November 2014

Estimating Prevalence from Complex Surveys

Sophie O'Brien
University of Massachusetts Amherst

Estimating Prevalence from Complex Surveys

A Thesis Presented

by

SOPHIE O'BRIEN

Submitted to the Graduate School of the University of Massachusetts Amherst in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

September 2014

Public Health

Biostatistics

Estimating Prevalence from Complex Surveys

A Thesis Presented

By

SOPHIE O'BRIEN

Approved as to style and content by:

_____

Edward J. Stanek III, Chair

_____

Rachel Volberg, Member

_____

Jing Qian, Member

_____

Edward J. Stanek III, Department Head

Department of Public Health

ACKNOWLEDGEMENTS

This project would not have been possible without the support of many people. Many thanks are due to my advisor, Ed J. Stanek III, for his patience and guidance throughout. I would also like to extend my gratitude to the members of my committee, Rachel Volberg and Jing Qian, for their helpful comments and suggestions. Thanks are also due to Martha Zorn for her support in all stages of this project and my professional development.

I want to thank the Massachusetts Gaming Commission for funding the SEIGMA project and allowing use of data in this manuscript.

A special thank you to all those whose support and friendship helped me to stay focused on this project and who have provided me with the encouragement to continue through hard times.

ABSTRACT

ESTIMATING PREVALANCE FROM COMPLEX SURVEYS

SEPTEMBER 2014

SOPHIE O'BRIEN, B.A., SAINT MICHAEL'S COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Edward J. Stanek III

Massachusetts passed legislation in the fall of 2012 to allow the construction of three casinos and a slot parlor in the state. The prevalence of problem gambling in the state and in areas where casinos will be constructed is of particular interest. The goal is to evaluate the change in prevalence after construction of the casinos, using a multi-mode address based sample survey. The objective of this thesis is to evaluate and describe ways of using statistical inference to estimates prevalence rates in finite populations. Four methods were considered in an attempt to evaluate the prevalence of problem gambling in the context of the gambling study. These methods were evaluated unconditionally and conditionally, controlling for gender, using mean square error (MSE) as a measure of accuracy. The simple mean, the post-stratified mean, the best linear unbiased predictor (BLUP), and the empirical best linear unbiased predictor (EBLUP) were considered in three examples: a simple population with N=5 taking samples of n=4; a population of N=20 taking samples of n=5; and a larger population of N=1,000 taking samples of n=200. Ten thousand

simulations were performed in SAS on the two larger populations and conclusions were made.

Conditional analyses of a population with N=1,000 and a crude problem gambling rate of 1.5, samples of n=200 led to the simple mean and the post-stratified mean to perform better in certain situations, as measured by their low MSE values. When there are less females than expected in a sample, the post-stratified mean produces a lower mean MSE over the 10,000 simulations. When there are more females than expected in a sample, the simple mean produces a lower mean MSE over the 10,000 simulations. Conditional analysis provided more appropriate results than unconditional analysis.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER I

INTRODUCTION

The University of Massachusetts Amherst School of Public Health & Health Sciences (SPHHS) has been engaged by the Massachusetts Gaming Commission (MGC) to carry out a comprehensive, multi-year research project, believed to be the first of its kind, on the economic and social impacts of introducing casino gambling in Massachusetts. The project fulfills Section 71 of the 2011 Expanded Gaming Act, which requires the MGC to establish "an annual research agenda" to assist in understanding the social and economic effects of the introduction of casino gambling in MA, and in making annual scientifically-based recommendations to the Legislature.

The study will focus particularly on problem gambling, but will also examine a wide array of social and economic effects of expanded gambling in Massachusetts. In addition to SPHHS, other key members of the research team include the UMass Donahue Institute, which will lead the economic and fiscal impact research, and NORC at the University of Chicago, which will lead primary data collection efforts for the SEIGMA study.

This project has been coined as a first-of-its-kind gambling monitoring system. This system will provide stakeholders in Massachusetts with an objective basis for strategic analysis and decision-making about the impacts of casinos at state, regional and local levels. The study will generate early warning signs of changes in social and economic impacts of new and existing forms of gambling in

Massachusetts, promote responsible gambling, and mitigate problem gambling through refinement of services. The agenda produces an analytic framework for socioeconomic impact studies and a multiple methods research strategy by employing primary and secondary data collection and analysis with quantitative and qualitative research methods.

There are three essential elements to the research agenda. One is the execution of a baseline prevalence study of problem gambling and determining existing prevention and treatment programs. Two is to document and monitor the social and economic effects of expanded gambling. Three is to facilitate independent studies in order to obtain scientific information relevant to enhancing responsible gambling and minimizing harmful effects.

CHAPTER II

SIGNIFICANCE

Massachusetts passed legislation in the fall of 2011 to allow the construction of three casinos and a slot parlor in the state. The prevalence of problem gambling in the state and in areas where casinos will be constructed is of particular interest. Evaluation of the change in prevalence after construction of the casinos is an important objective. Its significance is underscored by legislation that commits funds from casino operations to mitigate negative gambling impacts.

A multi-mode address-based sample (ABS) survey has been designed to collect 10,000 interviews from one adult per household based on a stratified sample of addresses from the US Postal system. The interviews accrue from responses to invitations by mail to complete an online survey, responses to a follow-up mailed copy of the survey, and responses to a telephone interview of a sub-sample of subsequent non-respondents. A survey of 5,000 members of an online panel in Massachusetts has also been conducted by an independent survey organization.

The challenge is to use primary data from these sources to estimate the prevalence of problem gambling in the state and in geographic regions adjacent to the casino sites once the licenses are awarded. Such challenges include defining a framework for estimation that accounts for the different survey modalities, defining populations for geographic target areas and linking survey data to these areas, accounting for survey weights and non-response, and accounting for important distributions of risk factors thought to be related to problem gambling prevalence.

This research focuses on the first challenge listed above. I will conduct and review competing approaches for estimation of the prevalence of problem gambling.

CHAPTER III

OBJECTIVE

The objective of this thesis is to evaluate and describe ways of using statistical inference to estimate prevalence rates in finite populations. We will consider four approaches controlling for gender, and methods will be analyzed using a simple example, with results compared via simulation using SAS 9.4.

We begin by discussing various techniques from the literature for estimation. Then we proceed to defining the population used in this paper. We describe the four methods used in detail, defining each aspect of each equation explicitly, where a simple example is included to portray ideas. Next we simulate data, first using the same population as the simple example, and then increasing the population to represent a more realistic situation. Evaluations are performed unconditionally and conditionally on gender. Limitations are noted and conclusions are made.

CHAPTER IV

LITERATURE REVIEW

There are many survey sampling estimation techniques that may be appropriate to use in the present context. One example is calibration. Calibration in survey sampling was formalized by Deville and Sarndal (1992) and Deville et al. (1993) to produce efficient estimators of totals for a set of variables of interest. The estimators are defined to minimize a distance with regard to initial sampling weights. Guggemos and Tille (2000) propose a new class of model-assisted estimators obtained by releasing a few calibration constraints and replacing them with penalty terms. They obtained a more flexible estimation procedure, a design-based alternative, by combining usual calibration and this 'relaxed' calibration.

Small area estimation has received considerable attention in recent years because of a growing demand for reliable small-area statistics. The direct-survey estimators, based only on the data from a given small area (or small domain), are likely to yield unacceptably large standard errors because of small sample size in the domain. Therefore, alternative estimators that borrow strength from other related small areas have been proposed in the literature to improve efficiency. These estimators use models, either implicitly or explicitly, that connect the small areas through supplementary (e.g. census and administrative) data. Prasad and Rao (1990) investigate three small-area models. These models are all special cases of a general mixed linear model involving fixed and random effects. A two-stage estimator of a small-area mean under each model is obtained, by first deriving the

best linear unbiased estimator (BLUP) assuming that the variance components that determine the variance-covariance matrix are known, and then replacing the variance components in the estimator with their estimators. Second-order approximation to the mean squared error of the two-stage estimator and the estimator of MSE approximation are obtained under normality. The MSE approximation provides a reliable measure of uncertainty associated with the two-stage estimator. It can also provide asymptotically valid confidence intervals on a small-area mean, as the number of small areas tends to increase to infinity.

Models for small area estimation based on a random effects specification typically assume population units in different areas are uncorrelated. However, they can be extended to account for the correlation between areas by assuming that area random effects are spatially correlated. Saei and Chambers (2003) suggest a simple variance-covariance structure for such a spatial correlation structure within the context of a linear model for the population characteristic of interest, and derive estimates of parameters and components of variance using maximum likelihood and restricted maximum likelihood methods. This allows empirical best linear unbiased predictions (EBLUP) for area totals to be computed for areas in sample as well as those not in sample. Their simulations indicate that their proposed method has the potential to lead to substantial increases in prediction efficiency for these areas where there is a strong spatial correlation in the data.

Sarndal (2011) extends the idea of balancing to the context of survey nonresponse. The set of respondents should be balanced vis-à-vis the whole probability sample. Here, balance is measured as 'lack of balance' with the opposite

sign, and is confined to the unit interval. At the data collection stage one may use aspects of responsive design to achieve good balance in an ultimate set of respondents. A pressing objective remains nevertheless for the estimation stage: to adjust for the bias that still affects the estimates. The size of the adjustment has remaining lack of balance as one of its factors; another is the strength of the relationship between the study variable y and the auxiliary vector x; the last is the degree to which large auxiliary variable mean difference between respondents and full sample is matched with large correlation between the auxiliary variable and the study variable. Further work would include more in-depth study of these factors and how they interact.

Weighting adjustments are commonly applied in surveys to compensate for nonresponse and bias, and to make weighted sample estimates conform to external values. Recent years have seen theoretical developments and increased use of methods that take account of substantial amounts of auxiliary information in making these adjustments. Kalton et al (2003) describe such methods as cell weighting, raking, generalized regression estimation, logistic regression weighting, mixtures of methods, and methods for restricting the range of the resultant adjustments, and how auxiliary variables may be chosen for use in the adjustments. By permitting more auxiliary information to be used for complex weighting adjustments, the variables have the potential to reduce biases arising from nonresponse and non-coverage. When a substantial amount of auxiliary information is available, a variety of alternative methods may be used. The choice of auxiliary

variables and of the mode in which they are employed in the adjustments may be of more significance than the choice of a particular method.

Theory and simulations show that, to reduce bias effectively without increasing variance, a covariate that is used for nonresponse weighting adjustment needs to be highly associated with both the response indicator and the survey outcome variable. In practice, these requirements pose a challenge that is often overlooked because those covariates are often not observed or may not exist.

