University of California

Los Angeles

# A Study on Conditional Likelihood Estimation for Survey Sampling

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

**Patrick Joseph McCarthy**

2013

## Abstract of the Thesis

# A Study on Conditional Likelihood Estimation for Survey Sampling

by

**Patrick Joseph McCarthy**

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Mark Handcock, Chair

The pursuit of accurate methods for generalizing attributes of a population from a sampled subset is a problem predating the discipline of statistics. Rather than attempting to characterize a population and so assume that the population perfectly represents its own generative process, a superpopulation approach considers the observed population as a sigma algebra of all possible data generated by a process and is focused upon estimating the parameters of the process rather than producing summary statistics.

This study briefly surveys the essentials of survey sampling and evaluates a new superpopulation-based approach put forth by Chaudhuri, Handcock and Rendall (2013), based upon the empirical likelihood of Owen (1989). Using the form of the Hájek estimator and informing it with conditional estimation on empirical likelihood, the approach is shown by simulation study to improve in both accuracy and variance against Hájek's estimator in cases where the values of interest and sampled auxiliary information have little or no correlation, and no improvement over existing methods of estimation otherwise.

The thesis of Patrick Joseph McCarthy is approved.

Susan Cochran

Nicolas Christou

Hongquan Xu

Mark Handcock, Committee Chair

University of California, Los Angeles

2013

To James, Josh and Katie, who added so much to two years and without whom I'd be so much less.

# Table of Contents

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

One of the chief occupations of a statistician is the description of a body of people, products, business indicators or other aspects of one's environment, and indeed the origin of the word "statistics" can be traced to 1748, when it was drived from the German equivalent of the Latin word for "state" dic (2013). Although estimates of center and spread are most accurately reproduced from a complete accounting of the topic of interest, cost and logistics often preclude building an exhaustive data set. The characteristics of a small subset of the data can, however, speak for the body from which it was derived via the practice of survey sampling, and it is by this means that estimates of quantities have been produced for tens or perhaps hundreds of years.

In this study, we examine a novel approach to survey sampling described by Chaudhuri, Handcock and Rendall Chaudhuri et al. (2013). The estimator used in this method attempts to estimate a parameter of a superpopulation by incorporating the value of interest in a sample as well as corresponding available auxiliary data. Although many such approaches exist in the literature, Chaudhuri, Handcock and Rendall incorporate the Empirical Likelihood of Owen Owen (2001) to minimize inaccuracies that may arise from conventional parametric estimation. The new estimation technique will be discussed in the context of existing practice, and finally tested with a simulation study on real data.

# CHAPTER 2

# An Overview of Survey Sampling

The method proposed by Chaudhuri, Handcock and Rendall describes a means for computing estimates from samples of larger population, although it does not dictate the entire procedure from beginning to end. In this chapter the fundamentals of simple random sampling, which provide our baseline, are described, followed by the sampling proportional-to-size technique suggested in the original paper. Additionally, estimators for computing means and variances from samples are described, including that which composes the framework for the CHR estimator.

## 2.1 Definitions and Key Concepts

Let a survey encompass a finite set of $N$ elements called a *finite population* (or simply population) which represents the complete body of interest. It is this population for which we would want to find summary information or parameters, where parameters are functions of the study variable values. A device called a sampling frame is constructed which allows for observation of individual population elements by relating elements of the population with sampling units in the frame, *e.g.* through a sampling algorithm. The sample consists of $N > n$ observed values of the sample units, and it is this sample that is used to construct point estimates of population characteristics of interest, and additionally calculate their precision. In some designs, sampling units contain not only variables of interest but also others, the *auxiliary variables.* Their values may not relate closely with the value being examined, or even relate at all, but through an understanding of their relation to other values in the finite population $U$ they may ease inference and increase precision Särndal & Wretman (1992). This last point will be addressed in detail in a later section.

To assemble a sample, elements are drawn from the finite population at random and we say that the $i$th element of the population $y_i$ is drawn with probability $\pi_i$ (or $p_i$ in the case of simple random sampling), and as the finite population is the set of all units eligible for sampling, $\sum_U \pi_i = 1$ necessarily. We also define $\pi_{ij} = P(y_i \in s)$ and $P(y_j \in s)$, the probability of both $i$ and $j$ being included in the sample together. This is sometimes called a *second-order inclusion probability* or *joint inclusion probability*, whereas $\pi_i$ represents a *first-order probability.* Elementary probability tells us that $P(A \cap B) = P(A) \times P(B)$ if and only if events $A$ and $B$ are independent, and as the dependence relationship comes up frequently in describing estimators we also define $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, or more broadly the covariance matrix of all possible combinations of $i$ and $j$ to be simply $\Delta$. By convention, $tr(\Delta)$ is the $N$-vector $\pi$. A last commonly-used term is the $\pi$-expanded value $y_i/\pi_i$ which we denote $\check{y}_i$. In the case of a dual subscript the check is understood to mean expansion by the second-order probability, as in $\check{\Delta}_{ij} = \Delta_{ij}/\pi_{ij}$.

Consider the example of a dataset describing counties in the United States with demographic information and also voter turnout during the 2004 U.S. Presidential Election Bureau (2009). This finite population consists of all 3,113 counties and county equivalents. For our purposes the total number of votes for the Democratic party candidate cast in all counties is our variable of interest, and auxillary data is composed of figures such as the total population by county and percent of a county's population that voted for a given party. The sampling frame is composed of all of the elements of the finite population (counties) and their values for all relevant variables, as well as some sampling design, that is the vector of first-order probabilities corresponding to the population elements.

Unlike many statistical pursuits such as regression or classical hypothesis testing, the practice of sampling is not interested in discovering the distribution of the data being examined or the relationships of variables, but rather estimates of the population derived from the samples and diagnostic measures on those samples describing their variance and bias. There are different approaches to obtaining these measurements, each with its own caveats.

## 2.2 Simple Random Sampling

The most straightforward method by which one may sample is Simple Random Sampling, in which every unit in the population $U = \{1, \cdots, k, \cdots, N\}$ has an equal probability of selection $p_i$, such that $\sum_U p_i = 1$. A common SRS estimator is that described by Särndal et al. Särndal & Wretman (1992) to estimate the population total is

$$\hat{t}_\pi = N\bar{y}_s = \frac{N}{n}\sum_s y_k,$$

which has an unbiased variance estimator

$$\widehat{Var}(\hat{t}) = N^2 \frac{1 - n/N}{n} S^2_{ys}.$$

Note that to estimate the average county turnout we can simply divide the point and variance estimators by the size of the finite population to get $\hat{t}$, and $\widehat{Var}(\hat{t}) = \frac{1-n/N}{n} S^2_{ys}$. One can also define the true variance of $\hat{t}$ as

$$Var(\hat{t}) = N^2 \frac{1 - n/N}{n} S^2_{yU},$$

however this is rarely useful in an estimation context as it depends upon the sample variance of the finite population rather than of the sample, and had the finite population information been available in the first place sampling would not be necessary. Another useful property of these variance estimators is they are always positive so long as $n < N$. A drawback, however, is these estimators are limited by the variation in the sample data itself. They can produce very accurate estimates if the data itself varies little, however if the data varies a lot the estimator's variance grows proportionately.

