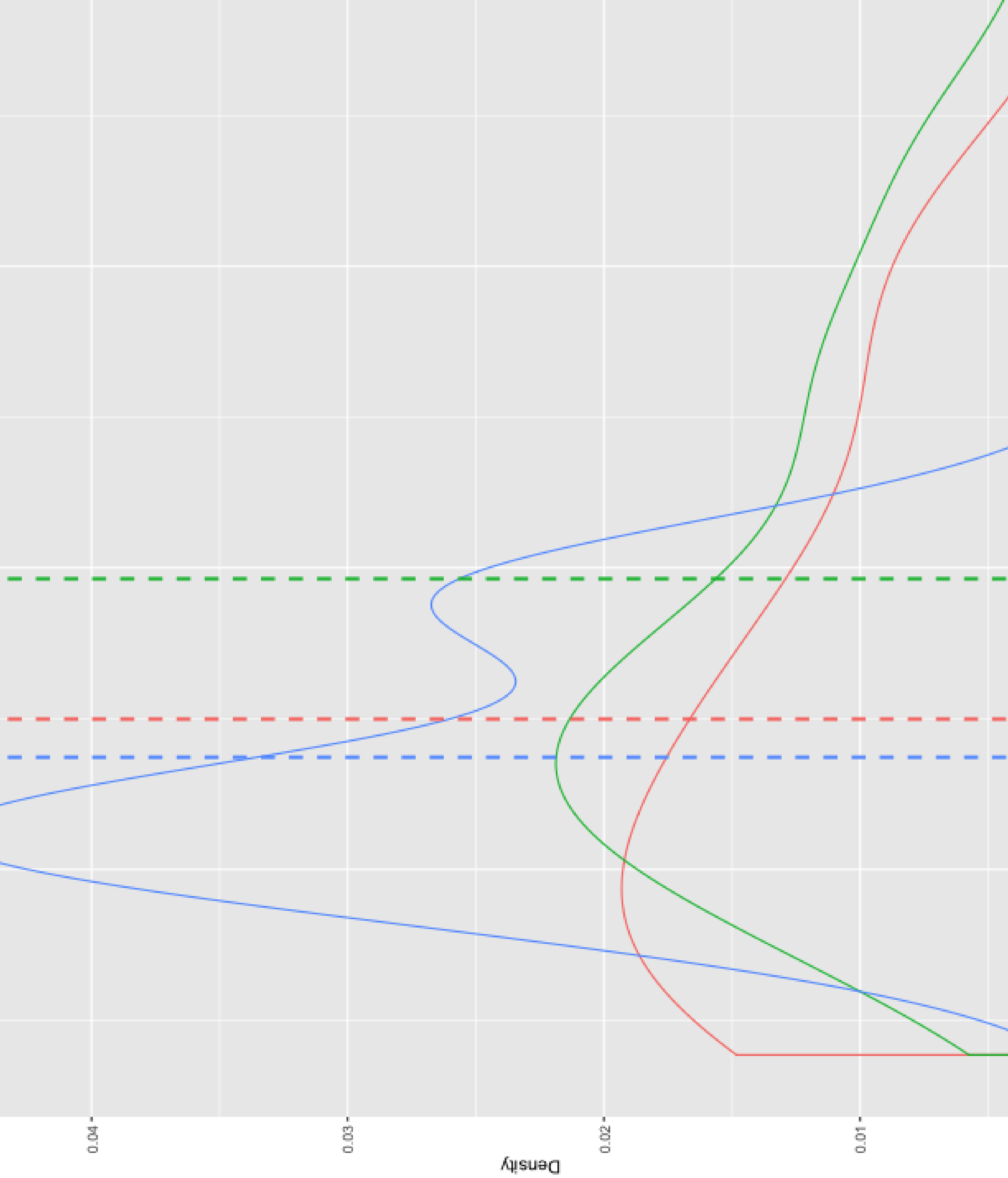
Effect	Smoker	Mean Progestrone (Standard Error)	p-value
Self-Report	1	47.44 (6.3)	0.1009
	0	58.33 (1.9)	
$NicAlert^{TM}$	1	48.77 (4.3)	0.0285
	0	59.27 (2.0)	

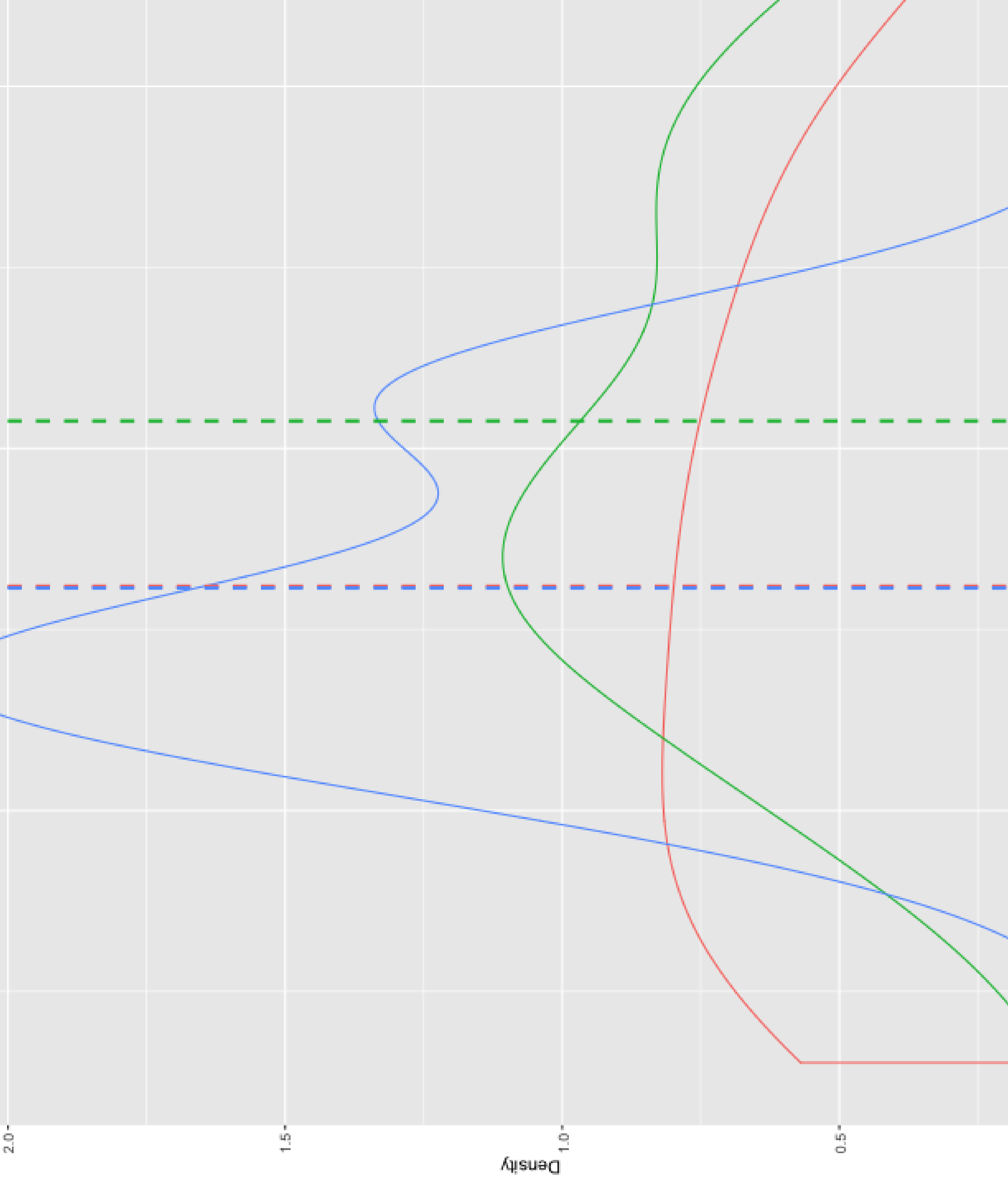
The kernel densities of the smoking status data can be seen in Figure 5.1. Since this data did not appear normally distributed, log-transformed progesterone levels were analyzed. The kernel densities of the log-transformed progesterone levels can be seen in Figure 5.2.

5.0.28 Intentional Inaccurate Response - Current Methods

Often, intentional inaccurate responses are ignored and all the self-report values are analyzed or the intentional inaccurate responses are identified and removed from the data before analysis. Cornell et al. (2012) Both of these current methods cause bias towards an incorrect answer since the intentional inaccurate responses do not occur randomly.

Linear regressions were run replicating these scenarios. First, all self-report values were used in a regression to predict progesterone. To finish the comparison with current methods, a regression was employed with self-report values, but the intentional inaccurate responses removed.


0.04
0.03
0.02
0.01
Density



Density
2.0
1.5
1.0
0.5

5.0.29 Intentional Inaccurate Resonse - Proposed Methods

Three alternative methods are proposed: Probability-Based Index, Heckman's model, and the EM algorithm.

Probability-Based Index.— The Probability-Based Index was introduced by Robinson-Cimpian. Robinson-Cimpian (2014) In his paper, he used it as a measure of the prevalence of low-frequency responses on screener items. This measure is similar to a weight given to each participant. These indices were included in statistical models and used as a covariate in the regression equations in order to adjusted for the intentional inaccurate responses.

With the smoking status data, two versions of the Probability-Based Index were considered. The purpose of the two versions was to see how much an affect the index had on the estimate depending on the quality of it's creation. The first version randomly assigned weight to each participant. This version represents an index that does not capture the intentional inaccurate responses well. The second version calculated a weight that was used as a measure of the prevalence of inconsistency with the self-report smoking variable. This version represents an index that accounts for the intentional inaccurate responses correctly. For both versions, a Probability-Based Index value was calculated for each participant and was used as a covariate in the regression equations in order to adjusted for the intentional inaccurate responses.

Heckman's Model.— Heckman's Sample Selection Model was created to handle censored or truncated dependent variables and is often used to handle non-ignorable incomplete data. With the smoking status data, Heckman's model is used to control for biases that could arise from the intentional inaccurate responses. Heckman's model is a two-step method. Consider the following equations.

$$\begin{aligned} y_{i1} &= x_{i1}\beta_1 + \epsilon_{i1} \\ y_{i2} &= x_{i2}\beta_2 + \epsilon_{i2}. \end{aligned}$$

When comparing the self-report values to the $NicAlert^{TM}$ values, some participants' classification did not agree. Participants were label as truthful or inaccurate. Using their $NicAlert^{TM}$ levels and truth status, a smoking score was created by summing each participants cotinine level and intentional inaccurate responder status. This smoking score was a continuous variable. It was then determined which variables predicted a participants smoking score. These explanatory variables were used in the selection model portion of Heckman's model.

The first step of Heckman's model, which is the selection equation, defines a dichotomous variable that indicates which group the participant is categorized:

$$\begin{aligned} y_{i2} &= x_{i2}\beta_2 + \epsilon_{i2} \\ y_{i2}* &= \begin{cases} 1, \text{if } y_{i2} \geq 0 \\ 0, \text{if } y_{i2} < 0 \end{cases} \end{aligned}$$

where y_{i2} is a smoking score, $y_{i2}*$ is an indicator for truth status, the x_{i2} are the explanatory variables of the smoking score, β_2 is a vector of parameter estimates, and ϵ_{i2} is an error term having a standard normal distribution. The first stage estimates β_2 using the probit maximum likelihood method. The second stage involves estimating an OLS regression of the response variable conditional on $y_{i2}*$ and x_{i1}. Vance and Buchheim (2005) Therefore,

$$E[y_{i1}|y_{i2}* = 1, x_{i1}] = x_{i1}\beta_1 + E[\epsilon_{i1}|y_{i2}* = 1] = x_{i1}\beta_1 + E[\epsilon_{i1}|\epsilon_{i2} \geq -x_{i2}\beta_2]$$

where y_{i1} is the progesterone levels and x_{i1} is the self-report smoking status variable of y_{i1}, β_1 is a vector of parameter estimates, and ϵ_{i1} is an error term having a standard normal distribution. Let ϵ_{i1} and ϵ_{i2} be correlated by ρ. Now,

$$E[y_{i1}|y_{i2}* = 1, x_{i1}] = x_{i1}\beta_1 + \rho\sigma_1\lambda_i$$

where σ_1 are the error variances of the OLS model, λ_i is the inverse Mills ratio such that

$$\lambda_i = \frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)}$$

where $\frac{\phi(x_{i2}\beta_2)}{\Phi(x_{i2}\beta_2)}$ is defined by the ratio of the density function of the normal distribution, ϕ, to its cumulative distribution function, Φ.

Two versions of Heckman's model were also considered. The first smoking score was created so that the scores were randomly assigned to each participant. The second smoking score was created so that honest participants were given a higher score than intentional inaccurate responders. It was then determined which variables predicted each of the smoking scores. The cutoff value for $y_{i2}*$ was chosen so that the majority of predicted y_{i2} values were appropriately identifying participants as truthful or inaccurate. Once x_{i2} were found, these variables were used in the second step of Heckman's model to estimate λ_i.

The EM Algorithm.— Dempster et al Dempster et al. (1977) use the EM algorithm to account for non-ignorable incomplete data. Little & Rubin Little and Rubin (2014) used the EM algorithm to account for non-ignorable incomplete data. In a similar manner, intentional inaccurate responses can be accounted for synonymous to incomplete data. In non-ignorable incomplete data, the data are not incomplete at random. In the smoking status data, the intentional inaccurate responses are not inaccurate at random. Participants consciously answered the smoking status question incorrectly. There is a pattern to the inaccurate responses that must be modeled. Since the incorrect responses are not inaccurate at random, the likelihood estimation needs to model the intentional inaccurate responses.

The smoking status data shows the self-report nonsmokers who are inaccurate seem have a different distribution than the self-report nonsmokers who reported the truth. For honest nonsmokers the mean progesterone is 59.27 with a variance of 377.8; on the other hand, intentional inaccurate nonsmokers have a mean measurement of 49.97 and a variance of 358.7, see Table 5.3.

With the transformed data honest nonsmokers the mean progesterone is 4.03 with a variance of 0.10; on the other hand, intentional inaccurate nonsmokers have a mean measurement of 3.85 and a variance of 0.14, see Table 5.4.