Surveys have recently begun to collect supplementary data, such as interviewer observations and other proxy measures of key survey outcome variables. To the extent that these auxiliary variables are highly correlated with the actual outcomes, these variables are promising candidates for nonresponse adjustment. Kreuter et al (2010) examine traditional covariates and new auxiliary variables from multiple surveys and provide empirical estimates of the association between proxy measures and response to the survey request as well as the actual survey outcome variables. Weighted and unweighted estimates under various nonresponse models are also considered. Results indicate difficulties in finding suitable covariates for nonresponse adjustment, leading the authors to emphasize the need to improve the quality of auxiliary data.

CHAPTER V

DEFINING THE POPULATION

The aim of this thesis is to evaluate different statistical methods for estimating the prevalence of problem gambling in the context of the gambling study. We define these methods in the context of a simple finite population. The first step in this process is to define the setting.

We begin by defining parameters for a finite population that closely resemble the setting encountered in the SEIGMA baseline survey. The population in the context of the gambling study consists of adults aged 18 and over in Massachusetts. Sampling of the subjects was via two stage cluster sampling, with clusters corresponding to households, and one adult selected at random per household. We investigate methods for a simpler sample design based on simple random sampling.

In order to illustrate the notation, we include a simple example of a finite population. We define λ as the label for an adult in the population, where a subject, λ_s, refers to a specific person. This population is given by

$$\Omega = \{\lambda_s : s = 1, \ldots, N\}.$$

Associated with each adult is his or her gender, *h*, where *h=1* represents males and *h=2* represents females. We define random variables that indicate a subject's gender, and are defined by $\lambda_s \to x_h(\lambda_s)$ where $x_h(\lambda_s)$=0 if subject λ_s is not of gender *h*, and $x_h(\lambda_s)$=1 if subject λ_s is of gender *h.* Using these random variables, subjects in the population can be divided into $H = 2$ strata, or domains,

$$\Omega_h = \left\{ \lambda_s \mid x_h\left(\lambda_s\right) = 1 \right\}$$

where $\Omega = \bigcup_{h=1}^{H} \Omega_h$ and *H=2.*

Associated with a subject is the subject response which corresponds to the subjects gambling status - a problem gambler or not a problem gambler. A subject's problem gambling status is calculated based on their responses to a set of questions in the SEIGMA baseline survey. These questions target typical problem gambling behaviors. We assume that problem gambling status can be observed based on the subject's responses to these questions. We define $\lambda_s \rightarrow f(\lambda_s)=y_s$ to denote a subject's observed problem gambling outcome, i.e. their response, where, y_s=0 if subject λ_s is not a problem gambler, and y_s=1 if subject λ_s is a problem gambler.

Table 1 summarizes the concepts just described. The simple finite population contains five subjects- three males and two females. Subject λ, lettered A through E, refers to the specific subject referenced by λ_s, for *s=1,…N=5*, with corresponding gender *g*. Problem gambling status for a subject λ_s is also listed. There are three problem gamblers and two non-problem gamblers in this example.

Table 1: Simple Finite Population

s	**Subject (λ)**	**Gender (*G*)**	x_s	**Problem Gambling Status (y_s)**	y_s
1	A	M	0	y_1	0
2	B	M	0	y_2	1
3	C	M	0	y_3	1
4	D	F	1	y_4	1
5	E	F	1	y_5	0

Using this notation, population parameters are defined and illustrated using Table 1. The population mean is denoted as

$$\mu = \frac{1}{N}\sum_{\lambda_s \in \Omega} y_s ,$$

and the population variance is expressed by, $\sigma^2 = \frac{N-1}{N}S^2$, where

$$S^2 = \frac{\sum_{\lambda_s \in \Omega}^{N} (y_s - \mu)^2}{N-1} .$$

Using the population in Table 1, the population mean is $\mu = \frac{3}{5} = 0.6$, and

$S^2 = \frac{1.2}{4} = 0.3.$

Two strata, or domains, i.e., *H*=2, are considered in this paper: males and females. Gender specific population parameters can also be defined. To do so, existing notation must be refined. Let N_h represent the number of subjects in Ω_h. Using similar notation, the population mean can be denoted as

$$\mu_h = \frac{1}{N_h}\sum_{\lambda_s \in \Omega_h} y_s ,$$

and variance can be expressed by, $\sigma_h^2 = \frac{N_h - 1}{N_h}S_h^2$, where

$$S_h^2 = \frac{1}{N_h - 1}\sum_{\lambda_s \in \Omega_h} \left(y_s - \mu_h\right) .$$

For the data in Table 1, $\mu_1 = \frac{2}{3} = 0.6667$, $\mu_2 = \frac{1}{2} = 0.5$, $S_1^2 = \frac{0.6668}{2} = 0.3334$, and $S_2^2 = \frac{0.5}{1} = 0.5$. We also define the variance of the gambling rate between domains as,

$$V^2 = \frac{1}{H-1}\sum_{h=1}^{H}(\mu_h - \bar{\mu})^2,$$

where $\bar{\mu} = \frac{1}{H}\sum_{h=1}^{H}\mu_h$, and $h = 1,\ldots,H = 2$. For the example in Table 1, since $\mu_1 = \frac{2}{3}$ and $\mu_2 = \frac{1}{2}, \bar{\mu} = \frac{7}{12}$, and

$$V^2 = \left[\left(\frac{2}{3} - \frac{7}{12}\right) + \left(\frac{1}{2} - \frac{7}{12}\right)\right]^2$$
$$= 0.0138$$

CHAPTER VI

MODELING GAMBLING STATUS OVER TIME

In a broader context, a subject's problem gambling status is not a static status. At any point in time, a person could or could not be a problem gambler. Extraneous variables in a person's life, such as financial situations, may affect a subject's gambling behavior, and lead to the subject changing their problem gambling status. As a result, the model for problem gambling status may take into account time.

In order to account for time, we must redefine our problem gambling status variable. Let $\lambda_s \rightarrow f_t(\lambda_s) = y_{st}$ over times t=1,2,...,T where

y_{st} = 0 if subject λ_s is not a problem gambler at time *t, and*

y_{st} = 1 if subject λ_s is a problem gambler at time *t.*

Thus, y_{st} denotes response for subject λ_s at time *t,* the condition of measurement. Table 2 below illustrates this concept.

Table 2: Illustration of Problem Gambling Responses over Time

s	**G**	y_{st}	**t=1**	**t=2**	**t=3**	**t=4**	**t=5**	**t=6**	**t=7**	**t=8**	**t=9**	**t=10**	**Average**
1	M	y_{1t}	1	0	1	0	0	0	0	0	0	0	**0.2**
2	M	y_{2t}	0	0	0	1	1	0	0	0	0	0	**0.2**
3	M	y_{3t}	0	0	0	0	1	0	0	0	0	1	**0.2**
4	F	y_{4t}	0	0	0	0	1	0	0	0	0	0	**0.1**
5	F	y_{5t}	0	1	0	0	0	0	0	0	0	0	**0.1**
Total			**0.2**	**0.2**	**0.2**	**0.2**	**0.6**	**0**	**0**	**0**	**0**	**0.2**	**0.16**

Table 2 consists of a simple finite population of 5 subjects. In this population, 3 subjects are male, 2 subjects are female and we observe problem gambling status

for each subject every year for ten years (t represents one calendar year). For example, subject 1 was a problem gambler at time 1. This is denoted by

$$y_{st} = y_{11} = 1.$$

On the contrary, at t=2, this same subject did not identify himself as a problem gambler:

$$y_{st} = y_{12} = 0.$$

This illustrates the concept of variability in a subject's response over time and is an important distinction to make. As Table 2 depicts, problem gambling status varies over time and it also varies by person; therefore, a person's true problem gambling status is a dynamic situation that may change from one time point to another. In the context of this paper, we assume that unless there is a change in condition, the chance that a subject of a given gender is a problem gambler is equal at any time. With this assumption, the expected value for a subject being a problem gambler can be defined. We define this expected value as the latent values for the subject. Since problem gambling status is not observed at all times, this latent value can only be estimated.

For simplicity, this thesis focuses on a cross-sectional examination of problem gambling status rather than pursuing a longitudinal study approach. It is important to note the difference, however, and this example was used to portray ideas.

CHAPTER VII

METHODS

The next step in estimating the prevalence of problem gambling in the context of the gambling study is to define the estimation methods. We evaluate four estimators of prevalence: the sample mean, a post-stratified estimator, the best linear unbiased prediction (BLUP), and the empirical BLUP, or EBLUP. These four methods are used to estimate prevalence in a simple random sample of the simple finite population example, and compared against each other. In theory, the estimator with the lowest mean square error (MSE) provides the most accurate estimate of the prevalence of problem gambling. This will be evaluated both unconditionally and conditionally; the motivation here is to make use of axillary variables which in this case is gender.

a. Simple Example

A simple example is discussed to make the ideas clear. We define a sample of subjects as $\Omega^{(d)}$, a subset of subjects from the total population. When *n=4*, there are *d=1,…,D=5* different sets of sample subjects. This is shown by the equation below,

$$d = 1, \ldots, D = {}_N C_n = \binom{N}{n} = \binom{5}{4} = 5.$$

For example, five potential combinations of four elements out of the set of subjects in population $\Omega = \{A, B, C, D, E\}$, are given by $\Omega^D = \left\{\Omega^{(d)}; d = 1, \ldots D = 5\right\}$ where

$$\Omega^{(1)} = \{A, B, C, D\} \qquad \Omega^{(2)} = \{A, B, C, \mathrm{E}\} \qquad \Omega^{(3)} = \{A, B, \mathrm{D}, \mathrm{E}\}$$

$$\Omega^{(4)} = \{A, \mathrm{C}, \mathrm{D}, \mathrm{E}\} \qquad \Omega^{(5)} = \{B, C, D, E\}.$$

Each sample set contains either three males and one female, or two males and two females. Table 1 contains information on which subject corresponds to what gender. When all samples are equally likely, the sampling is simple random without replacement sampling.

A sample of four subjects will be used as an example to illustrate the four estimation methods that will be defined below. Mean bias and MSE values will be computed for all methods defined for $\Omega^{(d)}$.

b. The Simple Mean

Recall that y_s refers to problem gambling status for subject λ_s. Thus, the sample mean for $\Omega^{(d)}$ is given by,

$$\hat{p}_{1d} = \bar{y}_d = \frac{1}{n} \sum_{\lambda_s \in \Omega^{(d)}} y_s .$$

We represent this random variable for a member of Ω^D by $\hat{P}_1$. The variance of $\hat{P}_1$ is given by,

$$Var(\bar{P}_1) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right),$$

where S^2 is estimated by $s_d{}^2 = \frac{1}{n-1}\sum_{\lambda_s \in \Omega^{(d)}} (y_s - \bar{y}_d)^2$, or $\hat{S}^2$.

c. Post-Stratified Estimator

A post-stratified estimator is the second method for evaluating the prevalence of problem gambling. It is a simple weighted estimator. In the context of this study, a post-stratified estimator is a weighted average of the gender specific estimates of problem gambling. Weights correspond to the proportion of men and women in the population. Estimates of problem gambling for a sample correspond to the sample proportion for each gender group.