Typically, the simplest sampling designs are conducted as sampling-with-replacement (WR), *i.e.* after a unit is selected for the sample it remains in the finite population to potentially be selected again. This presents disadvantages, especially where a finite population has many disparate elements, as the variation of the sample may not reflect the variation of the population. Despite this, several advantages keep SWR in use including unbiasedness, ease of calculation for estimation and variance as well as simple extensions to multilevel designs. Most discussion in this work will not center around SWR but instead its alternative sampling-without-replacement (WOR).

| County | Total Votes Cast | Democrat Votes | Prob. of Selection |
|---|---:|---:|---:|
| Autauga, AL | 20081 | 4758 | 0.00032 |
| Baldwin, AL | 69320 | 15599 | 0.00032 |
| Barbour, AL | 10777 | 4832 | 0.00032 |
| Bibb, AL | 7600 | 2089 | 0.00032 |
| Blount, AL | 21504 | 3938 | 0.00032 |
| ⋮ | ⋮ | ⋮ | ⋮ |

Table 2.1: A sample from the U.S. Counties data frame

Referring back to the U.S. counties example, we show the estimators as computed for 10,000 random samples, each consisting of 40 without-replacement draws with equal probability from the population (Table 2.1, Figure 2.2). In this estimate, the quantity of interest to be found $t$ is the total number of votes cast in all counties, the sample values $y_i$ are the votes cast in all of the counties, and the selection of probability $p$ is the same for all counties, $1/3{,}113 = 0.00032$. As the estimators are unbiased they surround the true values fairly accurately, especially

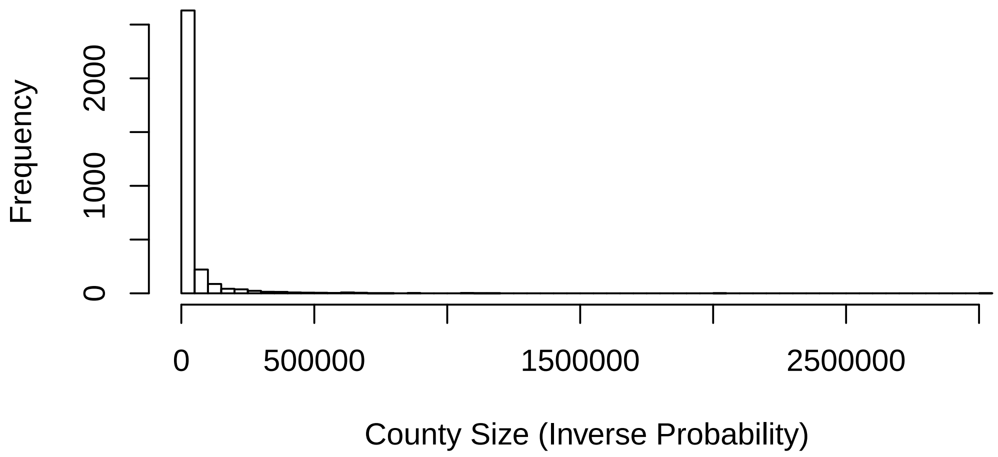


Figure 2.1: Frequencies of County Sizes

considering that the distribution of values has a very strong right skew with many small counties and very few counties of more than one million people (Figure 2.1). Despite this, the large variance of the data comes out (the maximum value of the variable of interest is more than an order of magnitude larger than the 75th percentile, which itself is approximately twice the value of the 25th percentile) and the estimator itself has a heavy skew (The median county has only 10,640 residents and even the 99th percentile is just short of 482,000), making it more difficult to trust the result of any given sample.

## 2.3 Sampling with Probability Proportional to Size

A common way to improve upon the precision of a simple random sample is to bring in additional information. Sampling with probability proportionate to size (PPS) is one way to do this. In PPS, an auxilliary variable is used to denote the "size" of the sample units in question, and rather than sample all units with

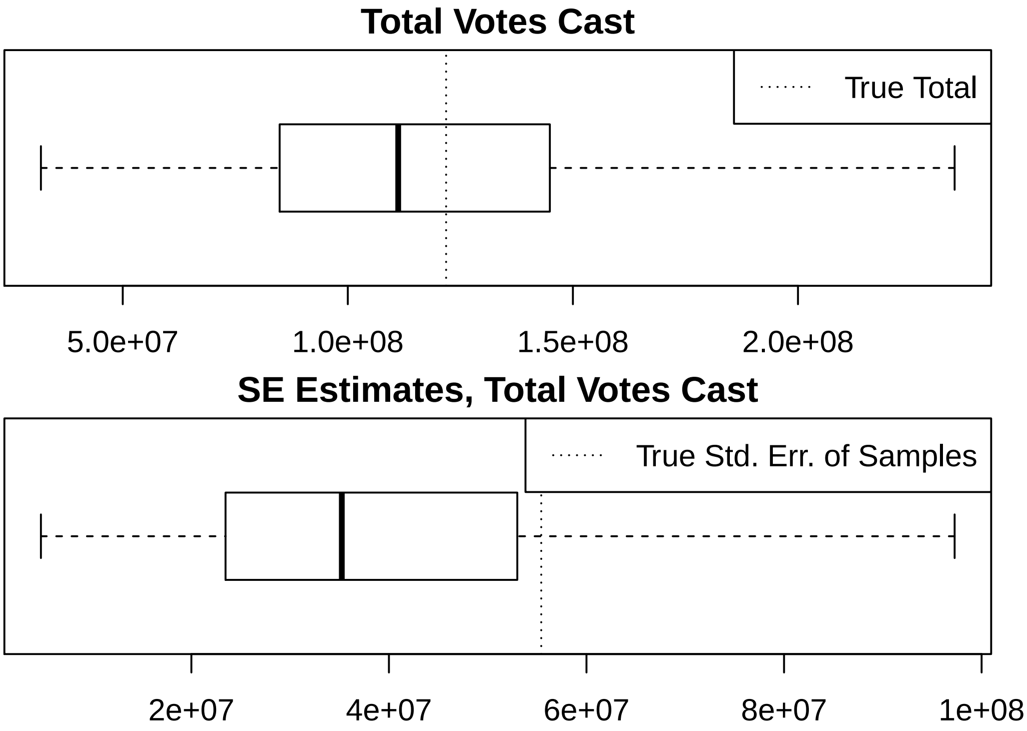


Figure 2.2: The distribution of 10,000 estimates of Total Votes and predicted std. errors. The top panel shows the summary plot of the estimates computed from different samples against the known true value. The bottom plot displays the corresponding estimators for the standard error (Equation 2.3) against their observed standard deviation, the ”true” estimate standard error.

equal probability as in SRS, this size is directly proportional to a probability of inclusion. One then draws samples using these probabilities, and estimation is conducted using formulas that include these probabilities. Typically, a large probability implies that a sampling unit represents a comparably large fraction of the population, and in constructing the estimator its value is weighted accordingly.