Table 5.3: Comparison of Self-Report Smoking to $NicAlert^{TM}$ values for Progesterone

Smokers	Inaccurate Nonsmokers	Honest Nonsmokers
47.44 (72.2)	49.97 (358.7)	59.27 (377.8)

True Smokers	True Nonsmokers
48.77 (213.1)	59.27 (377.8)

Besides the self-report nonsmokers, those participants who self-reported as smokers were also included in estimating progesterone differences in smoking status. Since this is an instance when inaccurate responders possibly have a different distribution than honest responders for progesterone level, the EM algorithm was be used to estimate the true mean for smokers (μ_S), the true mean for nonsmokers (μ_{NS}), and the probability of being an intentional inaccurate responder (π).

Consider two states for the self-reported non-smokers. For example,

Table 5.4: Comparison of Self-Report Smoking to $NicAlert^{TM}$ values for Log-transformed Progesterone

Smokers	Inaccurate Nonsmokers	Honest Nonsmokers
3.85 (0.03)	3.85 (0.14)	4.03 (0.10)

True Smokers	True Nonsmokers
3.85 (0.08)	4.03 (0.10)

$$\begin{aligned} z_1 &= (1,0) \rightarrow \text{intentionally inaccurate} \\ z_2 &= (0,1) \rightarrow \text{honest} \end{aligned}$$

Let y be progesterone level and n be the number of participants to self-report as nonsmokers. Using the EM methods from Dempster et al Dempster et al. (1977),

$$\begin{aligned} z_i &\sim Bern(r|\pi) \\ y_i|z_i &\sim u(y_i|r,\theta) \\ U(y|\theta) &= (\log u(y_i|(1,0),\theta), \log u(y_i|(0,1),\theta)) \\ V(\pi) &= (\log Bern((1,0)|\pi), \log Bern((0,1)|\pi)) \end{aligned}$$

Now, the log-likelihood for self-reported nonsmoker is

$$\log f_{NS}(x|\theta,\pi) = \sum_{i=1}^{n} z_i^T U(y_i|\theta) + z_i^T V(\pi)$$

Therefore, let m be the number of self-reported smokers and y be progesterone level. Assume the y_i are independently and identically drawn from a density $w(...|\theta)$. Let

$$W(y|\theta) = \log w(y|\theta).$$

Now, the log-likelihood for the self-reported smokers is

$$\log f_S(x|\theta) = \sum_{j=1}^{m} W(y_i|\theta).$$

Thus, the complete data log-likelihood is

$$\log f(x|\theta,\pi) = \sum_{j=1}^{m} W(y_i|\theta) + \sum_{i=1}^{n} z_i^T U(y_i|\theta) + z_i^T V(\pi).$$

Using this complete data log-likelihood, the EM algorithm was used to estimate the true mean for smokers (μ_S), the true mean for nonsmokers (μ_{NS}), and the probability of being an intentional inaccurate responder (π).

Meng & Rubin Meng and Rubin (1991) supplemented the EM algorithm with asymptotic variance-covariance matrix for estimates. All standard errors from the EM algorithm were calculated using the SEM.

Using the distributions from Table 5.5, the expectation of the smokers and nonsmokers were maximized iteratively until the estimate became stable at $1e-7$.

Table 5.5: Distribution of Progesterone Levels for 3 groups

Status	Mean (Variance)	n
Self-Report Non-smoker, but Smoker	3.85 (0.14)	10
Self-Report Non-smoker, True Non-smoker	4.03 (0.10)	89
Self-Report Smoker, True Smoker	3.85 (0.03)	9

5.0.30 Results

A linear regression was run that used smoking status as determined by $NicAlert^{TM}$ values and represents the truth. Using the truth, on average the transformed progesterone difference between smokers and non-smokers was -0.18. Smokers had a statistically significant, $p = 0.0244$, lower progesterone level compared to non-smokers. Next, a regression using smoking status determined by self-report values was considered. The average the progesterone difference between smokers and non-smokers was -0.17. However, due to the variability of the progesterone levels within non-smokers, the difference was not significant, $p = 0.1419$. Typically, only self-reported values are collected, but sometimes intentional inaccurate responses can be identified and removed. Removing the inaccurate self-report values, the difference in progesterone levels was about the same at -0.18, but was not found to be statistically significant, $p = 0.0959$. The comparisons of these regressions can be seen in Table 5.6.

Table 5.6: Comparison of Methods

Method	$\hat{\mu_S} - \hat{\mu_{NS}}$ (95% Confidence Interval)	p-value
Lab Values	-0.18 (-0.30,0.00)	0.0244
Self-Report	-0.17 (-0.40,0.10)	0.1419
Inaccurate Resp Removed	-0.18 (-0.40,0.00)	0.0959
Prob Based Index v1	-0.16 (-0.40,0.10)	0.1469
Prob Based Index v2	-0.18 (-0.40,0.00)	0.1017
Heckman's v1	-0.23 (-0.50,0.10)	0.1406
Heckman's v2	-0.13 (-0.40,0.10)	0.2722
EM	-0.16 (-0.40,0.00)	0.3471

As for the Probability-Based Index method, the average difference in progesterone levels for the first version is -0.16 and not significant, $p = 0.1469$. Similar results were seen in the second version with an estimated difference of -0.18 and $p = 0.1017$, see Table 5.6.

Using Heckman's model to analyze the self-report data, we find that the progesterone levels among self-reported smokers and non-smokers are not statistically different, $p = 0.1406$ and $p = 0.2722$; however, the difference is more than the true difference at -0.23 for the first version, but less than the true difference at -0.13 for the second version, see Table 5.6.

The EM algorithm estimates for the means and variances compared to the smoking status data means and variances can be seen in Table 5.7. The EM algorithm estimate for the mean progesterone for true smokers, 3.89, was close to the actual mean progesterone for smokers, 3.85. However, the EM algorithm estimated mean progesterone for nonsmokers is slightly higher than the actual mean, 4.05 and 4.03, respectively. When using the EM algorithm, the average difference in progesterone levels among smokers and non-smokers is -0.16 and not significant with $p = 0.3471$, see Table 5.6. With this method, the probability of an intentional inaccurate responder was also estimated, see Table 5.8.

Table 5.7: Comparison of EM estimates to Smoking Status Data

	True Smokers	True Nonsmokers
Pregnant Women Data	3.85 (0.08)	4.03 (0.10)
EM Algorithm Estimation	3.89 (0.08)	4.05 (0.11)

Table 5.8: EM algorithm estimate for probability of intentional inaccurate responder

	π	$\hat{\pi}$
Prob Intentional Inaccurate Responders	0.10	0.05 (0.04, 0.06)
Prob Not Intentional Inaccurate Responders	0.90	0.95 (0.94, 0.96)

5.0.31 Discussion

When intentional inaccurate responses are ignored and all the self-report data is analyzed, the estimated progesterone difference is larger and more variable than the truth. Using this method would cause a conclusion of a larger difference in progesterone level with no significance.

If investigators chose to ignore the noningorable inaccurate data by removing intentional inaccurate responses, once again, the difference of progesterone levels is not significance.

The concern for the Probability-Based Index method is the discrepancy in estimates depending upon the calculation of the index. Here the estimated difference was overestimated with no significance.

Heckman's model needed a continuous response variable in the selection equation. We were able to calculate a smoking score; however, this may not be available in all situations. Further, if the smoking score is not correctly capturing the intentional inaccuracies, this can create estimates far from the truth. The OLS estimates in the second step are very sensitive to the selection criteria.

Figure 5.1 shows the large variability in progesterone levels for honest nonsmokers and intentional inaccurate nonsmokers. This dispersion caused the respective distributions to overlap. Had the distributions of the self-report intentional inaccurate nonsmokers and the self-report honest nonsmokers been more distinct, the EM algorithm would have calculated estimates closer to the truth. However, the variances within these two groups of pregnant women were quite large, making it too difficult for the EM algorithm to specify between the inaccurate responses and the truthful responses.

Chapter 6

Conclusion and Future Research

As long as self-report data collection exists, intentional inaccurate responses will be an issue. This dissertation attempted to tackle this problem by providing methods for statistical inference by accounting for intentional inaccurate responses without removing data.

6.0.32 Conclusion

Using simulations and real data, the EM algorithm method focused on three groups of pregnant women. The first group were honest smokers, the second group were intentionally inaccurate nonsmokers, and the last group were honest nonsmokers.

Chapter 4 presented a simulation study that showed the dangers of not treating intentional inaccurate responders carefully in self-report data analysis. When intentional inaccurate responses are ignored and all the self-report data is analyzed, the estimated progesterone difference is not close to the truth when the CVs are small, no matter the sample size or probability of inaccurate response. Using this method would cause a conclusion of a larger difference in progesterone than that of the truth.

If investigators chose to ignore the non-ignorable inaccurate data by removing the intentional inaccurate responses, once again, the difference of progesterone levels were overestimated. The precision of the this method became worse as the sample size increased.

The concern for the Probability-Based index method is the discrepancy in estimates depending upon the calculation of the index. Version 1 represents a model using a covariate that does not capture the inaccurate responses well. Table 4.4 shows reverse conclusions when the probability of inaccurate response is high. This could lead to wildly incorrect conclusions. Version 2 represents a model using a covariate the does explain the intentional inaccuracies. Version 2 of the Probability-Based Index ability to estimate progesterone differences is comparable to the current methods.

Heckman's model has a similar situation to the Probability-Based Index method. In version 1, the inverse mills ratio calculated from $x_{i2}\beta_2$ in the selection equation did not account for the intentional inaccurate responders very well. Every simulation in which the CV was small, version 1 confidence intervals did not contain the lab value estimate. Version 2 confidence intervals contained the lab value estimate more often. Both versions; however, had very large confidence intervals hen the sample size was 55 and a large CV.

Of all the methods, the EM algorithm provided the closest estimation to the lab value. The success of the EM algorithm depends on the CV for larger sample sizes, 550 and 1100. It is documented that the EM algorithm estimates are not as precise when component densities in the mixture are not well separated. Redner and Walker (1984) It is of no surprise that the EM algorithm performs better with a small CV. Even with large CVs; however, the EM algorithm and lab value confidence intervals still overlapped. This did not change with sample size or probability of intentional inaccurate response. It is important to note that the EM algorithm struggled to converge when the probability of intentional inaccurate response was high and a large CV.

In conclusion, the EM algorithm is an overall better method for estimating the difference in progesterone levels between smokers and nonsmokers even with different sample sizes, CVs, or probabilities of intentional inaccurate response.

Chapter 5 displayed analyses with real data. When intentional inaccurate responses are ignored and all the self-report data are analyzed, the estimated progesterone difference is larger and more variable than the truth. Using this method would cause a conclusion of a larger difference in progesterone level with no significance.

If investigators chose to ignore the noningorable inaccurate data by removing intentional inaccurate responses, the difference of progesterone levels is close, but with no significance.

The concern for the probability-based index method is the discrepancy in estimates depending upon the calculation of the index. Here the estimated difference was overestimated with no significance.

Heckman's model needed a continuous response variable in the selection equation. We were able

to calculate a smoking score; however, this may not be available in all situations. Further, if the smoking score is not correctly capturing the intentional inaccuracies, this can create estimates far from the truth. The OLS estimates in the second step are very sensitive to the selection criteria.
Figure 5.1 shows the large variability in progesterone levels for honest nonsmokers and intentional inaccurate nonsmokers. This dispersion caused the respective distributions to overlap. Had the distributions of the self-report intentional inaccurate nonsmokers and the self-report honest nonsmokers been more distinct, the EM algorithm would have calculated estimates closer to the truth. However, the variances within these two groups of pregnant women were quite large, making it too difficult for the EM algorithm to specify between the inaccurate responses and the truthful responses.

6.0.33 Future Research

In both Chapters 4 and 5, the probability of an intentional inaccurate response followed a Bernoulli distribution. Most likely, in practice, a set of covariates will be needed in order to determine this probability. It would be interesting to use a set of demographics or particular questions in the survey to predict the probability of an intentional inaccurate responder.
This dissertation focused on a continuous dependent variable. Since survey data can often be categorical, a natural extension to this research is to the look at using the EM algorithm for questions where the response variable is categorical. The examples used here assumed the response variable followed a normal distribution. A more compelling situation might be to see how the EM algorithm fairs assuming the dependent variable followed a multinomial distribution.
Also, this dissertation looked at intentional inaccurate responders that only existed in the self-report nonsmoker category. The next step would be to see how the EM algorithm predicts intentional inaccurate responses with multiple states. For example, instead of intentional inaccuracies in the nonsmokers, there could be inaccurate responses in the self-report smokers, too. Here, the self-report question was dichotomous, but this can be extended to questions with multiple answer options, resulting in many states.