In order to define the post-stratified mean, first gender specific weights are defined. As stated above, weights correspond to gender in the *population.* Thus, the weight for males i.e., *h=1*, is defined by $w_1 = \frac{N_1}{N}$, and the weight for females i.e., *h=2*, is defined as $w_2 = \frac{N_2}{N}$. We assume these weights are known.

Next, we define the sample means for males and females. The equations are similar to the population mean for males and females, but altered to account for the *sample*. Let $n_{hd} = \sum_{\lambda_s \in \Omega^{(d)}} x_h(\lambda_s)$ denote the number of subjects in the sample

$\Omega^{(d)}$ in domain h . Using this notation, the sample mean for domain h is defined as

$$\overline{y}_{hd} = \frac{1}{n_h} \sum_{\lambda_s \in \Omega^{(d)}} x_h(\lambda_s) y_s \, .$$

We represent the random variables for the proportion of gamblers in domain h in a sample by $\overline{Y}_h$, for $h = 1, \ldots, H$.

Using this notation, the post-stratified estimator can now be defined (Lohr, 2009). The estimate of μ properly weighted to account for gender differences is

$$\hat{P}_2 = \sum_{h=1}^{H=2} w_h \overline{Y}_h \, .$$

The variance of $\hat{P}_2$ depends on whether or not we condition on the number of males and females in the sample. If we condition on these numbers, then

$$Var(\hat{P}_2) = \sum_{h=1}^{H} w_h^2 \operatorname{var}\left(\overline{Y}_h\right),$$

where when $n > n_h > 0$,

$$\operatorname{var}\left(\overline{Y}_h\right) = \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right).$$

d. Best Linear Unbiased Prediction

The third estimator is based on the best linear unbiased prediction (BLUP) (Robinson, 1991). This estimator also takes into account gender by estimating gender specific rates, where gender is considered as a random effect. We use

population weights with BLUP estimates similar to post-stratification. We define two estimators using both known and estimated variance components.

BLUP is similar to a simple weighted estimator with its use of weights, but differentiates itself by estimating gambling rates for domains as if they are random effects. First, we define gender specific constants, or shrinkage factors, where $k_h = \frac{V^2}{V^2 + \frac{S_h^2}{n_h}}$ denotes the shrinkage constant for domain *h*. These shrinkage constants omit the finite population correction factors for domain h when evaluating $\text{var}\left(\overline{Y}_h\right)$ (as in Stanek and Singer, 2004). Let the weighted least squares estimate of the mean be given by, $\overline{Y}^* = \sum_{h=1}^{H} w_h^* \overline{Y}_h$, where

$$w_h^* = \frac{1 \big/ \text{var}\left(\overline{Y}_h\right)}{\sum_{h=1}^{H} 1 \big/ \text{var}\left(\overline{Y}_h\right)}.$$

These weights assume independence of sample domain means.

The best linear unbiased predictor for domain *h* is given by $P_h^* = \overline{Y}^* + k_h(\overline{Y}_h - \overline{Y}^*)$. Accordingly, our best linear unbiased prediction estimator is represented as,

$$\hat{P}_3 = \sum_{h=1}^{2} w_h P_h^* .$$

e. Empirical Best Linear Unbiased Prediction

The empirical BLUP is the last estimator we consider of the prevalence of problem gambling. It is similar to the BLUP but instead of various combinations of population means and population variances, the EBLUP uses combinations of *sample* means and *sample* variances to arrive at its estimate. The EBLUP, consequently, uses estimated variance components.

The empirical best linear unbiased predictor replaces μ^* and k_h using estimates of V^2, S_h^2 and w_h^*. The estimate of V^2 is given by $\hat{V}^2$, where for sample $\Omega^{(d)}$, $\hat{V}^2$ is

$$\hat{V}_d^2 = \frac{1}{H-1}\sum_{h=1}^{H}(\bar{y}_{hd} - \hat{\bar{\mu}}_d)^2$$

where $\hat{\bar{\mu}}_d = \frac{1}{H}\sum_{h=1}^{H}\bar{y}_{hd}$. We estimate the gender-specific shrinkage constant for domain *h* by

$$\hat{k}_h = \frac{\hat{V}^2}{\hat{V}^2 + \frac{\hat{S}_h^2}{n_h}},$$

where for sample $\Omega^{(d)}$ and $n_h > 1$, $\hat{S}_h^2$ is given by $\hat{s}_h^{(d)2} = \frac{1}{n_h - 1}\sum_{s\in\Omega_h^{(d)}}(y_s - \bar{y}_{hd})^2$ for sample $\Omega^{(d)}$. Similarly as with the BLUP, this shrinkage constant omits the population correction factors for domain *h*. Thus, the empirical best linear unbiased prediction for domain *h* is given by $\hat{P}_h$ where for sample $\Omega^{(d)}$, $\hat{P}_h$ is given for sample

$\Omega^{(d)}$ by $\hat{p}_{hd} = \hat{\bar{\mu}}_d + \hat{k}_{hd}(\bar{y}_{hd} - \hat{\bar{\mu}}_d)$, leading to the EBLUP (Kleffe and Rao, 1992) given by

$$\hat{P}_4 = \sum_{h=1}^{2} w_h \hat{P}_h .$$

f. An Example

To make the ideas clear, an example will be used to illustrate these four methods. The simple finite population described earlier and depicted in Table 1 will be used, where N=5. In this situation, problem gambling status cannot be examined because the population is too small to realistically draw conclusions. This example will only evaluate general gambling as a means to illustrate the estimators.

As stated previously, a sample of four subjects is taken from the finite population. There are $\binom{N}{n} = 5$ sets of subjects. We refer to a particular set of sample subjects by the index *d=1,…,D=5*. Gambling status for each subject in a set is given in Table 1. Using these sample sets and equations defined in the Methods section, we can illustrate the estimates for the sample mean, post-stratified estimator, best linear unbiased prediction, and the empirical BLUP.

In this simple example, we will consider two BLUPs. From some perspectives, such as the Bayesian perspective, the population mean may be considered 'known', for example from a prior distribution. With this perspective, the BLUP could be evaluated using the population mean as opposed to the WLS estimate of the

population mean. In order to distinguish these two BLUPs we represent $\hat{P}_3$ to represent the BLUP based on the population mean by $\hat{P}_{3a}$ and $\hat{P}_{3b}$ to represent the BLUP based on the WLS mean.

To make this clear, these estimators will be written out explicitly. The BLUP based on the population mean, $\hat{P}_{3a}$, is represented as,

$$\hat{P}_{3a} = \sum_{h=1}^{2} w_h P_{ha}^*$$

where $P_{ha}^* = \mu + k_h(\bar{Y}_h - \mu)$. The BLUP based on the WLS mean, $\hat{P}_{3b}$ is represented as,

$$\hat{P}_{3b} = \sum_{h=1}^{2} w_h P_{hb}^*$$

where $P_{hb}^* = \bar{Y}^* + k_h(\bar{Y}_h - \bar{Y}^*)$. In Table 3a, the corresponding shrinkage constants, means, and predictors used to calculate the two BLUPs, are provided. Shrinkage constants are the same for both BLUPS, the distinguishing factor is their predictors which are denoted appropriately *a* and *b* in their subscripts.

Table 3a: Ingredients for BLUP's

Sample (d)	Set $\Omega^{(d)}$	k_1	k_2	P_{1a}^*	P_{2a}^*	P_{1b}^*	P_{2b}^*
1	{ABCD}	0.1103	0.02686	0.60735	0.61074	0.78544	0.8055
2	{ABCE}	0.1103	0.02686	0.60735	0.58388	0.42912	0.3889
3	{ABDE}	0.0763	0.05231	0.59237	0.59477	0.5	0.5
4	{ACDE}	0.0763	0.05231	0.59237	0.59477	0.5	0.5
5	{BCDE}	0.0763	0.05231	0.63053	0.59477	0.81504	0.7841

Table 3a shows a slight difference in predictors for $\hat{P}_{3a}$ and $\hat{P}_{3b}$. The predictor, $\hat{P}_{3b}$, seems to produce slightly lower estimates. Analysis of bias and MSE later will bring this to fruition. In subsequent tables, $\hat{P}_{3a}$ and $\hat{P}_{3b}$ will be referred to as BLUPa and BLUPb.

Table 3b provides the ingredients necessary to compute the post-stratified mean and the EBLUP. These two tables were included to provide background into the subsequent calculations for estimating the prevalence of gambling.

Table 3b: Ingredients for Post-Stratified Mean and EBLUP

Sample (d)	Set $\Omega^{(d)}$	$\overline{Y}_{h=1}$	$\overline{Y}_{h=2}$	$\hat{V}^2$	$\hat{S}^2_{h=1}$	$\hat{S}^2_{h=2}$	$\hat{k}_{h=1}$	$\hat{k}_{h=2}$	$\hat{w}^*_{h=1}$	$\hat{w}^*_{h=2}$	$\hat{\mu}^*_{h=1}$	$\hat{\mu}^*_{h=2}$	$\hat{p}_{h=1}$	$\hat{p}_{h=2}$
1	{ABCD}	0.75	0.667	2.7778	0.333	0	0.962	1	0	0	0	0	0.641	1
2	{ABCE}	0.5	0.667	0.4444	0.333	0	0.8	1	0	0	0	0	0.533	0
3	{ABDE}	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.25	0.375	0.375
4	{ACDE}	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.25	0.25	0.375	0.375
5	{BCDE}	0.75	1	2.25	0	0.5	1	0.9	0	0	0	0	1	0.45

Using results from Tables 3a and 3b, Table 4 below depicts unconditional results for the 5 estimators. After the four estimators were calculated, their respective expected values, bias, and mean square errors were calculated over all samples. The mean square error is defined as,

$$MSE_i = \frac{\sum_{d=1}^{D} (\hat{P}_{id} - \mu)^2}{D}.$$

where *i=1,..,4* represents the i^{th} estimator, and $d = 1,\ldots,D$, *indexes* the samples, $\Omega^{(d)}$ of subjects. Bias is defined as $B_i = E(\hat{P}_i) - \mu$ where $E\left(\hat{P}_i\right) = \sum_{d=1}^{D} \frac{\hat{P}_{id}}{D}$.