The procedure according to Hansen and Hurwitz (1943) is one of the earliest, making use of the relative size of the sample elements in a with-replacement scheme Brewer (1983). An element $i$ was selected with probability $z_i/\sum_i z_i$, where $z_i$ is the measure of size of that element. Despite the advantages of with-replacement methods, this estimator proved less efficient than without-replacment schemes developed afterward Brewer (1983). One such later method commonly used is that of Horvitz and Thompson (1952) which estimates a population total

$$\hat{t}_{HT} = \sum_s \frac{y_i}{\pi_i} = \sum_s \check{y} \tag{2.1}$$

without bias where $\pi_i$ remains the probability of inclusion of the $i$th element in the sample Horvitz & Thompson (1952). Its variance estimator

$$Var(\hat{t}_{HT}) = \sum\sum_U \Delta_{ij}\check{y}_i\check{y}_j \tag{2.2}$$

is approximated by

$$\widehat{Var}(\hat{t}_{HT}) = \sum\sum_s \check{\Delta}_{ij}\check{y}_i\check{y}_j, \tag{2.3}$$

which is unbiased when all $\pi_{ij} > 0$ Särndal & Wretman (1992). Among other attractive properties, the HT estimator is especially interesting for our purposes because the auxiliary information offered by $\pi_i$ allows us to reduce the variance. For example, if there was a proportionality constant $c$ such that $c\pi_i = y_i$ for all $i$ then all terms $y_i/\pi_i$ would reduce to $c$ and the variance of the estimator would become 0. Therefore, in cases where the probability of inclusion and the value of interest are closely correlated the HT estimator provides excellent estimation.

Correspondingly, this relationship has little benefit when the variable of interest and the weight are uncorrelated. In this case, one may make use of an estimator sometimes attributed to Hájek Särndal & Wretman (1992)

$$\tilde{t}_{Hj} = N\tilde{y} = N\frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k}. \tag{2.4}$$

Särndal *et al.* give an approximate (Taylor-linearized) variance instead of a true variance for computational ease Särndal & Wretman (1992),

$$Var(N\tilde{t}_{Hj}) = N^2 AVar(\tilde{y}) = \sum\sum_U \Delta_{ij}\left(\frac{y_i - \bar{y}_U}{\pi_i}\right)\left(\frac{y_j - \bar{y}_U}{\pi_j}\right) \tag{2.5}$$

with variance estimator

$$\widehat{Var}(\tilde{t}_{Hj}) = N^2 AVar(\tilde{y}) = \sum\sum_s \check{\Delta}_{ij}\left(\frac{y_i - \tilde{y}_s}{\pi_i}\right)\left(\frac{y_j - \tilde{y}_s}{\pi_j}\right) \tag{2.6}$$

The Hájek estimator is intended to measure means rather than totals, and so to compute totals with them the formulas as given contain the coefficients $N$ for the estimator and $N^2$ for variance. (This is the opposite of the adjustment made to the HT estimator to obtain means.) Of particular note is the term $\sum 1/\pi$ in the estimator. Whereas the HT is analogous to a sum, Hájek's is a weighted average, with $\sum 1/\pi$ acting as a stand-in for population size.

Returning to the example of the 2004 election, let the size of a county be defined by the total number of votes cast by that county, and let the probability of selection be defined by the reciprocal of that number. Consider a series of samples drawn from the population using a PPS-friendly sampling algorithm (see section 5.1) with these probabilities, each consisting of 40 of the 3,113 U.S. Counties. The 2004 election was contentious in many counties and so a larger voter turnout in the population often indicated a larger turnout among Democratic voters. In this dataset they are in fact highly correlated, with correlation coefficient $\rho = 0.978$. For comparison, a variable uncorrelated with size was generated for each member of the population as a random Gaussian with mean 0, actual correlation $\rho = -0.02$.

Figures  2.3 and  2.4 show the results of the estimators for the collection of samples and their standard errors.  As expected, HT is a clear winner in the correlated case with a median estimate closer to the true population value and a smaller variance. Though Hájek's estimator isn't terrible, the diversity of the population units and non-centrality of the population mean produce a variance much greater than with HT. By contrast, in the uncorrelated case Hájek comes out ahead albeit by a smaller margin. This illustrates the general case that Horvitz-Thompson performs well when absolute correlation is far from zero, and Hájek's performs well when correlation is closer to zero.