Appendix A

Appendix

A.1 R code

R language:

```
library(sampleSelection)
library(ggplot2)
library(gridExtra)
library(xtable)
library(lsr)
library(functional)
library(plyr)

#number of simulations - 310
((qnorm(.975)*4.7)/(10.49*.05))^2

#simulate data
getmixdata=function(n,m,k,probs,muvec,sigmasqvec,trmuvec,trsigmasqvec)
{
popnumbers=sample(c(1:k),size=n,replace=T,prob=c(probs,1-probs))
yvec=rnorm(n,muvec[popnumbers],sqrt(sigmasqvec[popnumbers]))
yvec0=rnorm(m,trmuvec,sqrt(trsigmasqvec))
smokers=c(yvec[popnumbers==1],yvec0)
nonsmokers=c(yvec[popnumbers==2])

#probability based screening index
pbsi1=rep(0,n)
race1=rep(0,n)
mar1=rep(0,n)
edu1=rep(0,n)
inc1=rep(0,n)
w1depress1=rep(0,n)
age1=rep(0,n)
bmi1=rep(0,n)
edin1=rep(0,n)
sleep1=rep(0,n)
smoke1=rep(0,n)
nic1=rep(0,n)
for (i in 1:n)
{
if (popnumbers[i]==1)
{
pbsi1[i]=length(popnumbers[popnumbers==1])/(n+m)
race1[i]=1
mar1[i]=sample(c(0:1),size=1,prob=c(0.16,1-0.16))
edu1[i]=sample(c(0:1),size=1,prob=c(0.11,1-0.11))
inc1[i]=sample(c(0:2),size=1,prob=c(0.44,0.55,0))
w1depress1[i]=sample(c(0:1),size=1,prob=c(0.67,1-0.67))
age1[i]=sample(c(17:42),size=1,
prob=c(.1111,0,0,0,.1111,.1111,.1111,.2222,0,.1111,0,.1111,0,0,0,0,0,0,0,.1111,0,0,0
edin1[i]=sample(c(0:22),size=1,prob=c(.4444,.1111,0,0,0,0,.1111,0,0,0,.2222,0,0,.111
sleep1[i]=sample(c(0:17),size=1,prob=c(0,0,0,.1250,0,.25,0,0,.25,.1250,0,0,.25,0,0,0
bmi1[i]=sample(c(17:50),size=1,prob=c(.1250,0,0,.25,0,0,0,0,.1250,0,0,0,0,.3750,0,0,0
smoke1[i]=1
nic1[i]=6
}
else
{
pbsi1[i]=length(popnumbers[popnumbers==2])/(n+m)
race1[i]=sample(c(0:1),size=1,prob=c(0.16,1-0.16))
mar1[i]=sample(c(0:1),size=1,prob=c(0.27,1-0.27))
edu1[i]=sample(c(0:1),size=1,prob=c(0.11,1-0.11))
inc1[i]=sample(c(0:2),size=1,prob=c(0.14,0.15,0.71))
w1depress1[i]=sample(c(0:1),size=1,prob=c(0.94,1-0.94))
age1[i]=sample(c(17:42),size=1,
prob=c(0,0,.0337,.0449,.0674,.0449,.0337,.0787,.0787,.0899,.0787,.0674,.0449,.0337,.0
edin1[i]=sample(c(0:22),size=1,prob=c(.1395,.1628,.1860,.1628,.0930,.0814,.0233,.0465
sleep1[i]=sample(c(0:17),size=1,prob=c(.0125,.0125,.0250,.0250,.0625,.0625,.1125,.07
bmi1[i]=sample(c(17:50),size=1,prob=c(.0941,0,0,.3882,0,0,0,0,.2941,0,0,0,0,.1059,0,0
smoke1[i]=0
nic1[i]=sample(c(0:1),size=1,prob=c(.6067,.3933))
}
}
pbsi2=rep(m/(n+m),m)
race2=sample(c(0:1),size=m,replace=T,prob=c(0.16,1-0.16))
mar2=sample(c(0:1),size=m,replace=T,prob=c(0.27,1-0.27))
edu2=sample(c(0:1),size=m,replace=T,prob=c(0.11,1-0.11))
inc2=sample(c(0:2),size=m,replace=T,prob=c(0.14,0.15,0.71))
w1depress2=sample(c(0:1),size=m,replace=T,prob=c(0.94,1-0.94))
age2=sample(c(17:42),size=m,replace=T,prob=c(0,.10,0,.10,0,.30,0,.10,.10,0,.10,.10,0,
edin2=sample(c(0:22),size=m,replace=T,prob=c(.1111,0,0,.1111,0,0,.2222,0,0,.1111,.111
sleep2=sample(c(0:17),size=m,replace=T,prob=c(0,0,0,0,0,0,.1111,0,.1111,0,.1111,.111
bmi2=sample(c(17:50),size=m,replace=T,prob=c(.1250,0,0,.25,0,0,0,0,.1250,0,0,0,0,.375
nic2=sample(c(3:6),size=m,replace=T,prob=c(.6,.1,.1,.2))
smoke2=rep(1,m)
srnsmk=rep(0,n)
srsmk=rep(1,m)

prog2=c(yvec,yvec0)
pbsi=c(pbsi1,pbsi2)
race=c(race1,race2)
mar=c(mar1,mar2)
edu=c(edu1,edu2)
inc=c(inc1,inc2)
depress=c(w1depress1,w1depress2)
age=c(age1,age2)
edin=c(edin1,edin2)
sleep=c(sleep1,sleep2)
bmi=c(bmi1,bmi2)
smoke=c(smoke1,smoke2)
srsmoke=c(srnsmk,srsmk)
nic=c(nic1,nic2)
truediff=mean(smokers)-mean(nonsmokers)

score=rep(NA,length(smoke))
```

```
pbi=rep(NA,length(smoke))
for (k in 1:length(smoke))
{
if (srsmoke[k]==1 & smoke[k]==1)
{
score[k]=20
pbi[k]=1-pi
}
else if (srsmoke[k]==0 & smoke[k]==0)
{
score[k]=20
pbi[k]=1-pi
}
else if (srsmoke[k]==0 & smoke[k]==1)
{
score[k]=-20
pbi[k]=pi
}
}

bigmama=cbind(prog2,pbsi,race,mar,edu,inc,depress,age,edin,sleep,bmi,smoke,srsmoke,n
return(list(yvec,yvec0,pbsi1,pbsi2,srnsmk,srsmk,truediff,bigmama))
}

morecol<-function(data)
{
smokescore<-rep(NA,length(data$score))
smokescore2<-rep(NA,length(data$score))
for (h in 1:length(data$score))
{
smokescore[h]=data$nic[h]+data$score[h]+rnorm(1)
smokescore2[h]=data$nic[h]+data$pbi[h]+rnorm(1)
}
sel<-predict(lm(smokescore~data$inc+data$depress+data$mar+data$age))
sel2<-predict(lm(smokescore2~data$nic))
choose<-rep(NA,length(sel))
choose2<-rep(NA,length(sel2))
for (j in 1:length(sel))
{
if (sel[j]>quantile(sel,.25))
{
choose[j]=1
}
else
{
choose[j]=0
}
if (sel2[j]>quantile(sel2,.25))
{
choose2[j]=1
}
else
{
choose2[j]=0
}
}
final=cbind(data,smokescore,sel,choose,sel2,choose2)
return(final)
}


emmix=function(yvec,yvec2,k,startprob,start1,start2,start3,start4,tol=1e-04,pause=F)
{
n=length(yvec)
m=length(yvec2)
#starting values
muvec=start1
sigmasqvec=start2
probs=startprob
muvec2=start3
sigmasqvec2=start4

#expected value of unobserved data
ezmatnorm=matrix(0,nrow=n,ncol=k)
ezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
ezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
ezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
rowsums=apply(ezmatnorm,1,sum)
ezmat=ezmatnorm/rowsums
llik=sum(log(rowsums))

ezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
rowsums2=apply(ezmatnorm2,1,sum)
ezmat2=ezmatnorm2/rowsums2
llik2=sum(log(rowsums2))

rowsumssmk=sum(c(ezmatnorm[,1],ezmatnorm2[,1]))
rowsumsnsmk=sum(ezmatnorm[,2])
ezmatsmk=c(ezmatnorm[,1],ezmatnorm2[,1])/rowsumssmk
ezmatnsmk=ezmatnorm[,2]/rowsumsnsmk


repeat
{
#ezmat contains the ez_ij values
#llik contains the loglikelihood
bothprobs=apply(ezmat,2,mean)
probs=bothprobs[which.min(bothprobs)]
if (probs<0.01){probs<-0.05}
#ifelse(probs>0.01,probs,0.05)
muvec[1]=sum(ezmat[,1]*yvec)/sum(ezmat[,1])
if (muvec[1]<=0){muvec[1]<-0.001}
muvec[2]=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvec[1]=sum((ezmat[,1]*(yvec-muvec[1])^2))/sum(ezmat[,1])
sigmasqvec[2]=sum((ezmat[,2]*(yvec-muvec[2])^2))/sum(ezmat[,2])

muvec2=sum(ezmat2[,1]*yvec2)/sum(ezmat2[,1])
```

```
sigmasqvec2=sum((ezmat2[,1]*(yvec2-muvec2)^2))/sum(ezmat2[,1])

muvecsmk=sum(c(ezmat[,1]*yvec,ezmat2[,1]*yvec2))/sum(c(ezmat[,1]),ezmat2[,1])
muvecnsmk=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvecsmk=sum(c(ezmat[,1]*((yvec-muvec[1])^2),ezmat2[,1]*((yvec2-muvec2)^2))/sum
sigmasqvecnsmk=sum(ezmat[,2]*((yvec-muvec[2])^2))/sum(ezmat[,2])

#compute loglikelihood for new iteration
newezmatnorm=matrix(0,nrow=n,ncol=k)
newezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
newezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
for (k in 1:length(newezmatnorm[,1]))
{
if (newezmatnorm[k,1]<=0.001){newezmatnorm[k,1]<-0.001}
}
newezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
check=newezmatnorm[apply(newezmatnorm,1,Compose(is.finite, all)),]
newrowsums=apply(check,1,sum)
newezmat=newezmatnorm/newrowsums
newllik=sum(log(newrowsums))

newezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
newrowsums2=apply(newezmatnorm2,1,sum)
newezmat2=newezmatnorm2/newrowsums2
newllik2=sum(log(newrowsums2))

newrowsumssmk=sum(c(newezmatnorm[,1],newezmatnorm2[,1]))
newrowsumsnsmk=sum(newezmatnorm[,2])
newezmatsmk=c(newezmatnorm[,1],newezmatnorm2[,1])/newrowsumssmk
newezmatnsmk=newezmatnorm[,2]/newrowsumsnsmk

smkdiff=muvecsmk-muvecnsmk
w1=sigmasqvecsmk/length(c(ezmat[,1],ezmat2[,1]))
w2=sigmasqvecnsmk/length(ezmat[,2])
smkdiffvar=w1+w2
denom=((w1^2)/(length(c(ezmat[,1],ezmat2[,1]))-1))+((w2^2)/(length(ezmat[,2])-1))

if ((newllik-llik)<tol) {break}
ezmat=newezmat
llik=newllik
ezmat2=newezmat2
llik2=newllik2
ezmatsmk=newezmatsmk
ezmatnsmk=ezmatnsmk
}
return(list(probs=c(probs,1-probs),muvec=muvec,sigmasqvec=sigmasqvec,muvec2=muvec2,si
muvecsmk=muvecsmk,muvecnsmk=muvecnsmk,sigmasqvecsmk=sigmasqvecsmk,sigmasqvecnsmk=sigm
smkdiff=smkdiff,smkdiffvar=smkdiffvar,w1=w1,w2=w2,denom=denom))
}



#plots
```

```
plotiter=function(y,y2,probs,muvec,sigmasqvec,muvectr,sigmasqvectr,pause=F)
{
xtemp=seq(min(y),max(y),(max(y)-min(y))/1000)
compvec1=probs*dnorm(xtemp,muvec[1],sqrt(sigmasqvec[1]))
compvec2=(1-probs)*dnorm(xtemp,muvec[2],sqrt(sigmasqvec[2]))
xtemp2=seq(min(y2),max(y2),(max(y2)-min(y2))/1000)
compvec3=dnorm(xtemp2,muvectr,sqrt(sigmasqvectr))
return(list(xtemp=xtemp,compvec1=compvec1,compvec2=compvec2,xtemp2=xtemp2,compvec3=co
}


semmix=function(yvec,yvec2,k,startprob,start1,start2,start3,start4,emmuvecsmk,emmuvec
{
n=length(yvec)
m=length(yvec2)
#starting values
muvec=start1
sigmasqvec=start2
probs=startprob
muvec2=start3
sigmasqvec2=start4

#expected value of unobserved data
ezmatnorm=matrix(0,nrow=n,ncol=k)
ezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
ezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
ezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
rowsums=apply(ezmatnorm,1,sum)
ezmat=ezmatnorm/rowsums
llik=sum(log(rowsums))

ezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
rowsums2=apply(ezmatnorm2,1,sum)
ezmat2=ezmatnorm2/rowsums2
llik2=sum(log(rowsums2))

rowsumssmk=sum(c(ezmatnorm[,1],ezmatnorm2[,1]))
rowsumsnsmk=sum(ezmatnorm[,2])
ezmatsmk=c(ezmatnorm[,1],ezmatnorm2[,1])/rowsumssmk
ezmatnsmk=ezmatnorm[,2]/rowsumsnsmk


repeat
{
#ezmat contains the ez_ij values
#llik contains the loglikelihood
bothprobs=apply(ezmat,2,mean)
probs=bothprobs[which.min(bothprobs)]
if (probs<0.01){probs<-0.05}
muvec[1]=sum(ezmat[,1]*yvec)/sum(ezmat[,1])
if (muvec[1]<=0){muvec[1]<-0.001}
muvec[2]=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvec[1]=sum((ezmat[,1]*(yvec-muvec[1])^2))/sum(ezmat[,1])
```

```
sigmasqvec[2]=sum((ezmat[,2]*(yvec-muvec[2])^2))/sum(ezmat[,2])

muvec2=sum(ezmat2[,1]*yvec2)/sum(ezmat2[,1])
sigmasqvec2=sum((ezmat2[,1]*(yvec2-muvec2)^2))/sum(ezmat2[,1])

muvecsmk=sum(c(ezmat[,1]*yvec,ezmat2[,1]*yvec2))/sum(c(ezmat[,1]),ezmat2[,1])
muvecnsmk=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvecsmk=sum(c(ezmat[,1]*((yvec-muvec[1])^2),ezmat2[,1]*((yvec2-muvec2)^2))/sum
sigmasqvecnsmk=sum(ezmat[,2]*((yvec-muvec[2])^2))/sum(ezmat[,2])

#compute loglikelihood for new iteration
newezmatnorm=matrix(0,nrow=n,ncol=k)
newezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
newezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
for (k in 1:length(newezmatnorm[,1]))
{
if (newezmatnorm[k,1]<=0.001){newezmatnorm[k,1]<-0.001}
}
newezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
check=newezmatnorm[apply(newezmatnorm,1,Compose(is.finite, all)),]
newrowsums=apply(check,1,sum)
newezmat=newezmatnorm/newrowsums
newllik=sum(log(newrowsums))

newezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
newrowsums2=apply(newezmatnorm2,1,sum)
newezmat2=newezmatnorm2/newrowsums2
newllik2=sum(log(newrowsums2))

newrowsumssmk=sum(c(newezmatnorm[,1],newezmatnorm2[,1]))
newrowsumsnsmk=sum(newezmatnorm[,2])
newezmatsmk=c(newezmatnorm[,1],newezmatnorm2[,1])/newrowsumssmk
newezmatnsmk=newezmatnorm[,2]/newrowsumsnsmk


newmuvecsmk=sum(c(newezmat[,1]*yvec,newezmat2[,1]*yvec2))/sum(c(newezmat[,1]),newezm
newmuvecnsmk=sum(newezmat[,2]*yvec)/sum(newezmat[,2])
newsigmasqvecsmk=sum(c(newezmat[,1]*((yvec-muvec[1])^2),newezmat2[,1]*((yvec2-muvec2