Table 4: Unconditional Estimates of P_1, P_2, P_3,, and P_4

Sample (d)	Set $\Omega^{(d)}$	Sample Mean	Post-Stratified Estimator	BLUPa	BLUPb	EBLUP
1	{ABCD}	0.75	0.8	0.60871	0.793474	0.784615
2	{ABCE}	0.5	0.4	0.59797	0.413053	0.32
3	{ABDE}	0.5	0.5	0.59333	0.5	0.375
4	{ACDE}	0.5	0.5	0.59333	0.5	0.375
5	{BCDE}	0.75	0.8	0.61623	0.802658	0.78
Expected Value	0.6	0.6	0.601911	0.601837	0.526923	
Bias	0	0	0.001911	0.001837	-0.07308	
MSE	0.015	0.028	0.0000828	0.026687	0.043886	

Each of the estimators is similar to each other across respective sample sets. This can be seen by simply looking across cells. It is also summarized by the expected value result given in the bottom half of the table. Expected values for the four estimates are identical for the sample mean and the post-stratified estimator. These estimators are also unbiased. BLUPa is more accurate (based on the MSE) and has a small bias. BLUPb also has a very low MSE value but a slightly smaller bias than BLUPa. The EBLUP produces the highest MSE overall and is also the most biased.

Next, the four estimation methods are evaluated conditionally on the number of females in a sample set of subjects. The estimators for set $\Omega^{(d)}$ have the same values as in Table 4. The expected values, bias and mean square errors change since

they are evaluated only over the samples with n_f females. Tables 5a and 5b summarize the results.

Table 5a: Conditional Estimates of P_1, P_2, $P_{3,}$, and P_4 given n_f=1

Sample (d)	Set $\Omega^{(d)}$	Sample Mean	Post-Stratified Estimator	BLUPa	BLUPb	EBLUP
1	{ABCD}	0.75	0.8	0.60871	0.79348	0.78462
2	{ABCE}	0.5	0.4	0.59797	0.41305	0.32
Expected Value		0.625	0.6	0.603337	0.60323	0.55231
Bias		0.025	0	0.003337	0.00323	-0.0477
MSE		0.0156	0.04	0.0000289	0.03618	0.05397

Table 5a illustrates the conditional estimates for the sample set of subjects containing only one female. Expected values for all estimators except the EBLUP remain virtually unchanged from Table 4. Bias has increased for the simple mean, BLUPa and BLUPb. The EBLUP and post-stratified mean have the highest MSE values, while BLUPa and the simple mean have the lowest. MSE for BLUPb has increased slightly.

Table 5b: Conditional Estimates of P_1, P_2, $P_{3,}$, and P_4 given n_f=2

Sample (d)	Set $\Omega^{(d)}$	Sample Mean	Post-Stratified Estimator	BLUPa	BLUPb	EBLUP
3	ABDE	0.5	0.5	0.59333	0.5	0.375
4	ACDE	0.5	0.5	0.59333	0.5	0.375
5	BCDE	0.75	0.8	0.61623	0.802658	0.78
Expected Value		0.5833	0.6	0.600961	0.600886	0.51
Bias		-0.01667	0	0.000961	0.000886	-0.09
MSE		0.0138	0.02	0.000117	0.020356	0.03645

Table 5b illustrates the conditional estimates for the sample set of subjects containing two females. Except for the post-stratified mean with a static expected value, all other expected values decreased with this analysis. Bias has slightly increased in the simple mean and the EBLUP, but it has decreased in BLUPa. All MSE values have decreased except for BLUPa. The EBLUP produces the highest MSE value again, while BLUPa still produces the lowest value despite its marginal increase.

In the finite population example of five samples from the simple population of five subjects, four methods were compared against each other using a simple example. The method producing the lowest mean square error score, both unconditionally and conditionally is BLUPa, the best linear unbiased predictor using the population mean. We will use this BLUP moving forward in the analysis.

This simple example depicted a very limited situation. The sampling fraction, given by 0.8, is extremely high and very unrealistic. Also, due to the particular example, all samples contain at least one male or at least one female. This is unlikely to be true in practice. In other settings, samples containing only males or samples containing only females will occur.

Another discussion point is in regards to the post-stratified mean results. Holt and Smith (1979) conclude that the post-stratified mean performs better when examined conditionally, as opposed to unconditionally. In this simple example, the post-stratified mean was examined both conditionally and unconditionally and

produced some of the highest MSE values in all analyses. This same result does not occur in a more realistic setting discussed subsequently in this paper.

CHAPTER VIII

SIMULATION

Thus far, a simple finite population has been described and four estimation techniques have been defined to estimate the prevalence of problem gambling in the context of the gambling study. We evaluated these four methods through the simple population example which examined general gambling. This population consisted of $N = 5$ subjects, of which 3 were male and 2 were female, with randomly selected samples of $n = 4$ subjects. We illustrate a second example where N=20 with 12 males and 8 females. This example is small enough to be studied by complete enumeration of all possible samples, and also illustrates the simulation; however, it is still too small to address problem gambling. This will be analyzed in the last example.

We will now describe the population consisting of $N = 20$ subjects, N_1=8 males and N_2 =12 females. Randomly selected samples consisting of $n = 5$ subjects are taken which result in a total of $\binom{N}{n}$ possible samples, where

$$\begin{aligned}\binom{20}{5} &= \frac{20(19)(18)(17)(16)}{(5)(4)(3)(2)1} \\ &= \frac{4(19)(3)(17)(4)}{1} \\ &= 15504\end{aligned}$$

possible samples. The possible samples can be enumerated and estimators can be evaluated for each sample. We also illustrate how the possible samples can be

simulated and the simulation can be used to evaluate the properties of estimators constructed over possible samples.

a. Assumptions

In order to account for varying data population combinations in our simulation, the post-stratified estimator, BLUP, and EBLUP must be clearly re-defined for all possible sample sets. Complications occur when $n_h \leq 1$ for a given gender. Without or with few sample members of a domain, we must make assumptions about the domain mean and variance.

The first assumption we must make is for samples containing only one gender. In situations where the samples of n=5 contain 5 males or 5 females, we assume that males and females have the same gambling rate. This allows us to use the sample mean as the estimator for the prevalence of gambling.

The second assumption we must make is for samples containing only one subject of a given gender. In situations where the samples of n=5 contain 4 males and1 female or 4 females and 1 male, the variance in a domain cannot be properly estimated for that one subject. Consequently for the EBLUP, we assume the shrinkage constant, $\hat{k}_h$, is 1 for that subject, leaving the means to cancel out in the estimator. Thus, the EBLUP reduces to simply the sample mean for that one subject of a given gender, multiplied by the population weight.

b. Initial Results of Population with N=20

In this population, we set the overall gambling rate is set to be 55% (11 gamblers out of the 20 total subjects). Among females of the population the gambling rate is 58.333% (7/12), and among males of the population the gambling rate is 50% (4/8). The rates are similar to rates observed in preliminary survey data results (first 2 months) from the gambling study on response to the gambling question "In the past 12 months, how often have you purchased lottery tickets?" Possible answers were: 'never'; 'less than once a month'; '1-3 times a month'; and '1-4+ times a week'. Subjects who answered 'never' were coded as non-gamblers and subjects who gave one of the other three answers were coded as gamblers.

Using this population, samples of $n = 5$ subjects were created and the prevalence of gambling was estimated for each sample using the techniques discussed previously. First, the makeup of the sample by gender and gambling status was calculated. In order to calculate the number of males or females in each of the 15,504 samples, the following table was constructed:

Table 6: Construction of Samples

	M	**F**	**Total**
Sample	x	(5- x)	5
Remainder	(8-x)	(7+x)	15
Population	8	12	20

This leads to,

$$\binom{N_{males}}{x}\binom{N_{females}}{n_{sample}-x},$$

which equals the number of distinct sets with *x=0,…,5* male subjects. For example, to enumerate the number of samples containing 2 males, we have

$$\binom{8}{2}\times\binom{12}{3}=28\times 220=6160.$$

There are 6,160 samples containing 2 males and 3 females. To enumerate the number of male gamblers in these samples, we use similar notation,

$$\binom{n_{Mgamblers}}{y}\times\binom{N_{males}-n_{Mgamblers}}{x-y}.$$

This calculates the number of male gamblers in a particular sample with *y=0,…,5* male gamblers with *x* males. To enumerate the number of male gamblers in the 792 samples containing 2 males we calculate,

$$\begin{aligned}\binom{4}{0}\times\binom{4}{2}&=1\times 6=6\\ \binom{4}{1}\times\binom{4}{1}&=4\times 4=16.\\ \binom{4}{2}\times\binom{4}{0}&=6\times 1=6\end{aligned}$$

The sum of these three calculations is 28. Dividing 28 into the total number of samples, 6,160, gives us 220. If we multiple 220 by 6 (0 gamblers), 16 (1 gambler), and 6 (2 gamblers) we arrive at 1,320, 3,520, and 1,320 for 0, 1, and 2 gamblers,

respectively. The calculations for all possible combinations of males and male gamblers per sample are summarized in Table 7.

Table 7: Number of Samples Containing *x* Males and *y* Gamblers

#Males	#Samples	#Gamblers					
		0	**1**	**2**	**3**	**4**	**5**
0	792	792					
1	3,960	1,980	1,980				
2	6,160	1,320	3,520	1,320			
3	3,696	264	1,584	1,584	264		
4	840	12	192	432	192	12	
5	56	0	4	24	24	4	0

Similar logic was applied to calculate the number of samples containing *x=0,1,2,3,4 or 5* females and their corresponding gambling status. Table 8 summarizes these results.

Table 8: Number of Samples Containing *x* Females and *y* Gamblers

#Females	#Samples	#Gamblers					
		0	**1**	**2**	**3**	**4**	**5**
0	56	56					
1	840	420	420				
2	3,696	560	1,960	1,176			
3	6,160	280	1,960	2,940	980		
4	3,960	40	560	1,680	1,400	280	
5	792	1	35	210	350	175	21

Using this breakdown of samples by gender and gambling status, estimates were then constructed for each group of samples. For example, separate estimates were constructed for the 792 samples containing 5 females and 0 males, for the 21 samples containing 5 gamblers, the 175 samples containing 4 gamblers, and so on. The complete table is provided in Appendix A. Table 9 is an excerpt from the

extensive table provided in the appendix. It will be used to illustrate the distribution of samples by gambling status and for the calculation of weights.