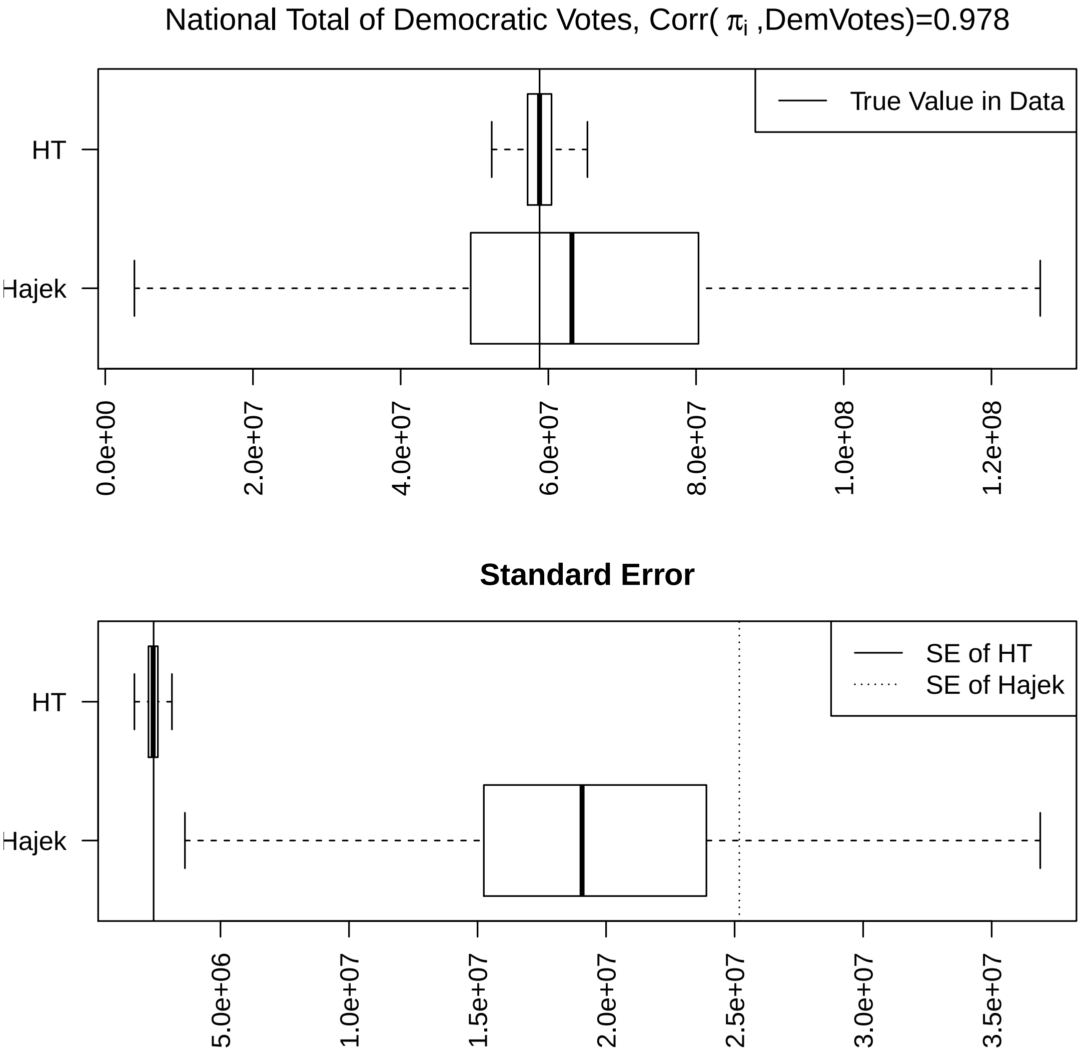


Figure 2.3: PPS Estimation of Total, Closely Correlated to Size

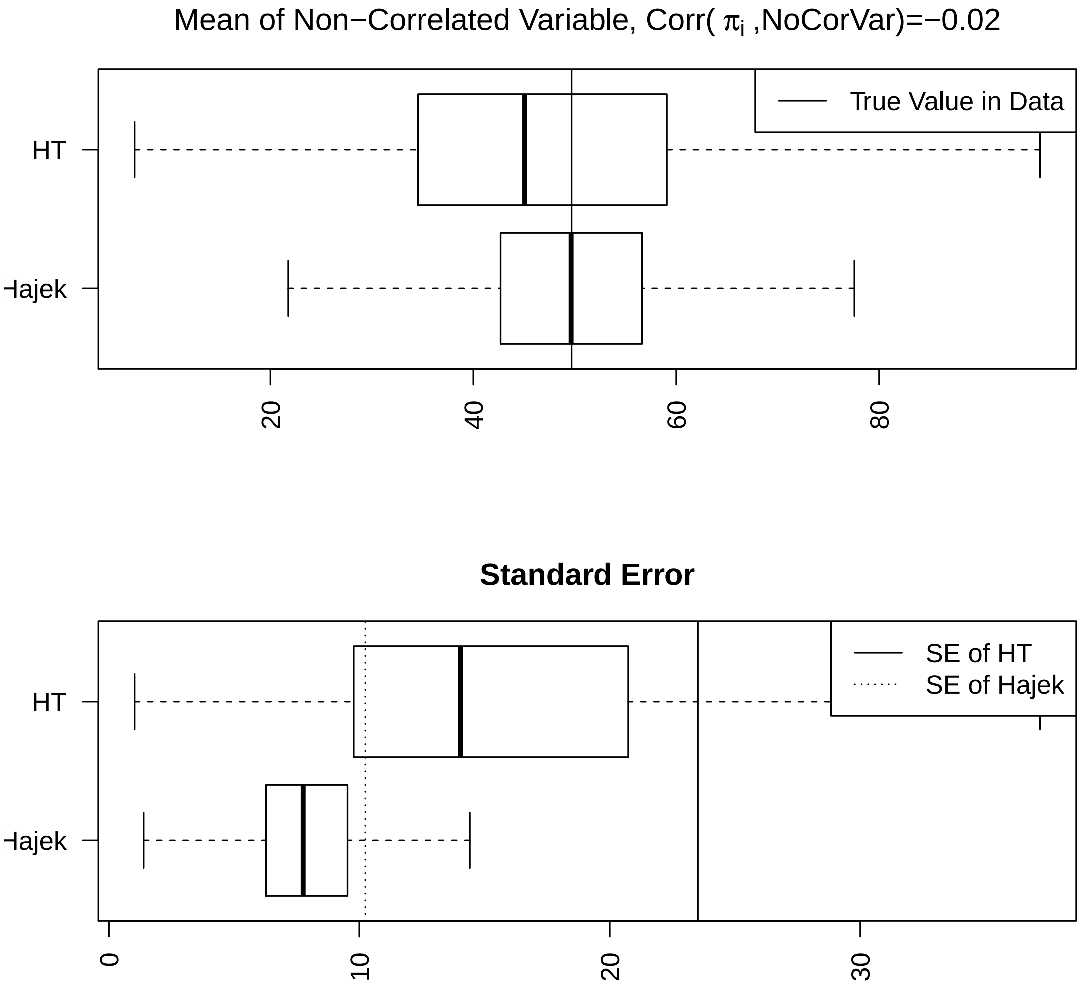


Figure 2.4: PPS Estimation of Mean, Not Correlated to Size

The additional stability provided by auxiliary information comes at the cost of additional assumptions. To maintain an unbiased estimation for variance $\pi_{ij} > 0$ necessarily for all $i, j$, and to ensure that the variance does not produce a negative value the Sen-Yates-Grundy (SYG) condition that $\pi_{ij} - \pi_i\pi_j < 0$ for all $i \neq j$ must always be met, requiring careful selection of a compatible sampling algorithm Särndal & Wretman (1992). Finally, as most PPS schemes are sampled without replacement, there must be a means by which to update the selection probabilities to new values conditioned on the sampled elements being removed from consideration for future draws. This can be especially cumbersome for large $N$.

All of the estimators discussed to this point are *design-based*. As the sample size approaches the population size, the formulas will produce increasingly accurate representations of the population, and where $N = n$ they will give the desired for the population of interest exactly. An alternative, the superpopluation approach, is discussed in a later section.

## 2.4 Multistage Designs

When a population of interest exhibits high variability it can be difficult to compute estimates with desirably small precision. To counter this, the finite population can be divided into smaller, more homogenous groups called Primary Sampling Units (PSUs) which are themselves sampled, and then the contents of the PSU are either enumerated exhaustively, sampled, or divided further into Secondary or even Tertiary Sampling Units which themselves can be enumerated or sampled Lumley (2011).

A common type of multistage design is cluster sampling, in which PSUs (and optionally sampling units within them) are sampled, after which the lowest-level sampling unit is enumerated exhaustively Lumley (2011). This is in effect what our election example has accomplished, in that samples of 40 counties are selected, and every vote within them is tallied. (By contrast, an exit poll of some percentage of a county's voters would not be considered a true cluster design). Although not the focus of this study, multistage sampling (and designs that use it, termed 'complex designs') is a primary application of PPS designs in general and the Chaudhuri-Handcock-Rendall estimator in particular.

# CHAPTER 3

## An Overview of Empirical Likelihood

As an analyst works to understand data, a commonly applied tool is parametric estimation. The data are assumed to come from a common generative process, in which each data point represents a random variable following a common distribution $F(\theta)$, for unknown parameter $\theta$. To assign the parameter a value, the probabilities of observing the values of each observation are computed and multiplied together to produce the probability of observing the sample. This product is a likelihood function such as (3.1), which is then optimized with regards to $\theta$:

$$L(\theta) = \prod_{i=1}^{n} f(x_1;\theta) \cdot f(x_2;\theta) \dots f(x_n;\theta). \tag{3.1}$$

This is a useful approach when the form of $f(\cdot)$ is known, and especially so when the computation of the likelihood or log-likelihood produces an easily-managed expression. When the form is not known independently of the data or the computations are difficult, however, it can be unclear how to proceed. A solution to this problem can be found in *empirical likelihood*.