newsigmasqvecnsmk=sum(newezmat[,2]*((yvec-muvec[2])^2))/sum(newezmat[,2])

newbothprobs=apply(newezmat,2,mean)
newprobs=newbothprobs[which.min(newbothprobs)]

r11=(newmuvecsmk-emmuvecsmk)/(muvecsmk-emmuvecsmk)
r12=(newsigmasqvecsmk-emsigmasqvecsmk)/(muvecsmk-emmuvecsmk)
r13=(newmuvecnsmk-emmuvecnsmk)/(muvecsmk-emmuvecsmk)
r14=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(muvecsmk-emmuvecsmk)
r15=(newprobs[1]-startprob)/(muvecsmk-emmuvecsmk)
r21=(newmuvecsmk-emmuvecsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r22=(newsigmasqvecsmk-emsigmasqvecsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r23=(newmuvecnsmk-emmuvecnsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r24=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r25=(newprobs[1]-startprob)/(sigmasqvecsmk-emsigmasqvecsmk)
```

```
r31=(newmuvecsmk-emmuvecsmk)/(muvecnsmk-emmuvecnsmk)
r32=(newsigmasqvecsmk-emsigmasqvecsmk)/(muvecnsmk-emmuvecnsmk)
r33=(newmuvecnsmk-emmuvecnsmk)/(muvecnsmk-emmuvecnsmk)
r34=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(muvecnsmk-emmuvecnsmk)
r35=(newprobs[1]-startprob)/(muvecnsmk-emmuvecnsmk)
r41=(newmuvecsmk-emmuvecsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r42=(newsigmasqvecsmk-emsigmasqvecsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r43=(newmuvecnsmk-emmuvecnsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r44=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r45=(newprobs[1]-startprob)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r51=(newmuvecsmk-emmuvecsmk)/(probs[1]-startprob)
r52=(newsigmasqvecsmk-emsigmasqvecsmk)/(probs[1]-startprob)
r53=(newmuvecnsmk-emmuvecnsmk)/(probs[1]-startprob)
r54=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(probs[1]-startprob)
r55=(newprobs[1]-startprob)/(probs[1]-startprob)


if ((newllik-llik)<tol) {break}
ezmat=newezmat
llik=newllik
ezmat2=newezmat2
llik2=newllik2
ezmatsmk=newezmatsmk
ezmatnsmk=ezmatnsmk
rmat=matrix(c(r11,r12,r13,r14,r15,r21,r22,r23,r24,r25,r31,r32,r33,r34,r35,r41,r42,r43
}
return(rmat)
}

se <- function(x)
{
B=length(x)
aveB=mean(x)
subtr=rep(NA,B)
for (i in 1:B)
{
subtr[i]=(x[i]-aveB)^2
}
return(sqrt(sum(subtr)/(B-1)))
}
##########################################################################################
simulation<-function(n,m,numsim,pi,var1,var2,es)
{
yvecsim=list()
emsim=list()
semsim=list()
data=list()
newdata=list()
mrremdata=list()
cdvcsim=list()
Vsim=list()
labval=rep(NA,numsim)
plabval=rep(NA,numsim)
selfrep=rep(NA,numsim)
```

```
pselfrep=rep(NA,numsim)
mrrem=rep(NA,numsim)
pmrrem=rep(NA,numsim)
pbsc=rep(NA,numsim)
ppbsc=rep(NA,numsim)
pbic=rep(NA,numsim)
ppbic=rep(NA,numsim)
eqn1=list()
eqn2=list()
eqn3=list()
eqn4=list()
heck=rep(NA,numsim)
pheck=rep(NA,numsim)
heck2=rep(NA,numsim)
pheck2=rep(NA,numsim)
emdiff=rep(NA,numsim)
pemdiff=rep(NA,numsim)
empi=matrix(rep(NA,2*numsim),nrow=numsim,ncol=2)
eta=rep(NA,numsim)
for (i in 1:numsim)
{
yvecsim[[i]]=getmixdata(n,m,2,pi,c(50.0,60.0),var1,40.0,var2)
data[[i]]=as.data.frame(yvecsim[[i]][[8]])
labval[i]<-coef(summary(glm(prog2~smoke,data=data[[i]])))[2,1]
plabval[i]<-coef(summary(glm(prog2~smoke,data=data[[i]])))[2,4]
selfrep[i]<-coef(summary(glm(prog2~srsmoke,data=data[[i]])))[2,1]
pselfrep[i]<-coef(summary(glm(prog2~srsmoke,data=data[[i]])))[2,4]
mrremdata[[i]]<-data[[i]][data[[i]]$score==20,]
mrrem[i]<-coef(summary(glm(prog2~srsmoke,data=mrremdata[[i]])))[2,1]
pmrrem[i]<-coef(summary(glm(prog2~srsmoke,data=mrremdata[[i]])))[2,4]
pbsc[i]<-coef(summary(glm(prog2~srsmoke+pbsi,data=data[[i]])))[2,1]
ppbsc[i]<-coef(summary(glm(prog2~srsmoke+pbsi,data=data[[i]])))[2,4]
pbic[i]<-coef(summary(glm(prog2~srsmoke+pbi,data=data[[i]])))[2,1]
ppbic[i]<-coef(summary(glm(prog2~srsmoke+pbi,data=data[[i]])))[2,4]
newdata[[i]]=morecol(data[[i]])
eqn1[[i]]<-glm(choose~bmi,family=binomial(link='probit'),data=newdata[[i]])
eqn3[[i]]<-glm(choose2~bmi,family=binomial(link='probit'),data=newdata[[i]])
newdata[[i]]$IMR<-dnorm(eqn1[[i]]$linear.predictors)/pnorm(eqn1[[i]]$linear.predictor
newdata[[i]]$IMR2<-dnorm(eqn3[[i]]$linear.predictors)/pnorm(eqn3[[i]]$linear.predicto
#print(cbind(newdata[[i]]$IMR,newdata[[i]]$sel,newdata[[i]]$choose))
eqn2[[i]]<-summary(lm(prog2~srsmoke+IMR,data=newdata[[i]],subset=(choose==1)))
eqn4[[i]]<-summary(lm(prog2~srsmoke+IMR2,data=newdata[[i]],subset=(choose2==1)))
heck[i]<-coef(eqn2[[i]])[2,1]
pheck[i]<-coef(eqn2[[i]])[2,4]
heck2[i]<-coef(eqn4[[i]])[2,1]
pheck2[i]<-coef(eqn4[[i]])[2,4]
emsim[[i]]=emmix(yvecsim[[i]][[1]],yvecsim[[i]][[2]],2,pi,c(50.0,60.0),var1,40.0,var2
emdiff[i]=emsim[[i]]$smkdiff
empi[i,]=emsim[[i]]$probs
semsim[[i]]=semmix(yvecsim[[i]][[1]],yvecsim[[i]][[2]],2,pi,c(55.0,65.0),var1,45.0,va
cdvcsim[[i]]=matrix(c((1/emsim[[i]]$sigmasqvecsmk),0,0,0,0,0,(1/(2*emsim[[i]]$sigmasq
0,0,0,0,0,(1/(2*emsim[[i]]$sigmasqvecnsmk^2)),0,0,0,0,0,
((1/emsim[[i]]$probs[1])+(1/(1-emsim[[i]]$probs[1])))),5,5)
#print(semsim[[i]])
```

```
if (diag(5)[1,1]-semsim[[i]][1,1]==1)
{Vsim[[i]]=matrix(rep(NA,25),ncol=5)}
else{Vsim[[i]]=solve(cdvcsim[[i]]) + solve(cdvcsim[[i]])%*%semsim[[i]]%*%solve(diag(5
#print(emsim[[i]]$smkdiff)
if (is.na(diag(Vsim[[i]])[1])== 'TRUE' | is.na(diag(Vsim[[i]])[3])== 'TRUE' | diag(Vsin
{pemdiff[i]==NA}
else{pemdiff[i]=2*pt(-abs(emsim[[i]]$smkdiff/(sqrt((diag(Vsim[[i]])[1]/m)+(diag(Vsim
#print(c(diag(Vsim[[i]]), diag(cdvcsim[[i]]), sqrt((diag(Vsim[[i]])[1]/m)+(diag(Vsim[[
eta[i]<-etaSquared(aov(data[[i]][,1]~data[[i]][,12]),type=3,anova=FALSE)[,1]
}

estimates<-c(mean(labval),mean(selfrep),mean(mrrem),mean(pbsc),mean(pbic),mean(heck)
seestimates<-c(se(labval),se(selfrep),se(mrrem),se(pbsc),se(pbic),se(heck),se(heck2)
cilow<-c(mean(labval)-qt(.975,(n+m-2))*se(labval),mean(selfrep)-qt(.975,(n+m-2))*se(
mean(mrrem)-qt(.975,(n+m-2))*se(mrrem),mean(pbsc)-qt(.975,(n+m-2))*se(pbsc),
mean(pbic)-qt(.975,(n+m-2))*se(pbic),mean(heck)-qt(.975,(n+m-2))*se(heck),
mean(heck2)-qt(.975,(n+m-2))*se(heck2),mean(emdiff)-qt(.975,(n+m-2))*se(emdiff),
mean(empi[,1])-qt(.975,(n+m-2))*se(empi[,1]),mean(empi[,2])-qt(.975,(n+m-2))*se(empi
cihigh<-c(mean(labval)+qt(.975,(n+m-2))*se(labval),mean(selfrep)+qt(.975,(n+m-2))*se
mean(mrrem)+qt(.975,(n+m-2))*se(mrrem),mean(pbsc)+qt(.975,(n+m-2))*se(pbsc),
mean(pbic)+qt(.975,(n+m-2))*se(pbic),mean(heck)+qt(.975,(n+m-2))*se(heck),
mean(heck2)+qt(.975,(n+m-2))*se(heck2),mean(emdiff)+qt(.975,(n+m-2))*se(emdiff),
mean(empi[,1])+qt(.975,(n+m-2))*se(empi[,1]),mean(empi[,2])+qt(.975,(n+m-2))*se(empi
numsig<-c(length(plabval[plabval<0.05]),length(pselfrep[pselfrep<0.05]),length(pmrren
length(ppbsc[ppbsc<0.05]),length(ppbic[ppbic<0.05]),length(pheck[pheck<0.05]),
length(pheck2[pheck2<0.05]),length(pemdiff[pemdiff<0.05 & !is.na(pemdiff)]),sum(is.na
persig<-c(length(plabval[plabval<0.05])/length(plabval),
length(pselfrep[pselfrep<0.05])/length(pselfrep),
length(pmrrem[pmrrem<0.05])/length(pmrrem),length(ppbsc[ppbsc<0.05])/length(ppbsc),
length(ppbic[ppbic<0.05])/length(ppbic),length(pheck[pheck<0.05])/length(pheck),
length(pheck2[pheck2<0.05])/length(pheck2),
length(pemdiff[pemdiff<0.05 & !is.na(pemdiff)])/(length(pemdiff)-sum(is.na(pemdiff)))
samplesize<-rep((n+m),length(estimates))
pis<-rep(pi,length(estimates))
effectsize<-rep(es,length(estimates))
method<-c('Lab␣Values','Self-Report','Misch␣Resp␣Removed','Prob␣Based␣Index␣v1','Prob
simtab<-cbind(samplesize,pis,effectsize,method,estimates,seestimates,cilow,cihigh,nur
return(list(yvecsim,labval,selfrep,mrrem,pbsc,pbic,heck,heck2,emdiff,empi,simtab,mear
}

sim.1<-simulation(50,5,1000,0.1,c(5,6),4,1)
sim.2<-simulation(50,5,1000,0.1,c(350,375),325,0)
sim.3<-simulation(50,5,1000,0.2,c(5,6),4,1)
sim.4<-simulation(50,5,1000,0.2,c(350,375),325,0)
sim.5<-simulation(50,5,1000,0.4,c(5,6),4,1)
sim.6<-simulation(50,5,1000,0.4,c(350,375),325,0)
sim.7<-simulation(100,10,1000,0.1,c(5,6),4,1)
sim.8<-simulation(100,10,1000,0.1,c(350,375),325,0)
sim.9<-simulation(100,10,1000,0.2,c(5,6),4,1)
sim.10<-simulation(100,10,1000,0.2,c(350,375),325,0)
sim.11<-simulation(100,10,1000,0.4,c(5,6),4,1)
sim.12<-simulation(100,10,1000,0.4,c(350,375),325,0)
sim.13<-simulation(500,50,1000,0.1,c(5,6),4,1)
sim.14<-simulation(500,50,1000,0.1,c(350,375),325,0)
sim.15<-simulation(500,50,1000,0.2,c(5,6),4,1)
sim.16<-simulation(500,50,1000,0.2,c(350,375),325,0)
sim.17<-simulation(500,50,1000,0.4,c(5,6),4,1)
sim.18<-simulation(500,50,1000,0.4,c(350,375),325,0)
sim.19<-simulation(1000,100,1000,0.1,c(5,6),4,1)
sim.20<-simulation(1000,100,1000,0.1,c(350,375),325,0)
sim.21<-simulation(1000,100,1000,0.2,c(5,6),4,1)
sim.22<-simulation(1000,100,1000,0.2,c(350,375),325,0)
sim.23<-simulation(1000,100,1000,0.4,c(5,6),4,1)
sim.24<-simulation(1000,100,1000,0.4,c(350,375),325,0)
sim<-rbind(sim.1[[11]],sim.2[[11]],sim.3[[11]],sim.4[[11]],sim.5[[11]],sim.6[[11]],si
sim.13[[11]],sim.14[[11]],sim.15[[11]],sim.16[[11]],sim.17[[11]],sim.18[[11]],sim.19

sim
write.csv(sim,"/Users/Kristen/Desktop/dissertation/simulationTable.csv")

sims<-cbind(sim[,1],sim[,2],sim[,3],sim[,4],format(round(as.numeric(sim[,5]),digits=
paste("(",format(round(as.numeric(sim[,7]),digits=2),nsmall=2),",",format(round(as.n
sim[,9],format(round(as.numeric(sim[,10]),digits=2),nsmall=2))
colnames(sims)<-c("Sample␣Size","Pi","Effect␣Size","Method","Estimate","95%␣CI","Num
write.csv(sim,"/Users/Kristen/Desktop/dissertation/simulationTable2.csv")
print(xtable(sims),include.rownames=FALSE)

est<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_est.csv",header=T
print(xtable(est),include.rownames=FALSE)

numsig<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_numsig.csv",hea
print(xtable(numsig),include.rownames=FALSE)

pipi<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_pi.csv",header=T
print(xtable(pipi),include.rownames=FALSE)

notcon<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_notconverge.cs
print(xtable(notcon),include.rownames=FALSE)


sim.25<-simulation(50,5,1000,0.1,c(100,110),90,1)
sim.26<-simulation(50,5,1000,0.1,c(4500,5000),4000,0)
sim.27<-simulation(50,5,1000,0.2,c(100,110),90,1)
sim.28<-simulation(50,5,1000,0.2,c(4500,5000),4000,0)
sim.29<-simulation(50,5,1000,0.4,c(100,110),90,1)
sim.30<-simulation(50,5,1000,0.4,c(4500,5000),4000,0)
sim.31<-simulation(100,10,1000,0.1,c(100,110),90,1)
sim.32<-simulation(100,10,1000,0.1,c(4500,5000),4000,0)
sim.33<-simulation(100,10,1000,0.2,c(100,110),90,1)
sim.34<-simulation(100,10,1000,0.2,c(4500,5000),4000,0)
sim.35<-simulation(100,10,1000,0.4,c(100,110),90,1)
sim.36<-simulation(100,10,1000,0.4,c(4500,5000),4000,0)
sim.37<-simulation(500,50,1000,0.1,c(100,110),90,1)
sim.38<-simulation(500,50,1000,0.1,c(4500,5000),4000,0)
sim.39<-simulation(500,50,1000,0.2,c(100,110),90,1)
sim.40<-simulation(500,50,1000,0.2,c(4500,5000),4000,0)
sim.41<-simulation(500,50,1000,0.4,c(100,110),90,1)
sim.42<-simulation(500,50,1000,0.4,c(4500,5000),4000,0)
sim.43<-simulation(1000,100,1000,0.1,c(100,110),90,1)