After all the estimators were calculated separately, they were multiplied by weights. Weights were constructed based on the proportion of samples in relation to the total number of samples. To exemplify this, we use the same example as above. For the 6,160 samples containing 2 males and 3 females, the 1,320 samples with 0 male gamblers produce 60 samples with 0 female gamblers, 420 samples with 1 female gambler, 630 samples with 2 female gamblers, and so on. To calculate appropriate weights, the number of female samples containing *y=0,…,5* gamblers is divided by the total number of samples which is 15,504. So, the first weight provided in Table 9 refers to the 60 samples with 0 male gamblers and 0 female gamblers in a sample of 2 males and 3 females total. Dividing 60 by 15,504 gives us 0.00387. For the 420 samples containing 0 male gamblers and 1 female gambler, $420/15504 = 0.02709$.

This table also includes a column containing conditional weights that sum to one. These are weights used later to evaluate the estimators conditionally. Table 9 summarizes these results.

Table 9: Construction of Weights by Female Gambling Status

#Males	#Females	#Samples	#Male Gamblers	#Female Samples	#Female Gamblers	WEIGHT	Conditional Weight
2	3	1320	0	60	0	0.00387	0.00974
				420	1	0.02709	0.06818
				630	2	0.04063	0.10227
				210	3	0.01354	0.03409
		3520	1	160	0	0.01032	0.02597
				1120	1	0.07224	0.18182
				1680	2	0.10836	0.27273

#Males	#Females	#Samples	#Male Gamblers	#Female Samples	#Female Gamblers	WEIGHT	Conditional Weight
				560	3	0.03612	0.09091
		1320	2	60	0	0.00387	0.00974
				420	1	0.02709	0.06818
				630	2	0.04063	0.10227
				210	3	0.01354	0.03409
TOTALS		**6160**		**6160**		**0.39732**	**1.00000**

Using the sample distribution and weights exemplified in Table 9, expected values, bias and mean square errors were then calculated to evaluate the four estimators. Table 10, below, summarizes the results. In this finite population of N=20 subjects, 12 females and 8 males, the gambling rate was set to be 55 percent. The results of this distribution lead to bias in three estimators, as was expected for the BLUP and EBLUP. Overall, the BLUP still produces the lowest MSE value.

Table 10: Unconditional Estimates Summary

	Simple Mean	Post-Stratified Mean	BLUP	EBLUP
Expected Value	0.55	0.551522	0.551820	0.5551
Bias	0	0.001522	0.001820	0.0051
MSE	0.03908	0.049098	0.001767	0.06055

The simple mean remains unbiased as it was in the previous simple example with *n=5* total subjects. The post-stratified mean, the BLUP and the EBLUP are all biased due to the assumptions just discussed. The EBLUP produces the highest MSE value which is due to the fact that its variance cannot be properly estimated in all cases since it depends on the sample composition. Ultimately, the BLUP produces the lowest mean square error.

Lastly, the variance of the sample mean was calculated to support the unbiased MSE value of 0.03908. The variance with correction factor was calculated using,

$$\sigma^2_{\bar{Y}} = (1-\frac{n}{N})\frac{1}{n}S^2_w$$

where $S^2_w = \frac{1}{N-1}\sum_{s=1}^{N}(y_s - \mu)^2$. The first part of this equation reduces to 0.26053 and the second part, S^2_w, reduces to 0.15. Multiply 0.26053 by 0.15 and you arrive at 0.03908 which is equivalent to the unbiased MSE produced by the simple mean. This verifies that the mean square error is unbiased.

Now the same estimators are examined, conditional on gender. Tables 11 and 12 summarize the results for each sample conditional on the number of females in the sample. The numbers of samples containing 0, 1, 2, 3, 4, or 5 females are included in the table as well as overall unconditional estimates for comparison purposes.

Table 11 provides conditional bias results for each of the four estimators. Bias is the same across all estimators for samples containing 5 females and for samples containing 0 females. This result is the direct outcome of assuming equal gambling rates among males and females for samples containing only one gender. It is interesting to note that although the simple mean had the lowest bias (zero) for the unconditional results, the post-stratified mean produces the lowest bias in three out of the four remaining sample groups. The simple mean produces the lowest bias only for the group containing three females and two males.

Table 11: Conditional Estimates Bias Summary

#Females	#Samples	Simple Mean	Post-Stratified Mean	BLUP	EBLUP
Overall	15,5504	0	0.001522	0.001820	0.0051
5	792	0.03334	0.03334	0.03334	0.03334
4	3,960	0.01666	-0.000002	0.000784	0.00055
3	6,160	0	-0.000004	0.000455	0.00537
2	3,696	-0.01667	-0.000002	-0.000201	0.00498
1	840	-0.03334	-0.00001	-0.000699	0.00189
0	56	-0.05	-0.05	-0.05	-0.05

Table 12 provides conditional MSE results for each of the four estimators. Again, MSE values are the same across estimators for samples of 5 females and sample containing 0 females. For the remaining four sample groups, the BLUP produces the lowest MSE values. It is interesting to note that there is no shifting between estimators among the samples. BLUP yields the lowest MSE values across the board, conditionally and unconditionally, with the exception of samples containing all or no females.

Table 12: Conditional Estimates MSE Summary

#Females	#Samples	Simple Mean	Post-Stratified Mean	BLUP	EBLUP
Overall	15,5504	0.03908	0.049098	0.001767	0.06055
5	792	0.03205	0.03205	0.03205	0.03205
4	3,960	0.03856	0.055909	0.000049	0.06289
3	6,160	0.04101	0.041005	0.00005	0.05504
2	3,696	0.03938	0.049297	0.000041	0.06488
1	840	0.03369	0.093213	0.00003	0.10017
0	56	0.02393	0.02393	0.02393	0.02393

Conditional weighted MSE values were calculated to verify results. This was performed by multiplying each conditional MSE value shown in Table 12 by its

corresponding number of samples, provided in this table, summing over all conditional groups and dividing by the total number of samples. This number should equal the unconditional estimate result for that estimator. Using the simple mean as an example, multiply 0.03205 by 792, 0.03856 by 3,960, 0.04101 by 6,160, and so on down the line. Summing of those results and dividing by 15,504 produces a value of 0.03908 which is equivalent to the unconditional estimate for the simple mean. The results are verified.

c. SAS Simulation Results

Simulations were performed using SAS, version 9.4. Two simulations were conducted. The first had a population of *N=20* and samples of *n=5*. This simulation duplicated the finite population in the previous example and was simulated 15,504 times. The second simulation had a population of *N=1,000* and samples of *n=200*. This data was simulated 10,000 times unconditionally and then conditionally on the number of females in each sample. It is through this example that problem gambling will be addressed. MSE, bias, squared-bias, variance by gender, and variance of the post-stratified mean were calculated and conclusions were made.

i. First Simulation with N=20

Continuing with the current example population of *N=20* and samples of *n=5*, a SAS simulation program was written to simulate this data, 15,504 times. The idea

was to re-create conclusions found previously when enumerating the samples and calculations by hand. Table 13 provides SAS output for these results.

Table 13: Unconditional Bias and MSE Summary from SAS: 15,504 Simulations

Bias and MSE Summary
N=20;n=5;#F=12;#FG=7;seed=123456;

The MEANS Procedure

Variable	N	Mean	Std Dev
BiasYBAR	15504	0.0010449	0.1959539
BiasPostStrat	15504	0.0026294	0.2204970
BiasBLUP	15504	0.0017594	0.0414555
BiasEBLUP	15504	0.0064249	0.2449518
MSEybar	15504	0.0383965	0.0507693
MSEpoststrat	15504	0.0486227	0.0559229
MSEblup	15504	0.0017215	0.0119904
MSEeblup	15504	0.0600388	0.0624441

This table provides mean bias and MSE estimates for the four estimators. These results are in line with earlier results summarized in Table 8. The BLUP produces the lowest MSE value and the EBLUP produces the highest MSE with the most bias. The simple mean has the second lowest MSE value and the smallest bias. Standard deviations were used to calculate 95% confidence intervals for the mean MSE values. Table 14 lists the intervals and corresponding 'true values' which refer to results calculated previously (Table 10).

Table 14: 95% Confidence Intervals for Simulated Estimators

Estimator	95% CI	True Value
YBAR	(0.03759,0.03919)	0.03908
POST-STRAT	(0.04774,0.04950)	0.049098
BLUP	(0.00153,0.00191)	0.001767
EBLUP	(0.05905,0.06101)	0.06055

Table 14 illustrates the accuracy of our simulation program. All true values are included in the 95% confidence intervals using mean MSE and standard deviation values from SAS. Next, a plot of the mean MSE values for the four estimators is provided for visualization purposes. Figure 1 portrays a plot of mean MSE values multiplied by the number of females in the sample.

Figure 1: Mean MSE Conditional on Females: 15504 Simulations

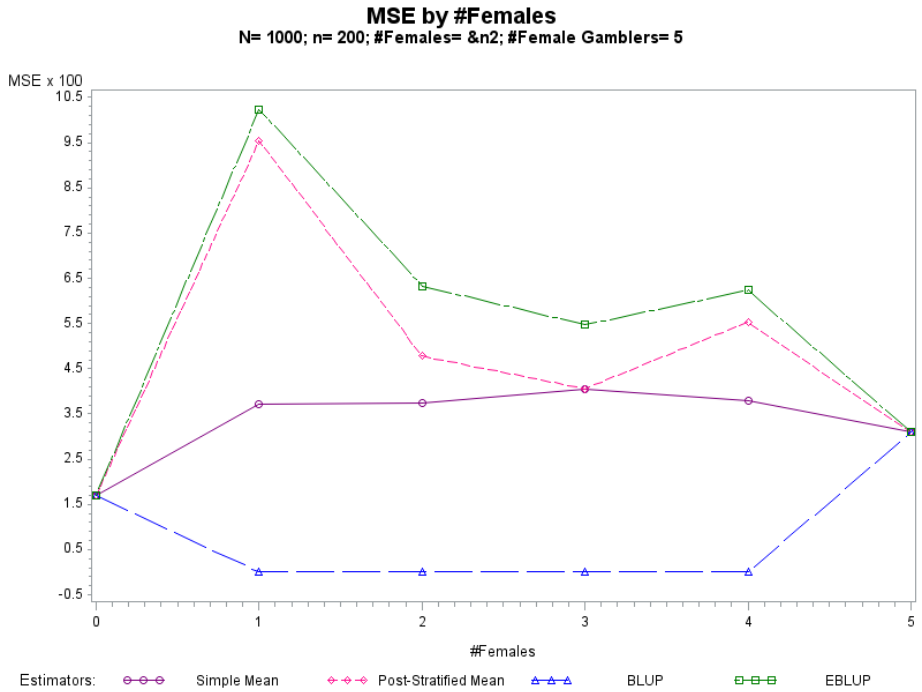


Figure 1 dramatically portrays the difference among the estimators. MSE values were multiplied by 100 to make the scale and plot less obscure. It is very easy to see that the BLUP (in blue) continually performs much better than the other estimators with its very low mean MSE. The EBLUP and post-stratified mean are symmetrically worse and the simple mean is in the middle consistently. The estimators have the same values for 0 females and for 5 females due to our assumption of equal gambling rates made previously.

ii. Second Simulation with N=1,000; Unconditional

Next, the population size and sample size were increased. Simulations were performed in SAS 10,000 times on a population of *N=1,000*, a simple random sample size of *n=200*, and a crude problem gambling rate of 1.5 percent. In this population, males and females both represent half of the population with 10 of the 500 males being problem gamblers and 5 of the 500 females being problem gamblers. This relative problem gambling rate of 2 between domains is closer to the expected problem gambling rate in the Massachusetts population. Table 15 portrays SAS output and results from this simulation.