Rather than attempting to maximize the functional form of a distribution, Empirical Likelihood (EL) treats the shape itself as a nuisance parameter to be profiled out in the pursuit of a desired statistic Pawitan (2001). Rather than defining a distribution as a functional form *a priori* and using that form to produce statistics, EL does not define the distribution but rather provides a basis for comparision and computing statistics in terms of the observed data, as well as producing a non-parametric likelihood function for which the probability of observing the present data can be optimized Owen (2001).

At its most general, the empirical likelihood is defined as

$$\sup_{\mathcal{F}_\theta} \prod_{i=1}^{n} p_i \tag{3.2}$$

where $\mathcal{F}_\theta$ is the family of all possible discrete functions over the observed values $x_1, \cdots, x_n$, and the variable $p_i$ refers to the probability of observing the value $x_i$ in the $i$th observation. This definition is a limiting property on the infinite or non-parametric model of the probability space as we only have $n$ observations and so can only produce an approximation. As $n$ approaches infinity the true shape of the infinite model emerges. Owen (2001) demonstrated that this supremum is definable as a discrete distribution

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}, \ -\infty < x < \infty \tag{3.3}$$

for a sample composed of random draws $X_1, X_2, \cdots X_n$ assumed to all be distributed according to the same CDF $F_0$ Owen (2001). This distribution is therefore the non-parametric maximum likelihood estimate (NPMLE). Correspondingly, the optimum nonparametric likelihood for the same observations can be described as

$$L(F) = \prod_{i=1}^{n} (F(X_i) - F(X_i-)), \tag{3.4}$$

where $F(x-)$ is understood to be $P(X < x)$. The expression $F(X_i) - F(X_i-)$ can thus be considered to take the place of $f(x_i;\theta)$ in (3.1) above as the "amount of density" occuring at that point.

It should be noted that this formulation depends upon the quantity $F(X_i) - F(X_i-)$ being nonzero everywhere in the domain of interest in order to produce a positive likelihood, a requirement which suggests the method is most capable for large samples which can be expected to sample every value in the support. Alternatively, Owen suggests limiting the family of possible functions $\mathcal{F}$ to the subset of $\mathcal{F}$ for which the support intersects the observed values.

Although taking a non-parametric approach suggests a lack of interest or even of belief in a governing parameter $\theta$, in fact the existence of a parameter is not prohibited or absent insomuch as it is a matter of perspective and derivation. In the standard parametric mode a parameter $\eta$ is sought through a function $\theta(\eta)$, and optimizing $\hat{\eta} = \eta_{MLE}$ produces the MLE $\hat{\theta} = \theta(\hat{\eta})$. In the non-parametric case, when we assume a true CDF $F_0$ and a relationship $\theta_0 = T(F_0)$ (in which we take $\theta = T(F)$ for a function $T$ of distributions), and so replacing $F$ with the optimal value $F_n$ the ECDF produces an empirical $\hat{\theta}$. In this case the functional $T$ is specifically one which produces a statistic (*e.g.* the mean) from a distribution, such as in the case $\hat{\theta} = \sum_{i=1}^{n} x_i dF(x_i)$. Replacing likewise $dF$ for $F_n$, $\hat{\theta}$ is the estimate of the MLE of the mean.

Calculating the empirical likelihood function is simplest when it can be directly computed as a product of each individual observation's probability, however when the observations contain ties (*i.e.* $X_i = X_j$ when $i \neq j$) the values are no longer distinct and the form must be altered. In order to preserve the convenient form of (3.2) , Owen recommends that one assign a weight $w_i$ to each observation $X_i$ such that $\sum w_i = p_j$, where $p_j = P(X_i = z_j)$ for some $z_j$ in the support. The likelihood can then be computed as

$$L(F) = \prod_{i=1}^{n} w_i. \tag{3.5}$$

Absent other constraints, this likelihood has been found to be maximized at $w_i = \frac{p_{j(i)}}{n_{j(i)}}$, those being the $p$ and $n$ in which the $i$th observation falls. Additionally, as the sum of the weights is necessarily between 0 and 1, one can describe the weights as constrained by an ($n$-1)-dimensional simplex

$$\mathcal{S}_{n-1} = \left\{ (w_1, w_2, \cdots w_n) | w_i \geq 0, \sum_{i=1}^{n} w_i = 1 \right\}. \tag{3.6}$$

The simplex is effectively a generalization of a triangle, in which the line between the triangle's centroid and a given point is a scale corresponding to the value of a single $w_i$ at a point, one $w_i = 1$ and the rest are 0, and every point along the edge of the shape represents a linear combination of weights totalling 1, whereas every point within the shape represents a linear combination of weights totalling less than one. Because on the edge the total is in fact 1, a degree of freedom is lost ($w_i = 1 - \sum_{j \neq i} w_j$) and so the simplex is denoted as ($n$-1)-dimensional. Although setting all weights to be equal produces an optimal likelihood, the value of the simplex is that it allows for optimization of the weights under additional constraints.

To contrast the parametric and non-parametric approaches, consider the example of the binomial distribution with the likelihood/probability mass function

$$\mathcal{L}(p|n,k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{3.7}$$

with unknown $p$, $n = 2$ and a series of observed $k = \{0, 1, 1, 2\}$. To estimate the mean via parametric likelihood, first one can recast the number of successes in each trial as an observation from a multinomial with log-likelihood

$$l(p) = \sum_{k=0}^{n} n_k \log(p_k)$$

where $n_k$ is the number of observations of value $k$, which can be solved for an optimal $\hat{p}$

$$\begin{aligned}\hat{p} &= \frac{\sum_k k n_k}{nN} \\ &= \frac{0 \cdot 1 + 1 \cdot 2 + 2 \cdot 1}{2 \cdot 4} \\ &= 1/2\end{aligned}$$

where $N = \sum_k n_k$, and finally use the formula for the binomial mean to obtain $n\hat{p} = 2 \cdot 1/2 = 1$.

By contrast, the EL approach is much simpler. The empirical likelihood is calculated as

$$L(F) = \prod_{i=1}^{n} w_i$$
$$= \left(\frac{1}{4}\right)^4$$

This doesn't immediately produce a mean, but with a functional $T = \sum_k p(k)k = \sum_k kw_k$ the mean can be produced as

$$0 \cdot 1/4 + 2(1 \cdot (1/4)) + 2 \cdot (1/4) = 1$$

It is important that the scope of these toy experiments be understood, as with a cursory reading appears little different than computing a sample mean. In the parametric case, by finding $\hat{p}$ the data are used to estimate the properties of the distribution from which both the data came, and which future data can be expected to follow. The EL case does produce a sample mean, but with the intent that future samples can be expected to follow the same distribution. One caveat is that a functional form is never suggested for the distribution $F$, and in fact where $w_i \neq p_i$ for any $i$, $F$ cannot even be asserted to be unique.