sim.44<-simulation(1000,100,1000,0.1,c(4500,5000),4000,0)
sim.45<-simulation(1000,100,1000,0.2,c(100,110),90,1)
sim.46<-simulation(1000,100,1000,0.2,c(4500,5000),4000,0)
sim.47<-simulation(1000,100,1000,0.4,c(100,110),90,1)
sim.48<-simulation(1000,100,1000,0.4,c(4500,5000),4000,0)
sim2<-rbind(sim.25[[11]],sim.26[[11]],sim.27[[11]],sim.28[[11]],sim.29[[11]],sim.30[
sim.37[[11]],sim.38[[11]],sim.39[[11]],sim.40[[11]],sim.41[[11]],sim.42[[11]],sim.43
write.csv(sim2,"/Users/Kristen/Desktop/dissertation/simulationTable2ndrun.csv")
sim2s<-cbind(sim2[,1],sim2[,2],sim2[,3],sim2[,4],format(round(as.numeric(sim2[,5]),d
paste("(",format(round(as.numeric(sim2[,7]),digits=2),nsmall=2),",",format(round(as.
sim2[,9],format(round(as.numeric(sim2[,10]),digits=2),nsmall=2))
colnames(sim2s)<-c("Sample␣Size","Pi","Effect␣Size","Method","Estimate","95%␣CI","Nu
write.csv(sim2,"/Users/Kristen/Desktop/dissertation/simulationTable22ndrun.csv")
print(xtable(sim2s),include.rownames=FALSE)

est2<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_est2ndrun.csv",h
print(xtable(est2),include.rownames=FALSE)

numsig2<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_numsig2ndrun.
print(xtable(numsig2),include.rownames=FALSE)

pipi2<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_pi2ndrun.csv",h
print(xtable(pipi2),include.rownames=FALSE)

notcon2<-read.csv("/Users/Kristen/Desktop/dissertation/simulationTable_notconverge2n
print(xtable(notcon2),include.rownames=FALSE)

mygraphs<-function(type1,type2,type3,type4,type5,type6,type7,xlimits,ylimits,xint,xlo
{
plot1<-qplot(type1,geom="histogram",binwidth = bw,main = "All␣Self-Report␣Data",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

plot2<-qplot(type2,geom="histogram",binwidth = bw,main = "IIR␣Removed",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

plot3<-qplot(type3,geom="histogram",binwidth = bw,main = "Prob␣Based␣Index␣v1",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

plot4<-qplot(type4,geom="histogram",binwidth = bw,main = "Prob␣Based␣Index␣v2",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

plot5<-qplot(type5,geom="histogram",binwidth = bw,main = "Heckman's␣v1",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

plot6<-qplot(type6,geom="histogram",binwidth = bw,main = "Heckman's␣v2",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi
```

```
plot7<-qplot(type7,geom="histogram",binwidth = bw,main = "EM␣Algorithm",
xlab = "Progesterone␣Diff",
fill=I("blue"),col=I("blue"),xlim=xlimits,ylim=ylimits) + geom_vline(xintercept = xi

grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, plot7, ncol=4)
}

graphsim1<-mygraphs(sim.1[[3]],sim.1[[4]],sim.1[[5]],sim.1[[6]],sim.1[[7]],sim.1[[8]]
ylim=c(0,105),as.numeric(sim.1[[11]][1,5]),
as.numeric(sim.1[[11]][1,7]),as.numeric(sim.1[[11]][1,8]),0.25)

graphsim2<-mygraphs(sim.2[[3]],sim.2[[4]],sim.2[[5]],sim.2[[6]],sim.2[[7]],sim.2[[8]]
ylim=c(0,25),as.numeric(sim.2[[11]][1,5]),
as.numeric(sim.2[[11]][1,7]),as.numeric(sim.2[[11]][1,8]),0.25)

graphsim3<-mygraphs(sim.3[[3]],sim.3[[4]],sim.3[[5]],sim.3[[6]],sim.3[[7]],sim.3[[8]]
ylim=c(0,105),as.numeric(sim.3[[11]][1,5]),
as.numeric(sim.3[[11]][1,7]),as.numeric(sim.3[[11]][1,8]),0.25)

graphsim4<-mygraphs(sim.4[[3]],sim.4[[4]],sim.4[[5]],sim.4[[6]],sim.4[[7]],sim.4[[8]]
ylim=c(0,25),as.numeric(sim.4[[11]][1,5]),
as.numeric(sim.4[[11]][1,7]),as.numeric(sim.4[[11]][1,8]),0.25)

graphsim5<-mygraphs(sim.5[[3]],sim.5[[4]],sim.5[[5]],sim.5[[6]],sim.5[[7]],sim.5[[8]]
ylim=c(0,105),as.numeric(sim.5[[11]][1,5]),
as.numeric(sim.5[[11]][1,7]),as.numeric(sim.5[[11]][1,8]),0.25)

graphsim6<-mygraphs(sim.6[[3]],sim.6[[4]],sim.6[[5]],sim.6[[6]],sim.6[[7]],sim.6[[8]]
ylim=c(0,25),as.numeric(sim.6[[11]][1,5]),
as.numeric(sim.6[[11]][1,7]),as.numeric(sim.6[[11]][1,8]),0.25)

graphsim7<-mygraphs(sim.7[[3]],sim.7[[4]],sim.7[[5]],sim.7[[6]],sim.7[[7]],sim.7[[8]]
ylim=c(0,105),as.numeric(sim.7[[11]][1,5]),
as.numeric(sim.7[[11]][1,7]),as.numeric(sim.7[[11]][1,8]),0.25)

graphsim8<-mygraphs(sim.8[[3]],sim.8[[4]],sim.8[[5]],sim.8[[6]],sim.8[[7]],sim.8[[8]]
ylim=c(0,25),as.numeric(sim.8[[11]][1,5]),
as.numeric(sim.8[[11]][1,7]),as.numeric(sim.8[[11]][1,8]),0.25)

graphsim9<-mygraphs(sim.9[[3]],sim.9[[4]],sim.9[[5]],sim.9[[6]],sim.9[[7]],sim.9[[8]]
ylim=c(0,105),as.numeric(sim.9[[11]][1,5]),
as.numeric(sim.9[[11]][1,7]),as.numeric(sim.9[[11]][1,8]),0.25)

graphsim10<-mygraphs(sim.10[[3]],sim.10[[4]],sim.10[[5]],sim.10[[6]],sim.10[[7]],sim
ylim=c(0,25),as.numeric(sim.10[[11]][1,5]),
as.numeric(sim.10[[11]][1,7]),as.numeric(sim.10[[11]][1,8]),0.25)

graphsim11<-mygraphs(sim.11[[3]],sim.11[[4]],sim.11[[5]],sim.11[[6]],sim.11[[7]],sim
ylim=c(0,105),as.numeric(sim.11[[11]][1,5]),
as.numeric(sim.11[[11]][1,7]),as.numeric(sim.11[[11]][1,8]),0.25)

graphsim12<-mygraphs(sim.12[[3]],sim.12[[4]],sim.12[[5]],sim.12[[6]],sim.12[[7]],sim
ylim=c(0,25),as.numeric(sim.12[[11]][1,5]),
as.numeric(sim.12[[11]][1,7]),as.numeric(sim.12[[11]][1,8]),0.25)
```

```
graphsim13<-mygraphs(sim.13[[3]],sim.13[[4]],sim.13[[5]],sim.13[[6]],sim.13[[7]],sim
ylim=c(0,105),as.numeric(sim.13[[11]][1,5]),
as.numeric(sim.13[[11]][1,7]),as.numeric(sim.13[[11]][1,8]),0.25)

graphsim14<-mygraphs(sim.14[[3]],sim.14[[4]],sim.14[[5]],sim.14[[6]],sim.14[[7]],sim
ylim=c(0,25),as.numeric(sim.14[[11]][1,5]),
as.numeric(sim.14[[11]][1,7]),as.numeric(sim.14[[11]][1,8]),0.25)

graphsim15<-mygraphs(sim.15[[3]],sim.15[[4]],sim.15[[5]],sim.15[[6]],sim.15[[7]],sim
ylim=c(0,105),as.numeric(sim.15[[11]][1,5]),
as.numeric(sim.15[[11]][1,7]),as.numeric(sim.15[[11]][1,8]),0.25)

graphsim16<-mygraphs(sim.16[[3]],sim.16[[4]],sim.16[[5]],sim.16[[6]],sim.16[[7]],sim
ylim=c(0,25),as.numeric(sim.16[[11]][1,5]),
as.numeric(sim.16[[11]][1,7]),as.numeric(sim.16[[11]][1,8]),0.25)

graphsim17<-mygraphs(sim.17[[3]],sim.17[[4]],sim.17[[5]],sim.17[[6]],sim.17[[7]],sim
ylim=c(0,105),as.numeric(sim.17[[11]][1,5]),
as.numeric(sim.17[[11]][1,7]),as.numeric(sim.17[[11]][1,8]),0.25)

graphsim18<-mygraphs(sim.18[[3]],sim.18[[4]],sim.18[[5]],sim.18[[6]],sim.18[[7]],sim
ylim=c(0,25),as.numeric(sim.18[[11]][1,5]),
as.numeric(sim.18[[11]][1,7]),as.numeric(sim.18[[11]][1,8]),0.25)

graphsim19<-mygraphs(sim.19[[3]],sim.19[[4]],sim.19[[5]],sim.19[[6]],sim.19[[7]],sim
ylim=c(0,105),as.numeric(sim.19[[11]][1,5]),
as.numeric(sim.19[[11]][1,7]),as.numeric(sim.19[[11]][1,8]),0.25)

graphsim20<-mygraphs(sim.20[[3]],sim.20[[4]],sim.20[[5]],sim.20[[6]],sim.20[[7]],sim
ylim=c(0,25),as.numeric(sim.20[[11]][1,5]),
as.numeric(sim.20[[11]][1,7]),as.numeric(sim.20[[11]][1,8]),0.25)

graphsim21<-mygraphs(sim.21[[3]],sim.21[[4]],sim.21[[5]],sim.21[[6]],sim.21[[7]],sim
ylim=c(0,105),as.numeric(sim.21[[11]][1,5]),
as.numeric(sim.21[[11]][1,7]),as.numeric(sim.21[[11]][1,8]),0.25)

graphsim22<-mygraphs(sim.22[[3]],sim.22[[4]],sim.22[[5]],sim.22[[6]],sim.22[[7]],sim
ylim=c(0,25),as.numeric(sim.22[[11]][1,5]),
as.numeric(sim.22[[11]][1,7]),as.numeric(sim.22[[11]][1,8]),0.25)

graphsim23<-mygraphs(sim.23[[3]],sim.23[[4]],sim.23[[5]],sim.23[[6]],sim.23[[7]],sim
ylim=c(0,105),as.numeric(sim.23[[11]][1,5]),
as.numeric(sim.23[[11]][1,7]),as.numeric(sim.23[[11]][1,8]),0.25)

graphsim24<-mygraphs(sim.24[[3]],sim.24[[4]],sim.24[[5]],sim.24[[6]],sim.24[[7]],sim
ylim=c(0,25),as.numeric(sim.24[[11]][1,5]),
as.numeric(sim.24[[11]][1,7]),as.numeric(sim.24[[11]][1,8]),0.25)




library(sampleSelection)
library(ggplot2)
library(gridExtra)
library(xtable)
library(lsr)
library(functional)
library(plyr)
bmdata <- read.csv(file="/Users/Kristen/Desktop/dissertation/bigmamaEM.csv",na.string

#distribution of Big Mama data
false=bmdata$prog2[bmdata$smoke1==1 & bmdata$srsmoke==0]
mean(false,na.rm=TRUE)
var(false,na.rm=TRUE)
length(na.omit(false))

true=bmdata$prog2[bmdata$smoke1==0 & bmdata$srsmoke==0]
mean(true,na.rm=TRUE)
var(true,na.rm=TRUE)
length(na.omit(true))

smokes=bmdata$prog2[bmdata$smoke1==1 & bmdata$srsmoke==1]
mean(smokes,na.rm=TRUE)
var(smokes,na.rm=TRUE)
length(na.omit(smokes))

smokers=bmdata$prog2[bmdata$smoke1==1]
nonsmokers=bmdata$prog2[bmdata$smoke1==0]
mean(smokers,na.rm=TRUE)
var(smokers,na.rm=TRUE)
mean(nonsmokers,na.rm=TRUE)
var(nonsmokers,na.rm=TRUE)
mean(smokers,na.rm=TRUE)-mean(nonsmokers,na.rm=TRUE)

#probability-based screener index v1
n=sum(length(na.omit(false)),length(na.omit(true)),length(na.omit(smokes)))
pbsif=length(na.omit(false))/n
pbsit=length(c(na.omit(true),na.omit(smokes)))/n
pbsi1=c(rep(pbsif,length(na.omit(false))),rep(pbsit,length(na.omit(true))))
pbsi2=rep(pbsit,length(na.omit(smokes)))

#probability-based screener index v2
randnum=sample(1:n,n,replace=FALSE)
newpbsi=cbind(randnum,c(pbsi1,pbsi2))
newpbsi.sort=newpbsi[order(newpbsi[,1]),]

mr=na.omit(false)
tr=na.omit(true)
sm=na.omit(smokes)
prog=c(mr,tr,sm)
pbsi=c(pbsi1,pbsi2)
sr=c(rep(0,sum(length(na.omit(false)),length(na.omit(true)))),rep(1,length(na.omit(sm
smokescore2=bm$NIC+newpbsi.sort[,2]+rnorm(1,15,10)
sel<-predict(lm(bm$smokescore~bm$race+bm$w1edin,na.action=na.exclude))
sel2<-predict(lm(smokescore2~bm$race+bm$w1edin,na.action=na.exclude))
plot(sel2,smokescore2)
choose<-rep(NA,length(sel))
```

```
choose2<-rep(NA,length(sel2))
for (j in 1:length(sel))
{
choose[j]=ifelse(is.na(sel[j])|sel[j]<=12,0,1)
choose2[j]=ifelse(is.na(sel2[j])|sel2[j]<=7,0,1)
}
logprog=log(bmdata$prog2)

#data fix
bmdata=cbind(bmdata,newpbsi.sort[,2],smokescore2,sel,choose,sel2,choose2,logprog)
bm=as.data.frame(bmdata)


#analysis lab value
labval.bm<-round(coef(summary(glm(prog2~smoke1,data=bm)))[2,1],2)
selabval.bm<-round(coef(summary(glm(prog2~smoke1,data=bm)))[2,2],1)
plabval.bm<-round(coef(summary(glm(prog2~smoke1,data=bm)))[2,4],4)
cilabval.bm<-round(confint(glm(prog2~smoke1,data=bm),'smoke1'),1)

#analysis self-report
selfrep.bm<-round(coef(summary(glm(prog2~srsmoke,data=bm)))[2,1],2)
seselfrep.bm<-round(coef(summary(glm(prog2~srsmoke,data=bm)))[2,2],1)
pselfrep.bm<-round(coef(summary(glm(prog2~srsmoke,data=bm)))[2,4],4)
ciselfrep.bm<-round(confint(glm(prog2~srsmoke,data=bm),'srsmoke'),1)

#analysis mischievous responders removed
mrremdata.bm<-bm[bm$score==20,]
mrrem.bm<-round(coef(summary(glm(prog2~srsmoke,data=mrremdata.bm)))[2,1],2)
semrrem.bm<-round(coef(summary(glm(prog2~srsmoke,data=mrremdata.bm)))[2,2],1)
pmrrem.bm<-round(coef(summary(glm(prog2~srsmoke,data=mrremdata.bm)))[2,4],4)
cimrrem.bm<-round(confint(glm(prog2~srsmoke,data=mrremdata.bm),'srsmoke'),1)

#analysis probability based index v2
pbsc.bm<-round(coef(summary(glm(prog2~srsmoke+pbsi,data=bm)))[2,1],2)
sepbsc.bm<-round(coef(summary(glm(prog2~srsmoke+pbsi,data=bm)))[2,2],1)
ppbsc.bm<-round(coef(summary(glm(prog2~srsmoke+pbsi,data=bm)))[2,4],4)
cipbsc.bm<-round(confint(glm(prog2~srsmoke+pbsi,data=bm),'srsmoke'),1)

#analysis probability based index v1
pbic.bm<-round(coef(summary(glm(prog2~srsmoke+newpbsi.sort[,2],data=bm)))[2,1],2)