Table 15: Unconditional Bias and MSE Summary from SAS: 10,000 Simulations

Bias and MSE Summary for N= 1000; n= 200; Male/Female Ratio= 0.5; Crude Gambling Rate= 1.5

The MEANS Procedure

Variable	N	Mean	Std Dev
BiasYBAR ($\times 10^{-4}$)	10000	-1.59500	77.038
BiasPostStrat ($\times 10^{-4}$)	10000	-1.58985	77.244

Variable	N	Mean	Std Dev
BiasBLUP (x 10-4)	10000	-3.61174	19.750
BiasEBLUP (x 10-4)	10000	-13.357	78.950
MSEybar (x 10-6)	10000	5.9367	8.4341
MSEpoststrat (x 10-6)	10000	5.9685	8.4727
MSEblup (x 10-6)	10000	0.40305258	0.53221094
MSEeblup (x 10-6)	10000	6.4110	8.1193

Results shown in Table 15 magnify the results found in previous examples. First of all, all values for the mean and standard deviation columns were multiplied by 100,000 to make results more clear. The mean MSE for the BLUP is significantly lower than the other estimators. There is less of a drastic difference between the remaining 3 estimators: the simple mean and the post-stratified mean are especially close. One change to note is that the post-stratified estimator produces the lowest overall bias value now. Figure 2 below, provides a visualization of these results.

Figure 2: Mean MSE Conditional on Females: 10,000 Simulations

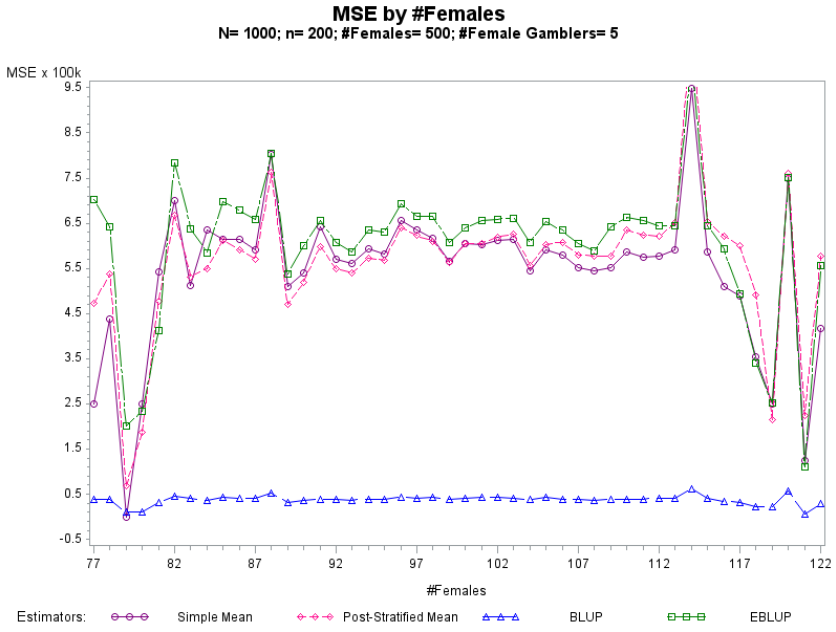
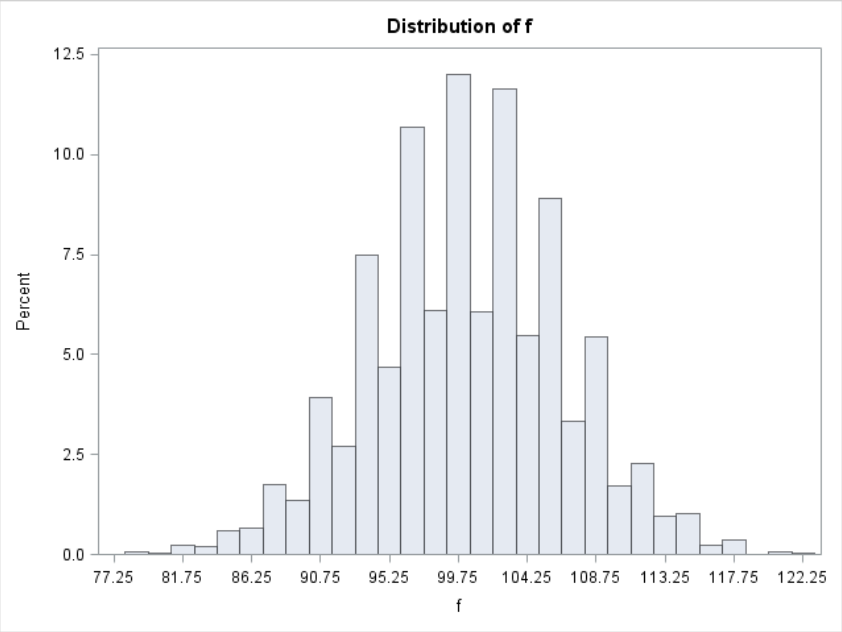


Figure 2 very clearly depicts results. Again, the MSE vertical scale axis was changed (multiplied by 100,000 this time) to provide a graph that is more readable. The BLUP produces the lowest MSE value by far, with the other three estimators producing similar values to each other, except much higher.

iii. Conditional Simulations of N=1,000

Due to the large population and sample size, it is not feasible to enumerate all possible samples by hand as we did earlier for the smaller population of *N=20*. This makes for difficulty in estimating the estimators conditionally. Figure 2 touched on this but a histogram produced in SAS and shown in Figure 3 below provides more insight into the composition of females in all the samples from this population of *N=1,000*.

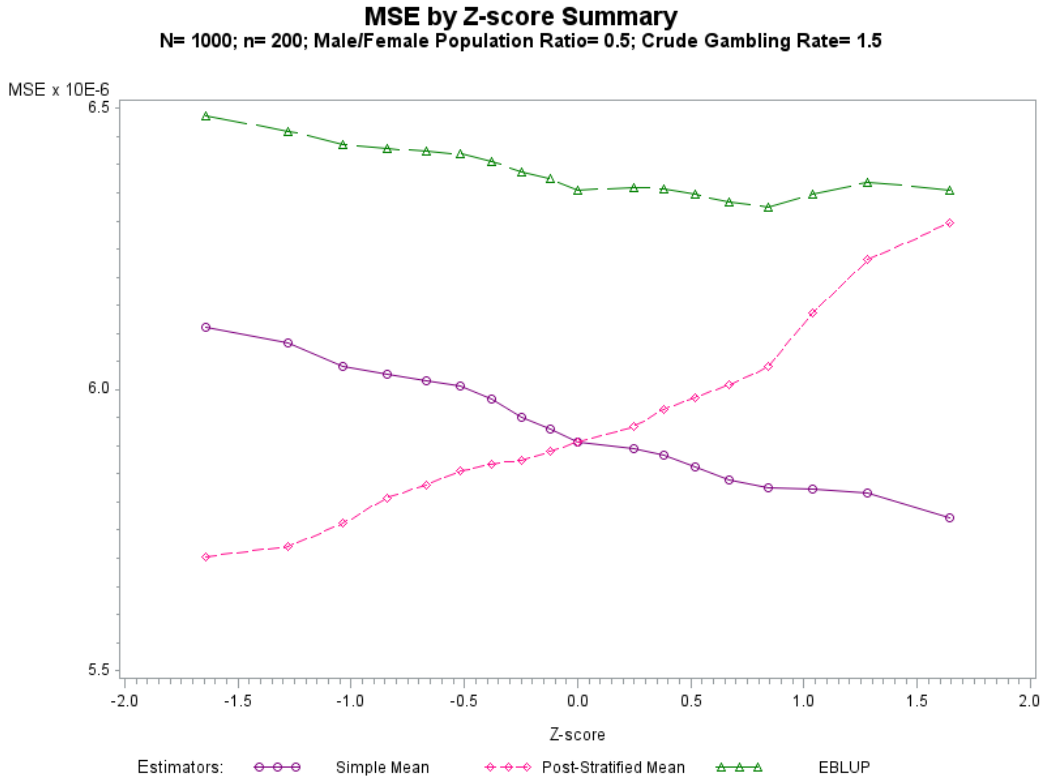
Figure 3: Histogram of #Females in 10,000 Simulated Samples of n=200



This histogram shows how the number of females in these samples ranges from about 77 to around 122. This is a wide range to analyze the estimators conditionally on, so we formed intervals based off 95% z-score estimates that included 5% of the samples and evaluated the estimators conditional on samples in a group. This span of the number of females in a sample ranges from 87 to 111 in 18 intervals. For example, the first 5% interval contains less than 88 females in a sample and this corresponds to a z-score of -1.64. The next 5% interval contains between 88 and 91 females, corresponding to a z-score of -1.28. This continues up to a z-score of 1.64 with the 5% interval containing between 109 and 111 females in the sample.

MSE, squared bias, and variance values were calculated for the simple mean, the post-stratified mean and the EBLUP. The BLUP will no longer be considered a viable estimator for the prevalence of problem gambling, since it depends on population estimates of the variance. Figure 4 depicts the MSE values for the simple mean, the post-stratified mean and the EBLUP by the 18 z-scores.