# CHAPTER 4

# The Composite Likelihood Estimator

Many pursuits in statistics are performed with the intent of discerning information about the true nature of things from data and assumptions about its source. Survey sampling is most commonly concerned with estimating means, populations, and other such descriptive figures while other endeavors such as regression attempt to discern relationships, or at least reasonable approximations thereof. The compound likelihood approach seeks to use sampling to describe a relationship, and so marry these two goals.

## 4.1 Definition of Terms

Chaudhuri, Handock and Rendall Chaudhuri et al. (2013) seek to understand the generative process of something, described by a response variable $Y$. It is assumed this can be explained by a parametric model or process, for example the simple regression model $\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\epsilon}$. Given a complete set of population data, one could simply perform the necessary calculations to determine $\boldsymbol{\beta}$, however by shifting the focus from the current population to those generated under different circumstatnces of time, location, etc. it immediately becomes clear that the finite population at hand is not the ”true” population of the study as an exhaustive analysis no longer offers a compete description. The understanding of the finite population as a realization of population random variables and the use of those to study their generating process is called a *superpopulation model* Särndal & Wretman (1992). Importantly however, the finite population is not often available in its entirety, and so this investigation must be conducted using sampling techniques. The approach often taken is a two-step approach described by Särndal *et al.* and used by Chaudhuri, Handcock and Rendall (2008), in which the parameter of interest ($\boldsymbol{\beta}$ in the regression above) is estimatated as a function of the sample data, and then that $\hat{\boldsymbol{\beta}}$ is used with probability information to estimate the true superpopulation parameter $\boldsymbol{\beta}$ Chaudhuri et al. (2008). The superpopulation approach should be constrasted with the design-based approach of the other estimators discussed. Whereas the design-based estimators get increasingly close to producing the desired value as the sample size approaches the size of the population, the superpopulation by definition attempts to specify the process from which the population is defined and as a result this goal is not realized by perfectly specifying the population at hand; it is one of many.

Let us generally refer to the superpopulation parameter of interest as $\theta$. Chaudhuri, Handcock and Rendall (hereafter CHR) estimate $\theta$ by generating a conditional likelihood $L_{CE}$ and maximizing, but first much groundwork must be laid. Consider a superpopulation model composed of $Y$ the response variable, a set $X$ of auxillary data which may or may not have a direct relationship with $Y$, and Z a matrix of design variables, and let $V = X \cup Y \cup Z$. Let $A \subseteq V$ be a collection of all explanatory variables. Additionally, let the finite population $\mathcal{P}$ consist of $N$ i.i.d. draws from model, and let a random subset $\mathcal{S}$ of $\mathcal{P}$ be the set of samples of size $n$. We will discuss $S$, an instance of such a sample.

As discussed above, the nature of a sample depends upon the means by which it was procured, the sampling design (here denoted $D$). For convenience, consider also a membership vector $I_S$ of length $N$, with sum $n$, for which the $i$th element is 1 when element $i$ of the population is included in $S$ and 0 otherwise. For our purposes $I_S$ can be read as "the sample $S$ being selected from the population $\mathcal{P}$". Finally, $\pi$ will denote the probability of inclusion— $\pi_S$ is the probability that the sample $S$ was selected, $\pi_i$ is the probability that unit $i$ was included in $S$, etc. Crucially, as the population consists of i.i.d random variables from the superpopulation (and therefore the population is not fixed), $\pi$ is a random variable.

## 4.2 Assumptions and Composite Likelihood

CHR rely upon several assumptions. First *conditional independence given the design* asserts that for all $S \subseteq \mathcal{P}$

$$\pi_S \perp (Y_{\mathcal{P}}, X_{\mathcal{P}})|D_{\mathcal{P}}, \forall S \subseteq \mathcal{P}. \tag{4.1}$$

Under this assumption, $E_{\mathcal{P}}[\pi_S|Y_{\mathcal{P}}, X_{\mathcal{P}}, D_{\mathcal{P}}] = E_{\mathcal{P}}[\pi_S|D_{\mathcal{P}}] = \pi[S, D_{\mathcal{P}}]$. Second CHR assume *conditional independence given the sampling probabilities*, in which the sample selected is dependent only upon the probability of selection, that is

$$I_S \perp (X_{\mathcal{P}}, Y_{\mathcal{P}}, D_{\mathcal{P}})|\pi_S, \forall S \subseteq \mathcal{P}. \tag{4.2}$$

This frees us to consider the sample (and therfore its contents) as solely influenced by one source, a natural conclusion which does much to ease computation. This enables the adoption of a principle from Pfeffermann and Sverchkov (2003), that

$$Pr_{\mathcal{P}}[I_s = 1, V_S] = E_{\mathcal{P}}[\pi_S|V_S]Pr_{\mathcal{P}}[V_S] \tag{4.3}$$

or in other words that the joint probability of drawing a sample from the population with the data in those samples being what it is, can be expressed as the population density of the values of $V$ in the sample $V_S$ multiplied by the inclusion probability of the sample given $V_S$ Pfeffermann & Sverchkov (2003). This construction will allow us to define a density of interest via Bayes' Rule.

Should the assumptions not be met the estimation can go awry. In the case of conditional independence given the design, if $\pi_S$ is not independent of $X_{\mathcal{P}}$ and $Y_{\mathcal{P}}$ then it isn't enough to build assumptions about our sample from the selection probabilities of the elements alone, and the EL weights put upon the observations will be misspecified. Fulfilling this assumption allows us to directly translate $P(X_i = x)$ into the probability of selection $\pi_i$ and leverage the elegant EL form. In the election example, this assumption may fail if in some counties with low Democrat turnout the county was not listed in the election totals. The probability of selecting the county, *i.e.* the proportion of all national votes cast in that county, would be linked to the auxiliary variable Democrat turnout. In the data as used, the selection of probability is always included, non-zero, and determined solely by the county's proportion of the national total of votes.

Likewise if conditional independence given the sampling probabilities is violated, then samples cannot be considered upon equal footing but will require investigation into the auxiliary variables that correspond with each new sample. This assumption is known to be met because the sampling algorithm used takes only the probabilities of selection as inputs. If there was a quota of some kind requiring an equal proportion of Democrat- and Republican-leaning states then the inclusion vector $I_S$ would rely upon both $\pi_i$ and $X_i$ for a county.

It is given that the $i$th element in $S$ was drawn from the finite population with probability $\pi_i$. Let $F_0$ be the distribution of $V_{\mathcal{P}}$, presently defined to a parameter. Taking into account the above expression, we can then define $F_S^{(i)}$ to be the conditional distribution of the values of sampled elements $V_i$ given inclusion of unit $i$ in the sample, with density $dF_S^{(i)}$. We can then reformulate the above to produce

$$dF_S^{(i)} = \frac{Pr_{\mathcal{P}}(I_{\{i\}} = 1, V_i)}{Pr_{\mathcal{P}}(I_{\{i\}} = 1)} = \frac{E_{\mathcal{P}}[\pi_i | V_i] dF_0(V_i)}{Pr_{\mathcal{P}}(I_{\{i\}} = 1)} \tag{4.4}$$

where the normalizing constant $Pr_{\mathcal{P}}(I_{\{i\}} = 1)$ can also be expressed as

$$\int E_P[\pi_i | V_i] dF_0(V_i) dV_i \tag{4.5}$$

via (4.3). The density defined in (4.4) is a central quantity in this study. By combining the density of $F_0$ and all the information of selection encapsulated in

$E_{\mathcal{P}}[\pi_i|V_i]$ according to (4.2), a likelihood based upon this density can be maximized to estimate the superpopulation parameter $\theta$. Before defining the likelihood, however, the issue the availability of population information must be addressed.