sepbic.bm<-round(coef(summary(glm(prog2~srsmoke+newpbsi.sort[,2],data=bm)))[2,2],1)
ppbic.bm<-round(coef(summary(glm(prog2~srsmoke+newpbsi.sort[,2],data=bm)))[2,4],4)
cipbic.bm<-round(confint(glm(prog2~srsmoke+newpbsi.sort[,2],data=bm),'srsmoke'),1)

#analysis heckman v2
eqn1.bm<-fitted(glm(choose~bmi,family=binomial(link='probit'),data=bm,na.action=na.e
bm$IMR<-dnorm(eqn1.bm)/pnorm(eqn1.bm)
eqn2.bm<-summary(lm(prog2~srsmoke+IMR,data=bm,subset=(choose==1)))
heck.bm<-round(coef(eqn2.bm)[2,1],2)
seheck.bm<-round(coef(eqn2.bm)[2,2],1)
pheck.bm<-round(coef(eqn2.bm)[2,4],4)
ciheck.bm<-round(confint(lm(prog2~srsmoke+IMR,data=bm,subset=(choose==1)),'srsmoke')

#analysis heckman v1
```

```
eqn3.bm<-fitted(glm(choose2~bmi,family=binomial(link='probit'),data=bm,na.action=na.
bm$IMR2<-dnorm(eqn3.bm)/pnorm(eqn3.bm)
eqn4.bm<-summary(lm(prog2~srsmoke+IMR2,data=bm,subset=(choose2==1)))
heck2.bm<-round(coef(eqn4.bm)[2,1],2)
seheck2.bm<-round(coef(eqn4.bm)[2,2],1)
pheck2.bm<-round(coef(eqn4.bm)[2,4],4)
ciheck2.bm<-round(confint(lm(prog2~srsmoke+IMR2,data=bm,subset=(choose2==1)),'srsmoke



#EM Algorith
yvecbm=list(c(mr,tr),sm)

plotiter2=function(y,probs,muvec,sigmasqvec,ezmat,pause=F)
{
xtemp=seq(min(y),max(y),(max(y)-min(y))/1000)
compvec1=probs*dnorm(xtemp,muvec[1],sqrt(sigmasqvec[1]))
compvec2=(1-probs)*dnorm(xtemp,muvec[2],sqrt(sigmasqvec[2]))
ytemp=compvec1+compvec2
#plot(xtemp,ytemp,type="l",col="blue")
plot(xtemp,compvec1,col="red",type="l",xlab="progesterone␣level",ylab="Non-Smoker␣De
lines(xtemp,compvec2,col="green")
legend("topright",legend=c("Mischievous␣Responders","True␣Non-Smokers"),lty=c(1,1),c
#lines(xtemp,compvec2,col="green")

colorvec=rgb(ezmat[,1],ezmat[,2],0)
points(y,abs(rnorm(length(y),0,max(ytemp)/30)),col=colorvec)
if (pause) {scan()}
}


emmix=function(yvec,yvec2,k,startprob,start1,start2,start3,start4,tol=1e-04,pause=F)
{
n=length(yvec)
m=length(yvec2)
#starting values
muvec=start1
sigmasqvec=start2
probs=startprob
muvec2=start3
sigmasqvec2=start4

#expected value of unobserved data
ezmatnorm=matrix(0,nrow=n,ncol=k)
ezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
ezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
ezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
rowsums=apply(ezmatnorm,1,sum)
ezmat=ezmatnorm/rowsums
llik=sum(log(rowsums))

ezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
rowsums2=apply(ezmatnorm2,1,sum)
```

```
ezmat2=ezmatnorm2/rowsums2
llik2=sum(log(rowsums2))

rowsumssmk=sum(c(ezmatnorm[,1],ezmatnorm2[,1]))
rowsumsnsmk=sum(ezmatnorm[,2])
ezmatsmk=c(ezmatnorm[,1],ezmatnorm2[,1])/rowsumssmk
ezmatnsmk=ezmatnorm[,2]/rowsumsnsmk


repeat
{
#ezmat contains the ez_ij values
#llik contains the loglikelihood
bothprobs=apply(ezmat,2,mean)
probs=bothprobs[which.min(bothprobs)]
if (probs<0.01){probs<-0.05}
#ifelse(probs>0.01,probs,0.05)
muvec[1]=sum(ezmat[,1]*yvec)/sum(ezmat[,1])
if (muvec[1]<=0){muvec[1]<-0.001}
muvec[2]=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvec[1]=sum((ezmat[,1]*(yvec-muvec[1])^2))/sum(ezmat[,1])
sigmasqvec[2]=sum((ezmat[,2]*(yvec-muvec[2])^2))/sum(ezmat[,2])

muvec2=sum(ezmat2[,1]*yvec2)/sum(ezmat2[,1])
sigmasqvec2=sum((ezmat2[,1]*(yvec2-muvec2)^2))/sum(ezmat2[,1])

muvecsmk=sum(c(ezmat[,1]*yvec,ezmat2[,1]*yvec2))/sum(c(ezmat[,1]),ezmat2[,1])
muvecnsmk=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvecsmk=sum(c(ezmat[,1]*((yvec-muvec[1])^2),ezmat2[,1]*((yvec2-muvec2)^2))/sum
sigmasqvecnsmk=sum(ezmat[,2]*((yvec-muvec[2])^2))/sum(ezmat[,2])

#compute loglikelihood for new iteration
newezmatnorm=matrix(0,nrow=n,ncol=k)
newezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
newezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
for (k in 1:length(newezmatnorm[,1]))
{
if (newezmatnorm[k,1]<=0.001){newezmatnorm[k,1]<-0.001}
}
newezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
check=newezmatnorm[apply(newezmatnorm,1,Compose(is.finite, all)),]
newrowsums=apply(check,1,sum)
newezmat=newezmatnorm/newrowsums
newllik=sum(log(newrowsums))

newezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
newrowsums2=apply(newezmatnorm2,1,sum)
newezmat2=newezmatnorm2/newrowsums2
newllik2=sum(log(newrowsums2))

newrowsumssmk=sum(c(newezmatnorm[,1],newezmatnorm2[,1]))
newrowsumsnsmk=sum(newezmatnorm[,2])
newezmatsmk=c(newezmatnorm[,1],newezmatnorm2[,1])/newrowsumssmk
```

```
newezmatnsmk=newezmatnorm[,2]/newrowsumsnsmk

smkdiff=muvecsmk-muvecnsmk
w1=sigmasqvecsmk/length(c(ezmat[,1],ezmat2[,1]))
w2=sigmasqvecnsmk/length(ezmat[,2])
smkdiffvar=w1+w2
denom=((w1^2)/(length(c(ezmat[,1],ezmat2[,1]))-1))+((w2^2)/(length(ezmat[,2])-1))

if ((newllik-llik)<tol) {break}
ezmat=newezmat
llik=newllik
ezmat2=newezmat2
llik2=newllik2
ezmatsmk=newezmatsmk
ezmatnsmk=ezmatnsmk
}
return(list(probs=c(probs,1-probs),muvec=muvec,sigmasqvec=sigmasqvec,muvec2=muvec2,si
muvecsmk=muvecsmk,muvecnsmk=muvecnsmk,sigmasqvecsmk=sigmasqvecsmk,sigmasqvecnsmk=sigm
smkdiff=smkdiff,smkdiffvar=smkdiffvar,w1=w1,w2=w2,denom=denom))
}

embm=emmix(yvecbm[[1]],yvecbm[[2]],2,0.0001,c(50,60),c(350,370),50,70,1e-08,pause=F)
#embm=emmix(yvecbm[[1]],yvecbm[[2]],2,0.15,c(50,60),c(350,360),50,70,plotiter2,1e-08
embm

plotiter=function(y,y2,probs,muvec,sigmasqvec,muvectr,sigmasqvectr,pause=F)
{
xtemp=seq(min(y),max(y),(max(y)-min(y))/1000)
compvec1=probs*dnorm(xtemp,muvec[1],sqrt(sigmasqvec[1]))
compvec2=(1-probs)*dnorm(xtemp,muvec[2],sqrt(sigmasqvec[2]))
xtemp2=seq(min(y2),max(y2),(max(y2)-min(y2))/1000)
compvec3=dnorm(xtemp2,muvectr,sqrt(sigmasqvectr))
return(list(xtemp=xtemp,compvec1=compvec1,compvec2=compvec2,xtemp2=xtemp2,compvec3=co
}

bmdata=plotiter(yvecbm[[1]],yvecbm[[2]],0.1,c(49.97,59.27),c(358.66,377.77),47.44,72
#bmdata=plotiter(yvecbm[[1]],yvecbm[[2]],0.15,c(49.97,60.28),c(358.66,361.60),47.44,
plot(bmdata$xtemp,bmdata$compvec1,col="red",type="l",lwd=2,xlab="progesterone␣level"
lines(bmdata$xtemp,bmdata$compvec2,col="green",lty=2,lwd=2)
lines(bmdata$xtemp2,bmdata$compvec3,col="blue",lty=4,lwd=2)
legend("topright",legend=c("Inaccurate␣Nonsmokers","True␣Nonsmokers","True␣Smokers")
lty=c(1,2,4),,lwd=c(2,2,2),col=c("red","green","blue"))

emest=plotiter(yvecbm[[1]],yvecbm[[2]],embm$probs[1],embm$muvec,embm$sigmasqvec,embm
plot(emest$xtemp,emest$compvec1,col="red",type="l",lwd=2,xlab="progesterone␣level",y
lines(emest$xtemp,emest$compvec2,col="green",lty=2,lwd=2)
lines(emest$xtemp2,emest$compvec3,col="blue",lty=4,lwd=2)
legend("topright",legend=c("Inaccurate␣Nonsmokers","True␣Nonsmokers","True␣Smokers")
lty=c(1,2,4),,lwd=c(2,2,2),col=c("red","green","blue"))

semmix=function(yvec,yvec2,k,startprob,start1,start2,start3,start4,emmuvecsmk,emmuvec
{
n=length(yvec)
m=length(yvec2)
```

```
#starting values
muvec=start1
sigmasqvec=start2
probs=startprob
muvec2=start3
sigmasqvec2=start4

#expected value of unobserved data
ezmatnorm=matrix(0,nrow=n,ncol=k)
ezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
ezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
ezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
rowsums=apply(ezmatnorm,1,sum)
ezmat=ezmatnorm/rowsums
llik=sum(log(rowsums))

ezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
rowsums2=apply(ezmatnorm2,1,sum)
ezmat2=ezmatnorm2/rowsums2
llik2=sum(log(rowsums2))

rowsumssmk=sum(c(ezmatnorm[,1],ezmatnorm2[,1]))
rowsumsnsmk=sum(ezmatnorm[,2])
ezmatsmk=c(ezmatnorm[,1],ezmatnorm2[,1])/rowsumssmk
ezmatnsmk=ezmatnorm[,2]/rowsumsnsmk


repeat
{
#ezmat contains the ez_ij values
#llik contains the loglikelihood
bothprobs=apply(ezmat,2,mean)
probs=bothprobs[which.min(bothprobs)]
if (probs<0.01){probs<-0.05}
muvec[1]=sum(ezmat[,1]*yvec)/sum(ezmat[,1])
if (muvec[1]<=0){muvec[1]<-0.001}
muvec[2]=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvec[1]=sum((ezmat[,1]*(yvec-muvec[1])^2))/sum(ezmat[,1])
sigmasqvec[2]=sum((ezmat[,2]*(yvec-muvec[2])^2))/sum(ezmat[,2])

muvec2=sum(ezmat2[,1]*yvec2)/sum(ezmat2[,1])
sigmasqvec2=sum((ezmat2[,1]*(yvec2-muvec2)^2))/sum(ezmat2[,1])

muvecsmk=sum(c(ezmat[,1]*yvec,ezmat2[,1]*yvec2))/sum(c(ezmat[,1]),ezmat2[,1])
muvecnsmk=sum(ezmat[,2]*yvec)/sum(ezmat[,2])
sigmasqvecsmk=sum(c(ezmat[,1]*((yvec-muvec[1])^2),ezmat2[,1]*((yvec2-muvec2)^2))/sum
sigmasqvecnsmk=sum(ezmat[,2]*((yvec-muvec[2])^2))/sum(ezmat[,2])

#compute loglikelihood for new iteration
newezmatnorm=matrix(0,nrow=n,ncol=k)
newezmatnorm2=matrix(0,nrow=m,ncol=1)
#fill in ezmatnorm with p_j N(mu_j,sigmasqvec_j)
newezmatnorm[,1]=probs*dnorm(yvec,muvec[1],sqrt(sigmasqvec[1]))
```

```
for (k in 1:length(newezmatnorm[,1]))
{
if (newezmatnorm[k,1]<=0.001){newezmatnorm[k,1]<-0.001}
}
newezmatnorm[,2]=(1-probs)*dnorm(yvec,muvec[2],sqrt(sigmasqvec[2]))
check=newezmatnorm[apply(newezmatnorm,1,Compose(is.finite, all)),]
newrowsums=apply(check,1,sum)
newezmat=newezmatnorm/newrowsums
newllik=sum(log(newrowsums))

newezmatnorm2[,1]=dnorm(yvec2,muvec2,sqrt(sigmasqvec2))
newrowsums2=apply(newezmatnorm2,1,sum)
newezmat2=newezmatnorm2/newrowsums2
newllik2=sum(log(newrowsums2))

newrowsumssmk=sum(c(newezmatnorm[,1],newezmatnorm2[,1]))
newrowsumsnsmk=sum(newezmatnorm[,2])
newezmatsmk=c(newezmatnorm[,1],newezmatnorm2[,1])/newrowsumssmk
newezmatnsmk=newezmatnorm[,2]/newrowsumsnsmk


newmuvecsmk=sum(c(newezmat[,1]*yvec,newezmat2[,1]*yvec2))/sum(c(newezmat[,1]),newezm
newmuvecnsmk=sum(newezmat[,2]*yvec)/sum(newezmat[,2])
newsigmasqvecsmk=sum(c(newezmat[,1]*((yvec-muvec[1])^2),newezmat2[,1]*((yvec2-muvec2
newsigmasqvecnsmk=sum(newezmat[,2]*((yvec-muvec[2])^2))/sum(newezmat[,2])

newbothprobs=apply(newezmat,2,mean)
newprobs=newbothprobs[which.min(newbothprobs)]

r11=(newmuvecsmk-emmuvecsmk)/(muvecsmk-emmuvecsmk)
r12=(newsigmasqvecsmk-emsigmasqvecsmk)/(muvecsmk-emmuvecsmk)
r13=(newmuvecnsmk-emmuvecnsmk)/(muvecsmk-emmuvecsmk)
r14=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(muvecsmk-emmuvecsmk)
r15=(newprobs[1]-startprob)/(muvecsmk-emmuvecsmk)
r21=(newmuvecsmk-emmuvecsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r22=(newsigmasqvecsmk-emsigmasqvecsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r23=(newmuvecnsmk-emmuvecnsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r24=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(sigmasqvecsmk-emsigmasqvecsmk)
r25=(newprobs[1]-startprob)/(sigmasqvecsmk-emsigmasqvecsmk)
r31=(newmuvecsmk-emmuvecsmk)/(muvecnsmk-emmuvecnsmk)
r32=(newsigmasqvecsmk-emsigmasqvecsmk)/(muvecnsmk-emmuvecnsmk)
r33=(newmuvecnsmk-emmuvecnsmk)/(muvecnsmk-emmuvecnsmk)
r34=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(muvecnsmk-emmuvecnsmk)
r35=(newprobs[1]-startprob)/(muvecnsmk-emmuvecnsmk)
r41=(newmuvecsmk-emmuvecsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r42=(newsigmasqvecsmk-emsigmasqvecsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r43=(newmuvecnsmk-emmuvecnsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r44=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r45=(newprobs[1]-startprob)/(sigmasqvecnsmk-emsigmasqvecnsmk)