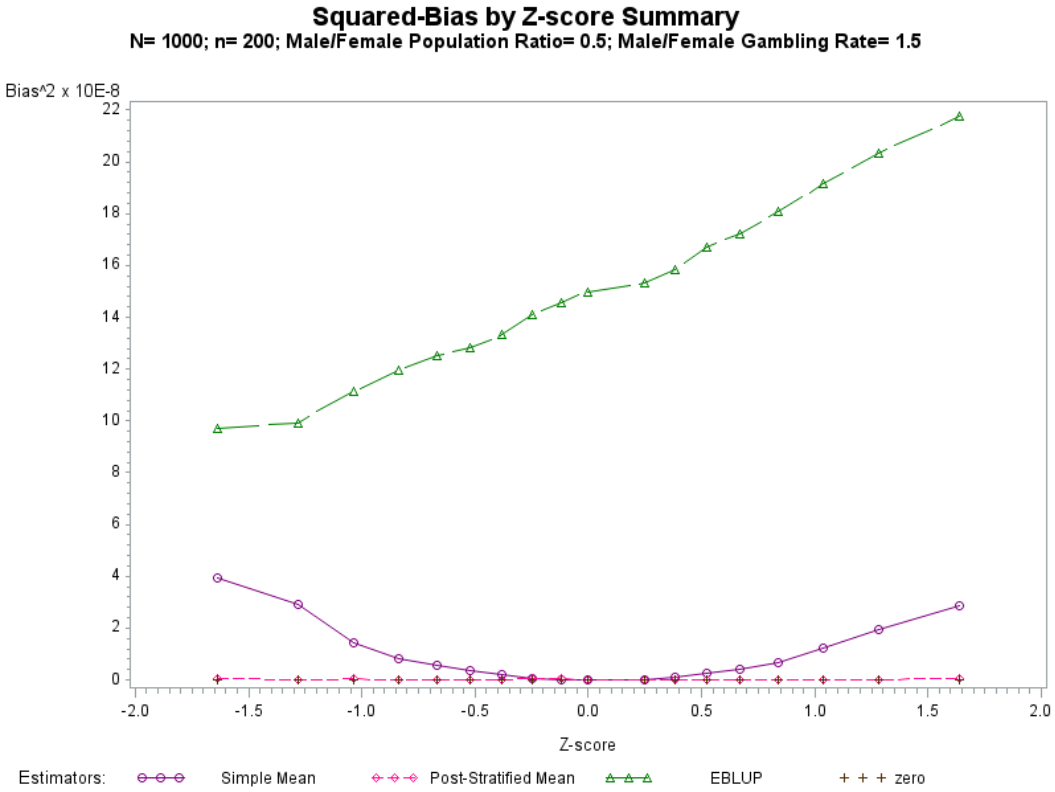
Figure 4: Mean MSE Conditional on Z-scores: 10,000 Simulations



MSE values were multiplied by 100,000 to adjust the scale in this figure. The plot shows an EBLUP that consistently performs the worst out of all the estimators again. The simple mean and post-stratified mean have an interesting relationship. When there are more males than expected in the sample (n_m>100), the post-stratified mean performs better by producing the lower mean MSE values. When there are more females than expected (n_f>100), the simple mean performs better. We examine plots of squared bias and variance to further analyze this result.

MSE is equal to the sum of squared bias plus the variance. These values were calculated and plotted to provide context for the mean MSE values portrayed in Figure 4. Below, Figure 5 illustrates squared bias for the three estimators.

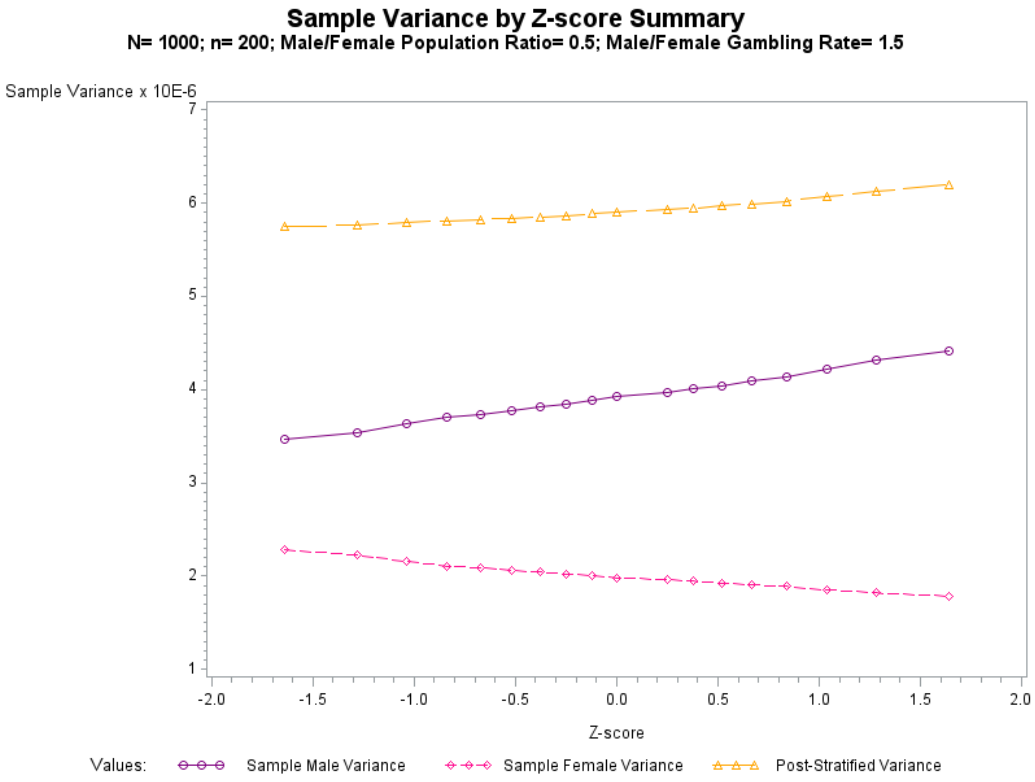
Figure 5: Squared-Bias Conditional on Z-Scores: 10,000 Simulations



The vertical axis in Figure 5 was multiplied by 10 million to adjust the scale (100 relative to the MSE plot). The result is a very clear larger bias in the EBLUP as compared to the other estimators. The simple mean produces a U-shaped plot for bias, which makes sense- when there are more males OR females expected in a sample, bias in the sample mean increases. The post-stratified mean has almost no bias which is further exemplified by the line included at zero. The green line at zero and the pink line for the post-stratified estimator overlap. As noted before, the vertical axis on this plot was magnified by 10 million so ultimately, there is very little bias overall in these estimators.

Figure 6 shows the sample variance for males, the sample variance for females, and the variance of the post-stratified mean, all conditional on sample size. The vertical axis was multiplied by ten thousand.

Figure 6: Sample Variance of the Mean Conditional on Z-Scores: 10,000 Simulations



The variance of the post-stratified mean in this plot was calculated using a correction factor equivalent to $\frac{N-n}{N-1}$. Including this value in the calculations for the variance of the post-stratified estimator leads the mean MSE values for the post-stratified estimator to be almost equivalent to its variance. This makes sense because there was relatively no bias.

The sample variance for the proportion of problem gamblers among males and the sample variance for the proportion of problem gamblers among females travel on opposite paths. When the number of males in a sample is greater than expected, the sample male variance is lowest. When the number of females in a sample is greater than expected, the sample variance for females is lowest. The opposition is easy to see in Figure 6 - when there are fewer females than expected in a sample, the sample female variance is at its highest point while the sample male variance is at its lowest point.

CHAPTER IX

DISCUSSION

In this paper, estimating the prevalence of problem gambling in the context of the gambling study was the objective. Four estimators were considered and evaluated using multiple examples and simulations, both conditional and unconditional on gender. Mean square errors were used to assess the accuracy of the simple mean, the post-stratified mean, the best linear unbiased predictor (BLUP) and the empirical BLUP (EBLUP).

Accuracy is measured by a low MSE. Initial results led to the BLUP producing the lowest MSE values. It may seem that this result would lead to recommending the use of BLUP in practice; however, the BLUP was later removed from analyses since it cannot be estimated without assumptions about the variance in practice. The empirical BLUP produced the highest MSE values. This is attributed to the fact that the EBLUP uses estimates of variances, leading the EBLUP to produce more variable results.

The sample mean and the post-stratified mean seem to estimate prevalence better in certain situations, as given by their low MSE values. The inverse relationship between the slopes of the simple mean and post-stratified mean portrayed in Figure 4 is interesting. As z-scores increase, or as the number of females in a sample becomes greater than their expected values, the mean square error for the simple mean decreases, while the mean square error for the post-stratified mean increases. A closer look into squared-bias and variance, the two

components of the MSE, provide insight. Squared bias is shown in Figure 5 on a scale that is 100 times smaller than the scale of the MSE's shown in Figure 4. A result of this small bias is not likely one that explains the relationship between the simple mean and the post-stratified mean.

This leads us to Figure 6 which compares the sample variances for males and females. This figure is on the same scale, (10^6), as Figure 4 and immediately, one can see an inverse relationship between the slopes that represent sample variances for males and females. This inverse relationship is similar to the MSE values shown in Figure 4 for the simple-mean and post-stratified mean.

Variance is proportional to the mean and the post-stratified mean accounts for different variances. When there are more males than expected in a sample, the post-stratified mean has a smaller variance for males due to the relatively larger number of males in the sample. This is why the variance for males is lower in this situation. When there are fewer males than expected in a sample, the post-stratified mean has a larger variance since there are fewer males in the sample and hence a relatively larger male variance, leading to the increasing variance portrayed in Figure 6.

The same is true for females - when there are fewer females than expected in a sample, the contribution to the variance of the post-stratified mean from females is larger, while when there are more females than expected in a sample, the contribution to the variance of the post-stratified mean is smaller.

Analysis of the variance of the simple mean and variance of the post-stratified mean is included to solidify results. These variance estimates were

calculated using population parameters of the prevalence of problem gambling as opposed to the sample estimates used in the simulation program. The goal here is to verify results- the results concluded based on 10,000 simulations of sample estimates should be almost equivalent to results based on population estimates, conditional on the number of males and females in a sample.

The variance of the simple mean, conditional on the number of males and females in a sample is defined by,

$$Var\left(\bar{Y} \mid n_m, n_f\right) = \left(1 - \frac{n}{N}\right)\frac{\bar{\pi}\left(1-\bar{\pi}\right)}{n}$$

where, $\bar{\pi} = \sum_{h=1}^{2} \frac{n_h}{n}\mu_h$, and $n_1 = n_m$, while $n_2 = n_f$. The variance of the post-stratified mean is defined by

$$Var\left(\bar{Y}_{PS}\right) = \sum_{h=1}^{2}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{\pi_h\left(1-\pi_h\right)}{n_h}.$$

In our last example with a population of N=1,000 and n=200, the crude problem gambling rate was 1.5%, a $\mu_1 = 0.02$ problem gambling rate among males and a $\mu_2 = 0.01$ problem gambling rate among females. Values of these two expressions for the variance were calculated in Excel and summarized in Figure 7.