Breaking down (4.4) and rearranging, it can be understood that the product of the sample density $dF_S^{(i)}$ and the probability of an observation's selection are together equivalent to the density of the finite population times the population expectation of sample probabilities, given the data matrix for the observation. This allows the conditional distribution of the sample data to be redefined in terms of both the finite population and also the indicator vector for inclusion in the sample. This is advantageous as it allows for the density to be defined while taking into account the probability of observation/selection through the $E_{\mathcal{P}}$ term, however at the same time it is problematic— the intent of a survey is conducted to gain insights without knowledge of the population itself. With this in mind, requiring population-level understanding is clearly counterproductive.

To circumvent this, we can redefine the necessary expectations in terms of population densities that are "masked" or hidden somehow. The expectation $E_{\mathcal{P}}[\pi_i|V_i] = [E_S(\pi_i^{-1}|V_i)]^{-1}$ will be defined as the *conditional visibility* $\nu_i$, following Patil and Rao's (1978) interpretation of sampling as enumerating a partially hidden population (Patil & Rao, 1978; Pfeffermann & Sverchkov, 1999). We define also $\Upsilon$ as $E_{\mathcal{P}}[\pi_i] = E_{F_0(V_t)}[\nu_i]$, a normalizer here called the visibility factor. This last term likewise follows from Patil and Rao, and together they can be imagined to form a pdf $f^w(x) = \frac{f^0(x)w(x)}{\Upsilon}$ where $w(x)$ is a weight applied to every value $x$ depending upon whether it was observed, and $\Upsilon$ normalizing the density to have an integral of 1. By substituting with prior expressions, we obtain

$$dF_S^{(i)} = \frac{\nu_i dF_0(V_i)}{\Upsilon_i} \tag{4.6}$$

Finally, a likelihood can be produced using the observed values of the sample. Invoking the assumption of $F_0$ being understood up to a parameter $\theta$ and modeling

$\nu_i$ with a parameter $\alpha$, a composite likelihood using all $V_i$, $i = 1, 2, \cdots n$ can be constructed as

$$L_{CE}(V, \alpha, \theta) = \prod_{i=1}^{n} dF_{\mathcal{S}}^{(i)}. \tag{4.7}$$

With this, $F_0$ can be optimized by finding the profile likelihood of $\theta$. The computation of $\Upsilon$ and even the optimization of $\theta$ itself can be complex however, and so the use of other methods to find and optimize this term are desirable.

## 4.3 Reinterpretation in Terms of Empirical Likelihood

By casting the problem in terms of conditional visibility, a conceptual link appears between the current formulation of a sample density and the empirical likelihood approach of Owen Owen (2001). Within the superpopulation model, we assume $F_0$ is specified by a parametric family $\mathcal{F}_\theta$. Suppose that for each $F \in \mathcal{F}$ we define the weights $F$ assigns on $V_i$ to be $w_i = F(\{V_i\})$, with a zero weight assigned to continuous functions of $F$. Additionally, the computations can be simpilifed further with another assumption which we call *label independence of visibility factors.* Restating (4.1) it can be asserted $\Upsilon_{i=E_{\mathcal{P}}[\pi(i,D_{\mathcal{P}})]}$, that the visibility factor depends upon the design variables rather than any value composing the sample, and so therefore all members of the population share the same visibility factor.

Approximating this factor by $\hat{\Upsilon} = \sum_{i=1}^{n} \nu_i w_i$, and substituting $w_i$ for $dF_0 = F_0(\{V_i\})$ we can easily recreate (4.7) as

$$L_{CE}(w, Z, \nu) = \prod_{i=1}^{n} \frac{w_i}{\sum_{i=1}^{n} \nu_i w_i} \tag{4.8}$$

which conveniently shares the form of the Hájek estimator. $\Upsilon$ must be treated carefully, as incorrect specification (e.g. as $\pi$,$\nu$) will not weight the parameters properly. As $\nu$ comes from $E_{\mathcal{P}}$ which itself is a maximization of $\alpha$, this term should be estimated separately from $w$. $\nu$ and $Z$ follow from the sample data, leaving us only to optimize for the vector $w_i$. Additionally, with the assumption that $Y$ and $A$ share a relationship via some function $\psi_\theta$

$$E_{F^\theta}[\psi_\theta(Y, A)] = 0 \tag{4.9}$$

the weights can be optimized further by employing this constraint in the n-dimensional simplex

$$\mathcal{W}_\theta = \left\{ w \in \Delta_{n-1} : \sum_{i=1}^{n} w_i \psi_\theta(Y_i, A_i) = 0 \right\} \forall \theta \in \Theta \tag{4.10}$$

(Population parameters understood to be known and fixed can also be incorporated in such a manner, however are beyond the scope of the present study.) A constrained estimator for $\theta$ can be produced according to Qin & Lawless (1994) Chaudhuri et al. (2008) as

$$\hat{\theta}_{CE} = \arg\max_{\theta \in \Theta} \left\{ \max_{w \in \mathcal{W}_\theta} L_{CE}(w, Z, \nu) \right\} \tag{4.11}$$

# CHAPTER 5

# Examination of the Estimator

We here empirically demonstrate the properties of the Chaudhuri-Handcock-Rendall (CHR) estimator. A simulation study was conducted in which samples were drawn from the election dataset used in prior examples, and the estimator and its variability are compared both to theoretical expectations of their performance and their true performance. Differing results arise across several values of interest, depending upon the relationship of that value to the probability of selection.

## 5.1 Generating Samples

To observe the properties of the estimator first-hand, many samples are drawn from a population with known characteristics, and then from each estimators are computed which are then compared to the known "truth" of the population. These estimators, while making use of the conditional empirical likelihood approach, are based upon the form of the Hájek estimators of mean and variance described in equations 2.4 and 2.6 and have their characteristic requirements. One, the Sen-Yates-Grundy criterion that $\pi_{ij} \neq \pi_i \pi_j$, is required to guarantee non-negative values for estimated variance Sen (1953) Yates & Grundy (1953). Consider again equation 2.6

$$\widehat{Var}(\tilde{t}_{Hj}) = N^2 AVar(\tilde{y}) = \sum \sum_s \check{\Delta}_{ij} \left( \frac{y_i - \tilde{y}_s}{\pi_i} \right) \left( \frac{y_j - \tilde{y}_s}{\pi_j} \right).$$

If $\pi_{ij} = \pi_i \pi_j$ then $\check{\Delta}_{ij} = 0$ and the variance expression collapses to 0, suggesting an unrealistically perfect estimator. Generating samples without this characteristic is non-trivial, requiring a special algorithm; in section 5.2 great care is taken also to confirm that these properties are as expected.