r51=(newmuvecsmk-emmuvecsmk)/(probs[1]-startprob)
r52=(newsigmasqvecsmk-emsigmasqvecsmk)/(probs[1]-startprob)
r53=(newmuvecnsmk-emmuvecnsmk)/(probs[1]-startprob)
r54=(newsigmasqvecnsmk-emsigmasqvecnsmk)/(probs[1]-startprob)
r55=(newprobs[1]-startprob)/(probs[1]-startprob)
```

```
if ((newllik-llik)<tol) {break}
ezmat=newezmat
llik=newllik
ezmat2=newezmat2
llik2=newllik2
ezmatsmk=newezmatsmk
ezmatnsmk=ezmatnsmk
rmat=matrix(c(r11,r12,r13,r14,r15,r21,r22,r23,r24,r25,r31,r32,r33,r34,r35,r41,r42,r43
}
return(rmat)
}

sembm=semmix(yvecbm[[1]],yvecbm[[2]],2,embm$probs[1],c(50,60),c(350,370),50,70,embm$
sembm

cdvcbm=matrix(c((1/embm$sigmasqvecsmk),0,0,0,0,0,(1/(2*embm$sigmasqvecsmk^2)),0,0,0,
((1/embm$probs[1])+(1/(1-embm$probs[1])))),5,5)

Vbm=solve(cdvcbm) + solve(cdvcbm)%*%sembm%*%solve(diag(5)-sembm)
Vbm
diag(Vbm)
sqrt(diag(Vbm))

tbm=embm$smkdiff/(sqrt((diag(Vbm)[1]/9)+(diag(Vbm)[3]/99)))
w1bm=Vbm[1]/9
w2bm=Vbm[2]/99
nubm=((w1bm+w2bm)^2)/(((w1bm^2)/9)+((w2bm^2)/99))

sebm=round(sqrt((diag(Vbm)[1]/9)+(diag(Vbm)[3]/99)),1)
sepibm=round(sqrt((diag(Vbm)[5])/108),2)

cilow.embm=embm$smkdiff-qt(.975,107)*sqrt((diag(Vbm)[1]/9)+(diag(Vbm)[3]/99))
cihigh.embm=embm$smkdiff+qt(.975,107)*sqrt((diag(Vbm)[1]/9)+(diag(Vbm)[3]/99))
ci.embm=round(c(cilow.embm,cihigh.embm),1)

pembm=round(2*pt(-abs(tbm),df=1),4)

cilow.pi1embm=embm$probs[1]-qnorm(.975)*sqrt((diag(Vbm)[5])/108)
cihigh.pi1embm=embm$probs[1]+qnorm(.975)*sqrt((diag(Vbm)[5])/108)
ci.pi1embm=round(c(cilow.pi1embm,cihigh.pi1embm),2)

cilow.pi2embm=embm$probs[2]-qnorm(.975)*sqrt((diag(Vbm)[5])/108)
cihigh.pi2embm=embm$probs[2]+qnorm(.975)*sqrt((diag(Vbm)[5])/108)
ci.pi2embm=round(c(cilow.pi2embm,cihigh.pi2embm),2)


##############################################################################################
#creating tables
estimates.bm<-c(labval.bm,selfrep.bm,mrrem.bm,pbic.bm,pbsc.bm,heck2.bm,heck.bm,round
seestimates.bm<-c(selabval.bm,seselfrep.bm,semrrem.bm,sepbic.bm,sepbsc.bm,seheck2.bm
ci.bm<-rbind(cilabval.bm,ciselfrep.bm,cimrrem.bm,cipbic.bm,cipbsc.bm,ciheck2.bm,cihec
sig.bm<-c(plabval.bm,pselfrep.bm,pmrrem.bm,ppbic.bm,ppbsc.bm,pheck2.bm,pheck.bm,pembm
method.bm<-c('Lab Values','Self-Report','Misch Resp Removed','Prob Based Index v1','
simtab<-cbind(method.bm, estimates.bm, seestimates.bm, ci.bm, sig.bm)

simtab
write.csv(simtab,"/Users/Kristen/Desktop/dissertation/simulationTableApplied.csv")

sims.bm<-cbind(simtab[,1],format(round(as.numeric(simtab[,2]),digits=2),nsmall=2),
paste("(",format(round(as.numeric(simtab[,4]),digits=2),nsmall=2),",",format(round(as
simtab[,6])
colnames(sims.bm)<-c("Method","Estimate","Standard Error","95% CI","p-value")
write.csv(sims.bm,"/Users/Kristen/Desktop/dissertation/simulationTableApplied2.csv")
print(xtable(sims.bm),include.rownames=FALSE)



#########################log prog##########################################################
#analysis lab value
labval.bmlog<-round(coef(summary(glm(logprog~smoke1,data=bm)))[2,1],2)
selabval.bmlog<-round(coef(summary(glm(logprog~smoke1,data=bm)))[2,2],1)
plabval.bmlog<-round(coef(summary(glm(logprog~smoke1,data=bm)))[2,4],4)
cilabval.bmlog<-round(confint(glm(logprog~smoke1,data=bm),'smoke1'),1)

#analysis self-report
selfrep.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=bm)))[2,1],2)
seselfrep.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=bm)))[2,2],1)
pselfrep.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=bm)))[2,4],4)
ciselfrep.bmlog<-round(confint(glm(logprog~srsmoke,data=bm),'srsmoke'),1)

#analysis mischievous responders removed
mrremdata.bmlog<-bm[bm$score==20,]
mrrem.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=mrremdata.bmlog)))[2,1],2)
semrrem.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=mrremdata.bmlog)))[2,2],1)
pmrrem.bmlog<-round(coef(summary(glm(logprog~srsmoke,data=mrremdata.bmlog)))[2,4],4)
cimrrem.bmlog<-round(confint(glm(logprog~srsmoke,data=mrremdata.bmlog),'srsmoke'),1)

#analysis probability based index v2
pbsc.bmlog<-round(coef(summary(glm(logprog~srsmoke+pbsi,data=bm)))[2,1],2)
sepbsc.bmlog<-round(coef(summary(glm(logprog~srsmoke+pbsi,data=bm)))[2,2],1)
ppbsc.bmlog<-round(coef(summary(glm(logprog~srsmoke+pbsi,data=bm)))[2,4],4)
cipbsc.bmlog<-round(confint(glm(logprog~srsmoke+pbsi,data=bm),'srsmoke'),1)

#analysis probability based index v1
pbic.bmlog<-round(coef(summary(glm(logprog~srsmoke+newpbsi.sort[,2],data=bm)))[2,1],2
sepbic.bmlog<-round(coef(summary(glm(logprog~srsmoke+newpbsi.sort[,2],data=bm)))[2,2]
ppbic.bmlog<-round(coef(summary(glm(logprog~srsmoke+newpbsi.sort[,2],data=bm)))[2,4]
cipbic.bmlog<-round(confint(glm(logprog~srsmoke+newpbsi.sort[,2],data=bm),'srsmoke')

#analysis heckman v2
eqn1.bmlog<-fitted(glm(choose~bmi,family=binomial(link='probit'),data=bm,na.action=na
bm$IMR<-dnorm(eqn1.bmlog)/pnorm(eqn1.bmlog)
eqn2.bmlog<-summary(lm(logprog~srsmoke+IMR,data=bm,subset=(choose==1)))
heck.bmlog<-round(coef(eqn2.bmlog)[2,1],2)
seheck.bmlog<-round(coef(eqn2.bmlog)[2,2],1)
pheck.bmlog<-round(coef(eqn2.bmlog)[2,4],4)
```

```
ciheck.bmlog<-round(confint(lm(logprog~srsmoke+IMR,data=bm,subset=(choose==1)),'srsmo

#analysis heckman v1
eqn3.bmlog<-fitted(glm(choose2~bmi,family=binomial(link='probit'),data=bm,na.action=n
bm$IMR2<-dnorm(eqn3.bmlog)/pnorm(eqn3.bmlog)
eqn4.bmlog<-summary(lm(logprog~srsmoke+IMR2,data=bm,subset=(choose2==1)))
heck2.bmlog<-round(coef(eqn4.bmlog)[2,1],2)
seheck2.bmlog<-round(coef(eqn4.bmlog)[2,2],1)
pheck2.bmlog<-round(coef(eqn4.bmlog)[2,4],4)
ciheck2.bmlog<-round(confint(lm(logprog~srsmoke+IMR2,data=bm,subset=(choose2==1)),'sr

#distribution of Big Mama data
false.log=bm$logprog[bm$smoke1==1 & bm$srsmoke==0]
mean(false.log,na.rm=TRUE)
var(false.log,na.rm=TRUE)
length(na.omit(false.log))

true.log=bm$logprog[bm$smoke1==0 & bm$srsmoke==0]
mean(true.log,na.rm=TRUE)
var(true.log,na.rm=TRUE)
length(na.omit(true.log))

smokes.log=bm$logprog[bm$smoke1==1 & bm$srsmoke==1]
mean(smokes.log,na.rm=TRUE)
var(smokes.log,na.rm=TRUE)
length(na.omit(smokes.log))

smokers.log=bm$logprog[bm$smoke1==1]
nonsmokers.log=bm$logprog[bm$smoke1==0]
mean(smokers.log,na.rm=TRUE)
var(smokers.log,na.rm=TRUE)
mean(nonsmokers.log,na.rm=TRUE)
var(nonsmokers.log,na.rm=TRUE)
mean(smokers.log,na.rm=TRUE)-mean(nonsmokers.log,na.rm=TRUE)

#em
mr.log=na.omit(false.log)
tr.log=na.omit(true.log)
sm.log=na.omit(smokes.log)
logprog=c(mr.log,tr.log,sm.log)
yvecbm.log=list(c(mr.log,tr.log),sm.log)
embm.log=emmix(yvecbm.log[[1]],yvecbm.log[[2]],2,0.3,c(3.8,4.0),c(0.14,0.10),3.8,0.03
sembm.log=semmix(yvecbm[[1]],yvecbm[[2]],2,embm.log$probs[1],c(50,60),c(350,370),50,7
sembm.log

cdvcbm.log=matrix(c((1/embm.log$sigmasqvecsmk),0,0,0,0,0,(1/(2*embm.log$sigmasqvecsmk
((1/embm.log$probs[1])+(1/embm.log$probs[1]))),5,5)

Vbm.log=solve(cdvcbm.log) + solve(cdvcbm.log)%*%sembm.log%*%solve(diag(5)-sembm.log)
Vbm.log
diag(Vbm.log)
sqrt(diag(Vbm.log))

tbm.log=embm.log$smkdiff/(sqrt((diag(Vbm.log)[1]/9)+(diag(Vbm.log)[3]/99)))
```

```
w1bm.log=Vbm.log[1]/9
w2bm.log=Vbm.log[2]/99
nubm=((w1bm.log+w2bm.log)^2)/(((w1bm.log^2)/9)+((w2bm.log^2)/99))


cilow.embm.log=embm.log$smkdiff-qt(.975,107)*sqrt((diag(Vbm.log)[1]/9)+(diag(Vbm.log
cihigh.embm.log=embm.log$smkdiff+qt(.975,107)*sqrt((diag(Vbm.log)[1]/9)+(diag(Vbm.log
ci.embm.log=round(c(cilow.embm.log,cihigh.embm.log),1)

pembm.log=round(2*pt(-abs(tbm.log),df=1),4)

cilow.pi1embm.log=embm.log$probs[1]-qnorm(.975)*sqrt((diag(Vbm.log)[5])/108)
cihigh.pi1embm.log=embm.log$probs[1]+qnorm(.975)*sqrt((diag(Vbm.log)[5])/108)
ci.pi1embm.log=round(c(cilow.pi1embm.log,cihigh.pi1embm.log),2)

cilow.pi2embm.log=embm.log$probs[2]-qnorm(.975)*sqrt((diag(Vbm.log)[5])/108)
cihigh.pi2embm.log=embm.log$probs[2]+qnorm(.975)*sqrt((diag(Vbm.log)[5])/108)
ci.pi2embm.log=round(c(cilow.pi2embm.log,cihigh.pi2embm.log),2)

#creating tables
estimates.bmlog<-c(labval.bmlog,selfrep.bmlog,mrrem.bmlog,pbic.bmlog,pbsc.bmlog,heck2
seestimates.bmlog<-c(selabval.bmlog,seselfrep.bmlog,semrrem.bmlog,sepbic.bmlog,sepbsc
ci.bmlog<-rbind(cilabval.bmlog,ciselfrep.bmlog,cimrrem.bmlog,cipbic.bmlog,cipbsc.bmlo
sig.bmlog<-c(plabval.bmlog,pselfrep.bmlog,pmrrem.bmlog,ppbic.bmlog,ppbsc.bmlog,pheck2
method.bmlog<-c('Lab Values','Self-Report','Misch Resp Removed','Prob Based Index v1
simtab.log<-cbind(method.bmlog,estimates.bmlog,seestimates.bmlog,ci.bmlog,sig.bmlog)

simtab.log
write.csv(simtab.log,"/Users/Kristen/Desktop/dissertation/simulationTableAppliedLog.c

sims.bmlog<-cbind(simtab.log[,1],format(round(as.numeric(simtab.log[,2]),digits=2),n
paste("(",format(round(as.numeric(simtab.log[,4]),digits=2),nsmall=2),",",format(roun
simtab.log[,6])
colnames(sims.bmlog)<-c("Method","Estimate","95% CI","p-value")
write.csv(sims.bmlog,"/Users/Kristen/Desktop/dissertation/simulationTableApplied2Log
print(xtable(sims.bmlog),include.rownames=FALSE)


#boxplot
ggplot(bm, aes(x=group, y=prog2, fill=group)) + geom_boxplot() +
stat_summary(fun.y=mean, geom="point", shape=5, size=4)
ggplot(bm, aes(x=group, y=logprog, fill=group)) + geom_boxplot() +
stat_summary(fun.y=mean, geom="point", shape=5, size=4)

#density plot w/ mean
mdat <- ddply(bm, "group", summarise, prog2.mean=mean(prog2))
mdat
logmdat <- ddply(bm, "group", summarise, logprog.mean=mean(logprog))
logmdat
ggplot(bm, aes(x=prog2, colour=group)) +
geom_density()
ggplot(bm, aes(x=prog2, colour=group)) +
geom_density() +
geom_vline(data=mdat, aes(xintercept=prog2.mean,  colour=group),
linetype="dashed", size=1) +
ggtitle('Kernel␣Densities␣of␣Smoking␣Status␣Data') +
labs(x='Progesterone␣Level',y='Density')

ggplot(bm, aes(x=logprog, colour=group)) +
geom_density()
ggplot(bm, aes(x=logprog, colour=group)) +
geom_density() +
geom_vline(data=logmdat, aes(xintercept=logprog.mean,  colour=group),
linetype="dashed", size=1) +
ggtitle('Kernel␣Densities␣of␣Transformed␣Smoking␣Status␣Data') +
labs(x='Log␣Progesterone␣Level',y='Density')

emest.log=plotiter(yvecbm.log[[1]],yvecbm.log[[2]],embm.log$probs[1],embm.log$muvec,e
mean(emest.log$compvec1)
mean(emest.log$compvec2)
mean(emest.log$compvec3)
iir<-rep('Intentional␣Inaccurate␣Responder',length(emest.log$compvec1))
tns<-rep('True␣Nonsmoker',length(emest.log$compvec2))
ts<-rep('True␣Smoker',length(emest.log$compvec3))
grp<-c(iir,tns,ts)
dens<-c(emest.log$compvec1,emest.log$compvec2,emest.log$compvec3)
lnprog<-c(emest.log$xtemp,emest.log$xtemp,emest.log$xtemp)
dat<-data.frame(lnprog,dens,grp)


ggplot(dat, aes(x=lnprog, y=dens, colour=grp)) +
geom_line() +
geom_vline(xintercept=embm.log$muvec[1],  colour="red",
linetype="dashed", size=1) +
geom_vline(xintercept=embm.log$muvec[2],  colour="green",
linetype="dashed", size=1) +
geom_vline(xintercept=embm.log$muvec2,  colour="blue",
linetype="dashed", size=1) +
ggtitle('EM␣Estimated␣Distribution␣of␣Smoking␣Status␣Data') +
labs(x='Log␣Progesterone␣Level',y='Density') +
scale_colour_discrete(name ="group")
```