Figure 7: Variance of Simple and Post-Stratified Mean

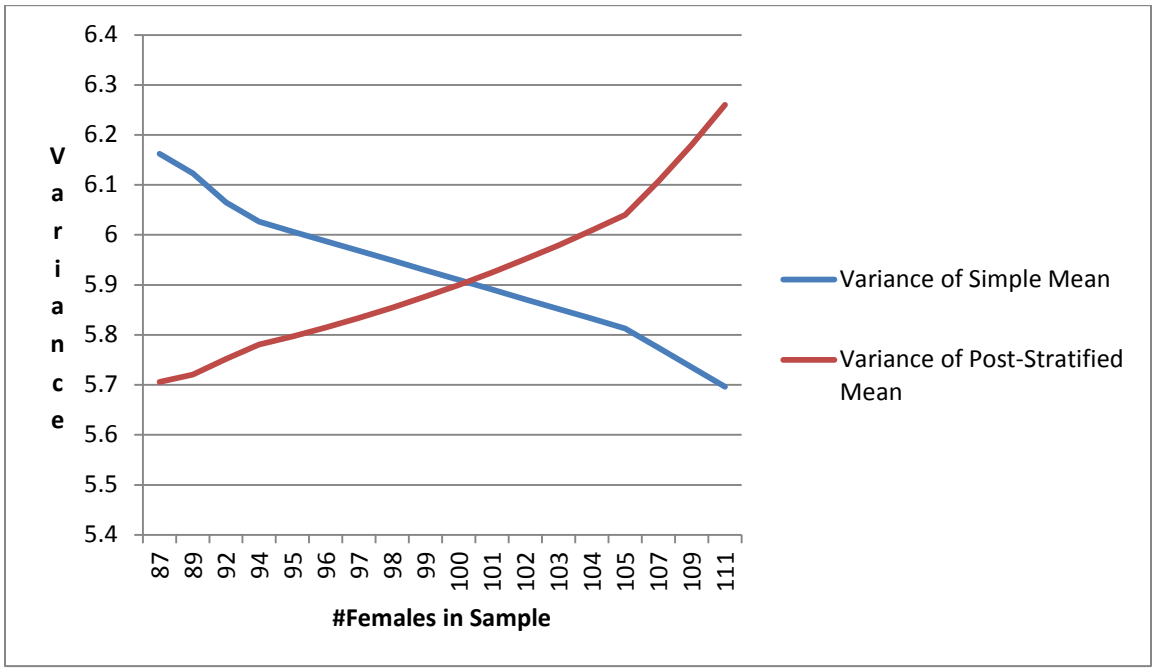


Figure 7 and Figure 4 are almost identical. When there are fewer females than expected in a sample, the post-stratified mean has a lower variance and thus mean square error (since there is relatively no bias). When there are more females than expected in a sample, the simple mean has a lower variance and thus mean square error. The result in Figure 7 are similar to the results concluded in Figures 4-6 based on estimates of the variances.

From conditional analysis results, we conclude that when the number of females in a sample is less than expected, the post-stratified mean produces the lowest MSE values, and is therefore a more accurate measure of the prevalence of problem gambling. When the number of females in a sample is greater than expected, the simple mean produces the lowest MSE values, therefore providing the more accurate measure of the prevalence of problem gambling.

CHAPTER X

LIMITATIONS

The example we considered in this investigation is very simple and extensions are needed to address the context of the gambling study realistically. For instance, we are only considering one factor - gender; however, this provides an example for subjects who will be divided by many groups such as gender, age, race, and ethnicity. In fact, post-stratification is used in setting weights in the gambling survey using 32 domains formed by a cross-classification of 2 gender x 4 age x 4 race/ethnicity groups.

This study assumes complete response. Weights in the gambling study include a non-response adjustment. We assume that subjects will accurately report their true gambling status (problem gambler, or not a problem gambler), and that the answer given is the true answer. We also assume that subjects provide unbiased information. Finally, we are assuming that a person's true state can be observed at a certain point in time.

The sample that will be considered in the gambling study will have an n=10,000 which is much larger than the final sample of n=200 considered in this thesis. We believe conclusions can be extended and generalized to larger populations and samples, though, since our example using n=200 was large enough to produce realistic results. Such results allowed us to drop assumptions made earlier in a sample of 4 or 5 where the sampling fraction was high (80%) and the possibility of producing samples with $n_h \leq 1$ for a given gender was realized. In our

simulation study using n=200, the probability of a sample containing $n_h \leq 1$ for a given gender was less than a z-score of -13.99, which is extremely unlikely and also more realistic.

This paper also only considers one problem gambling rate in its final simulation. Analyzing various crude problem gambling rates and other problem gambling rates between genders may produce different results. Although this study may fall short of accounting for the actual complex problem faced in estimating prevalence in the context of the gambling study, we discuss and define aspects of the real problem.

CHAPTER XI

CONCLUSION

This paper investigated a situation that we believed to be realistic in relation to the gambling study - a population consisting of 50% males and 50% females with 1.5 crude problem gambling rate. Conclusions drawn from this paper can be directly applied to analyses that will be performed on the gambling study survey data results, assuming there is no non-response. For example, one could calculate z-scores as we did, and use Figure 4 to see where that score falls on the plot. This would determine which estimator to use, the simple mean or the post-stratified mean. Further study is needed to evaluate the impact of non-response.

Suggestions for further analysis include running more simulations using a larger population and sample size, changing the crude problem gambling rate and the problem gambling rate between males and females, accounting for non-response, and examining other estimation techniques for the EBLUP. There is limited literature on the EBLUP currently, but exploration into different approaches to estimation and when to use such estimates in the EBLUP could lead to more stable and accurate estimates.

Lastly, the importance of conditional analysis is stressed. Holt and Smith (1979), mentioned previously, concluded that the post-stratified mean performed better when analyzed conditionally. Since the sample results are conditional on the number of males and females in the sample, one could argue that a conditional analysis framework is more appropriate than an unconditional analysis framework

because the former takes into account known measures i.e., the number of males and females in a sample. Using this information and conditioning on it during analyses provides more appropriate estimates since it matches the observed data.

We agree with the conclusions of Holt and Smith (1979). More appropriate inference is made when estimators are examined conditionally and more appropriate conclusions are made in the conditional analysis framework. In the setting studied where large populations and samples are being considered, an unconditional estimate corresponding to the mean does not account for the sample distribution of gender. Figure 4 depicts conditional analysis results for the four estimators. Apart from the BLUP that depends on the known parameters, the results clearly show patterns in a comparison of MSEs that selects the sample mean in some settings, and the post-stratified mean in others. The results in a study will allow the z-score to be evaluated, and hence the most accurate estimator can be selected. Thus, this paper recommends the use of a conditional analysis framework, and estimator choice based on a preliminary evaluation of the z-score for the sample domain.

APPENDIX A

TABLES

Table 16: Distribution of Samples by Males, Females, and Gamblers

#Samples	#Males	#Females	#Male Samples	#Male Gamblers	#Female Samples	#Female Gamblers
792	0	5	792	0	1	0
					35	1
					210	2
					350	3
					175	4
					21	5
3960	1	4	1980	0	20	0
					280	1
					840	2
					700	3
					140	4
			1980	1	20	0
					280	1
					840	2
					700	3
					140	4
6160	2	3	1320	0	60	0
					420	1
					630	2
					210	3
			3520	1	160	0
					1120	1
					1680	2
					560	3
			1320	2	60	0
					420	1
					630	2
					210	3
3696	3	2	264	0	40	0
					140	1
					84	2
			1584	1	240	0

#Samples	#Males	#Females	#Male Samples	#Male Gamblers	#Female Samples	#Female Gamblers
					840	1
					504	2
			1584	2	240	0
					840	1
					504	2
			264	3	40	0
					140	1
					84	2
840	4	1	12	0	6	0
					6	1
			192	1	96	0
					96	1
			432	2	216	0
					216	1
			192	3	96	0
					96	1
			12	4	6	0
					6	1
56	5	0	0	0	56	0
			4	1		
			24	2		
			24	3		
			4	4		
			0	5		

APPENDIX B

EXCEL PROGRAM NAMES

1. MethodsCalculatedSmallEx.xlsx: used to calculate first example with N=5 and n=4
2. SamplesExpanded.xlsx: used to calculate second example with N=20 and n=5
3. varComp.xlsx: used to compare the variance of the simple and post-stratified means in Discussion section

APPENDIX C

SAS PROGRAM NAMES

1. SOsim20.SAS: used to simulate data from population of N=20
2. SOsim1k.SAS: used to simulate data from population of N=1,000
3. CAT1.SAS-CAT18.SAS: used to calculate z-score categories
4. SOgraphs.SAS: used to plot N=1,000 conditionally using z-scores

BIBLIOGRAPHY

Arora, V., and P. Lahiri. "On the Superiority of the Bayesian Method over the BLUP in Small Area Estimation Problems." *SSC Annual Meeting: Proceedings of the Survey Methods Section* (1995): 39-45

Deville, Jean-Claude, and Carl-Erik Sarndal. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87.48 (1992): 376-82.

Deville, Jean-Claud, Carl-Erik Sarndal, and Sautory O. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88.423 (1993): 1013-1020.

Guggemos, Fabien, and Yves Tille. "Penalized Calibration in Survey Sampling: Design-based Estimation Assisted by Mixed Models." *Journal of Statistical Planning and Inference* 140.11 (2010): 3199-212.

Holt, D., and T.M.F Smith. "Post-Stratification." *Journal of the Royal Statistical Society* 142.1 (1979): 33-46.

Kalton, Graham, and Ismael Flores-Cervantes. "Weighting Methods." *Journal of Official Statistics* 19.2 (2003): 81-97.

Kreuter, Frauke, Kristen Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, et al. "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys." *Journal of the Royal Statistical Society* (2010): 389-407.

Lohr, Sharon L. *Sampling: Design and Analysis.* Boston, MA: Brooks/Cole, 2010. Print.

Prasad, N.G.N., and J.N.K. Rao. "The Estimation of the Mean Squared Error of Small-Area Estimators." *Journal of the American Statistical Association*85.409 (1990): 163-71.

Saei, Ayoub, and Ray Chambers. "Empirical Best Linear Unbiased Prediction for Out of Sample Areas." *Southampton Statistical Sciences Research Institute* (2003):

Sarndal, Carl-Erik. "The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation." *Journal of Official Statistics* 27.1 (2011): 1-21.

"Social and Economic Impacts of Gambling in Massachusetts." Massachusetts Gaming Commission, 6 Nov. 2013.