Bibliography

H Russell Bernard, Peter Killworth, David Kronenfeld, and Lee Sailer. The problem of informant accuracy: The validity of retrospective data. *Annual review of anthropology*, pages 495–517, 1984.

Kenneth RW Brewer and Robert W Mellor. The effect of sample structure on analytical surveys1, 2. *Australian Journal of Statistics*, 15(3):145–152, 1973.

Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder. The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292, 2006.

David Chan. So why ask me? are self-report data really that bad. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, pages 309–336, 2009.

Sven Cnattingius. The epidemiology of smoking during pregnancy: smoking prevalence, maternal characteristics, and pregnancy outcomes. *Nicotine & Tobacco Research*, 6(Suppl 2):S125–S140, 2004.

William G Cochran. *Sampling techniques.* John Wiley & Sons, 2007.

Dewey Cornell, Jennifer Klein, Tim Konold, and Francis Huang. Effects of validity screening items on adolescent survey data. *Psychological assessment*, 24(1):21, 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Xitao Fan, Brent C Miller, Kyung-Eun Park, Bryan W Winward, Mathew Christensen, Harold D Grotevant, and Robert H Tai. An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, 18(3):223–244, 2006.

RP Ford, DM Tappin, PJ Schluter, and CJ Wild. Smoking during pregnancy: how reliable are maternal self reports in new zealand? *Journal of epidemiology and community health*, 51(3): 246–251, 1997.

Sander Greenland, Kenneth J Rothman, and TL Lash. Modern epidemiology. 1998.

James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER, 1976.

Robert C Klesges, Margaret Debon, and Joanne White Ray. Are self-reports of smoking rate biased? evidence from the second national health and nutrition examination survey. *Journal of clinical epidemiology*, 48(10):1225–1233, 1995.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data.* John Wiley & Sons, 2014.

Judith Lumley, Catherine Chamberlain, Therese Dowswell, Sandy Oliver, Laura Oakley, and Lyndsey Watson. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database Syst Rev*, 3(3), 2009.

Xiao-Li Meng and Donald B Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

Joseph P Robinson-Cimpian. Inaccurate estimation of disparities due to mischievous responders several suggestions to assess conclusions. *Educational Researcher*, 43(4):171–185, 2014.

Bettylou Sherry, Maria Elena Jefferds, and Laurence M Grummer-Strawn. Accuracy of adolescent self-report of height and weight in assessing overweight status: a literature review. *Archives of pediatrics & adolescent medicine*, 161(12):1154–1161, 2007.

Kristin Siggeirsdottir, Thor Aspelund, Gunnar Sigurdsson, Brynjolfur Mogensen, Milan Chang, Birna Jonsdottir, Gudny Eiriksdottir, Lenore J Launer, Tamara B Harris, Brynjolfur Y Jonsson, et al. Inaccuracy in self-report of fractures may underestimate association with health outcomes when compared with medical record based fracture registry. *European journal of epidemiology*, 22 (9):631–639, 2007.

Pentti K Siiteri, Freddy Febres, LE Clemens, R Jeffrey Chang, Bernard Gondos, and Daniel Stites. Progesterone and maintenance of pregnancy: Is progesterone nature's immunosuppressant?*. *Annals of the New York Academy of Sciences*, 286(1):384–397, 1977.

Chris J Skinner, David Holt, and TM Fred Smith. *Analysis of complex surveys.* John Wiley & Sons, 1989.

C Vance and S Buchheim. On the application of heckman's sample selection model to travel survey data: some practical guidelines. *Proceedings of ETC 2005, Strasbourg, France 18-20 September 2005-Research to Inform Decision-making in Transport-applied Methods in Transport Planning-panel study*, 2005.

Appendix B

Vita

- **Education** PhD student in Statistics: University of Kentucky
  MS in Statistics: University of Kentucky, May 2012
  MA in Education: University of Kentucky, May 2005
  BA in Education: University of Kentucky, May 2004

- **Refereed Publications**

  1. Janes, J. G.; Garrett, K. S.; **McQuerry, K. J.**; Waddell, S.; Voor, M. J.; Reed, S. M.; Williams, N. M.; MacLeod, J. N. "Cervical Vertebral Lesions in Equine Stenotic Myelopathy." Veterinary pathology (2015):0300985815593127.
  2. Black, R.A.; Taraba, J.L.; Day, G.B.; Damasceno, F.A.; Wood, C.L.; **McQuerry, K.J.**; Bewley, J.M. "The relationship between compost bedded pack performance, management, and bacterial counts." Journal of Dairy Science 97(5), 2014, pp. 2669-2679.
  3. Smith, A.C.; Wood, C.L.; **McQuerry, K.J.**; Bewley, J.M. "Effect of a Tea Tree Oil and Organic Acid Footbath Solution on Digital Dermatitis in Dairy Cows." Journal of Diary Science 97(4), 2014, pp. 4041-4046.
  4. Janes, J.G.; Garrett, K.S.; **McQuerry, K.J.**; Pease, A.P.; Williams, N.M.; Reed, S.M.; and McLeod, J.N. "Comparison of MRI to standing cervical radiographs for evaluation of vertebral canal stenosis in equine Cervical Stenotic Myelopathy." Equine Veterinary Journal, 46(6), 2013, pp. 681-686.
  5. Liang, D.; Wood, C.L.; **McQuerry, K.J.**; Ray, D.L.; Clark, J.M.; Bewley, J.M. "Influence of breed, milk production, season, and ambient temperature on dairy cow reticulorumem temperature." Journal of Dairy Science 96(8), 2013, pp. 5072-5081.
  6. Sterrett, A.E.; Wood, C.L.; **McQuerry, K.J.**; Bewley, J.M. "Changes in teat-end hyperkeratosis after installation of an individual quarter pulsation milking system." Journal of Dairy Science 96(6), 2013, pp. 4041-4046.

- **Abstracts**

  1. Chavan,N.R.; Ashford K.; **McQuerry, K.J.**; McCubbin, A.; Barnett, J.; O'Brien, J. (2015). Smoking cessation in pregnancy: Impact of the bio-inflammatory milieu. 35th Annual Meeting Society of Maternal Fetal Medicine. San Diego, CA.
  2. Chavan,N.R.; Ashford K.; Meints, L.; **McQuerry, K.J.**; McCubbin, A.; Barnett, J.; O'Brien, J. (2015). Changes in inflammatory cytokines with sleep disturbances in pregnancy. 35th Annual Meeting Society of Maternal Fetal Medicine. San Diego, CA.
  3. Chavan,N.R.; Ashford K.; **McQuerry, K.J.**; McCubbin, A.; Barnett, J.; O'Brien, J. (2015). Weight gain in excess of IOM recommendations: The relationship to preterm birth and systemic inflammatory profile. 35th Annual Meeting Society of Maternal Fetal Medicine. San Diego, CA.
  4. Ashford, K.; O'Brien, J; **McQuerry, K.J.**; Barnett, J.; McCubbin, A.; Ferguson, J.; Ebersole, J. (2015). Cytokine concentrations and their associations with preterm birth: Comparison between serum and cervicovaginal measurements. 35th Annual Meeting Society of Maternal Fetal Medicine. San Diego, CA.
  5. Ashford, K.; Barnett, J.; McCubbin, A.; **McQuerry, K.J.**; Curry, T.; O'Brien, J. (2015). Tobacco use alters pregnancy biomarkers reflecting tissue function. 35th Annual Meeting Society of Maternal Fetal Medicine. San Diego, CA.

- **Presentations & Posters**

1. Jaromczyk, J.W., Moore, N., **McQuerry, K.J.** 2014. "Why What and How to publish in the Journal of the Kentucky Academy of Science." Centennial Meeting of the Kentucky Academy of Science. Lexington, KY.
2. **McQuerry, K.J.** 2013. "Using the Applied Statistics Laboratory and an Introduction to Experimental Design." University of Kentucky Plant and Soil Sciences Seminar Series. Lexington, KY.
3. Wadsworth, B.A.; Sterrett, A.E.; Wood, C.L.; **McQuerry, K.J.**; Clark, J.D.; Ray, D.L.; Bewley, J.M. 2013. "Characterization of lying time, milk yield, and rumination time with different freestall bases." Abstract 278. American Dairy Science Association Annual Meeting. Indianapolis, IN.
4. Sterrett, A.E., Wood, C.L., **McQuerry, K.J.**, Bewley, J.M.. 2012. "Potential utility of a parlor-based individual quarter milking system." Page 23 in Proceedings of the 38th of the International Committee for Animal Recording (ICAR) Session. Cork, Ireland.