The study sample consists of 2.24 million samples of 40 counties each, drawn from the 2004 election dataset used previously Commission (2006). The samples were drawn proportionately to county size which we define here as the total number of votes cast in that county across all presidential candidates.

Sampling with unequal probability while preserving the Sen-Yates-Grundy condition presents an additional challenge over simple sampling without replacement, as without due care the order in which units are sampled can have unforseen relationships upon the sampling probabilities. For instance, given a vector of sampling probabilities {.2,.2,.2,.4} a draw of the first element results in the renormalization of the remaining probabilities as {.25,.25,.5} where as a draw of the last results in probabilities of {.33,.33,.33}. In this instance, it is apparent that not only is recomputation necessary, but that the probability of inclusion of the next sample is affected.

Tillé presents a method in which the probabilities are easy to compute and do not depend upon draw order Tillé (1996). Consider a population $U$ from which $n$ units are to be sampled. First, an initial probability vector of values $\pi(i|k)$ is computed, representing for each $i$ the probability that the $i$th unit is selected given that it is being selected from a sample of size $k$. These values are proportional to the positive values $x_i$, $(i \in U)$ of some auxiliary variable $x$ and are computed as

$$\pi(i|k) = \frac{kx_i}{\sum_{i \in U} x_i}(i \in U).$$

If $\pi(i|k) \geq 1$, then set $\pi(i|k) = 1$ and repeat the procedure until all $\pi(i|k) \in [0, 1]$. After generating this initial vector, the first selection step is conducted which (because it's counted backward) is termed step $k = (N - 1)$. A unit is selected from $U$ with probability $1 - \pi(i|N - 1)$.

Each of the subsequent steps is subtly different from the first. At the beginning of step $k$, the sample is composed of $k+1$ units. The vector of probaiblities is recalculated as above with the reamining unselected units, and a unit is selected from the sample with probability

$$r_{ki} = 1 - \frac{\pi(i|k)}{\pi(i|k+1)}.$$

The selected unit is discarded. After selection only $k$ units remain in the sample, and the procedure stops at the end of step $n$, leaving behind the $n$ undiscarded units which then comprise our sample.

This algorithm, published by Tillé in the `sampling` package, was ported to C++ and run upon 30 Apple iMacs to produce 2.24 million samples for analysis in subsequent sections Tillé & Matei (2012).

## 5.2 Sen-Yates-Grundy Assumption

The CHR estimator dictates that the Sen-Yates-Grundy (SYG) condition be met, that is that the probability of two elements being included in the same sample is distinct from the product of two first order inclusion probabilities, $\pi_{ij} \neq \pi_i \pi_j$. Having gone to much trouble to obtain and execute a Tillé sampling procedure to achieve this purpose, it is important that we test this assertion.

At issue are three distinct matrices. The first, theoretical matrix is that specified by Tillé's sampling method, and is employed in the calculation of the HT, Hájek and CHR estimators. A second matrix is that observed by direct computation of the occurrence of pairs of samples, *i.e.* $\pi_{ij}$ is the number of times that unit $i$ and unit $j$ are sampled together divided by the number of possible opportunities. The third matrix is the outer product of the vector of first order selection probabilities $\boldsymbol{\pi}$ with itself. The theoretical matrix and the observed matrix will be compared to ensure that our sampling procedure produces a result of the expected form, and provided that it does the theoretical matrix will then be compared with the outer product to test the SYG condition.

Before beginning the analysis however, an issue of completeness must be addressed. The theoretical joint inclusion matrix provides a non-zero probability for all off-diagonal elements of the matrix, which is to say that any element can potentially be sampled with any other element with some probability. (Technically the diagonal should be empty as after drawing element $i$ without replacement $i$ cannot be selected again, however by convention we set the diagonal to the first order probabilities $\boldsymbol{\pi}$.) By contrast, as a result of some very small but non-zero joint probabilities the empirical joint inclusion matrix contains zeroes in approximately 2% of its cells even after millions of samples are drawn. As this will both force EL estimates to be zero as well as dilute many distance metrics, a workaround was implemented in the form of a general additive model Hastie & Tibshirani (1986).

The model is a Poisson regression (with form $E(Y|x) = e^{\theta x}$) taking the upper triangluar values $\pi_{ij}$ of the observed joint inclusion matrix from the simulation as the response. These elements of the theoretical matrix are predicted by a smoother function on the combination of $\pi_i$ and $\pi_j$ and offsets. The resultant model has 29 terms and an estimated degree of freedom of 28.999, but is not sufficiently insightful itself to merit reproduction. When the predictor $\log(\pi)$ is chosen the smoothing function in the GAM, borrowing strength, fits a response surface which contains no zeroes, but rather "smooths over" the observed zeros giving a close approximation of the expected observed values. This predicted fit will be the empirical matrix examined. The choice of Poisson regression is worth noting. Were the probabilities modeled simply as Gaussian around a mean rather than as a count variable, peculiar effects could arise such as expectations of $\pi_i$ close to, at, or even less than zero. As a Poisson however, the value is fixed as greater or equal to zero.

One of the most fundamental approaches to comparing discrete probability distributions is divergence measures. One of the most common, the Kullback-Leibler divergence, is the expectation of the sum of the log difference between like probabilitites belonging to distributions P and Q, defined as

$$D(P||Q) = \sum \log(P_i/Q_i)P_i$$

Though notably not a metric (as $D(P||Q) \neq D(Q||P)$) the difference of the measure from zero is a good indicator of deviance of distribution.

| | Theoretical $\pi_{ij}$ | Empirical $\pi_{ij}$ | $\pi_i\pi_j$ |
|---|---|---|---|
| Theoretical $\pi_{ij}$ | 0.0000 | 0.0024 | 89.6770 |
| Empirical $\pi_{ij}$ | 0.0024 | 0.0000 | 89.6763 |

Table 5.1: KL Divergence, Rows are Reference Matrix

Table 5.1 demonstrates that the theoretical and empirical matrices match each other quite closely, with a divergence across the $3113^2$ cells of less than 1/100 indicating that our theoretical matrix is a very close facsimilie to the expectation provided by Tillé's formula. This speaks well to the fidelity of the sampling algorithm and increases confidence in the samples gathered and the computations thereof. By contrast, either joint matrix when compared with the other products (that is, testing the assertion that $\pi_{ij} = \pi_i\pi_j$) demonstrates a much larger distance. From an information theoretic perspective, one can interpret this to say that it takes nearly 90 additional 'bits' of information to describe $Q$ with $P$.

| | Theoretical $\pi_{ij}$ | Empirical $\pi_{ij}$ | $\pi_i\pi_j$ |
|---|---|---|---|
| Theoretical $\pi_{ij}$ | 0 | 1.971 | 70.71 |
| Empirical $\pi_{ij}$ | 1.971 | 0 | 70.81 |

Table 5.2: Integrated Absolute Error, Columns are F-hat

Progressing from the KL divergence, the relative shape of the distributions can be examined with the Integrated Absolute Error or $L_1$ *Norm*, defined for use with the discrete case as

$$IAE = \sum |\hat{f} - f|$$

The result in Table 5.2 conveys much of the same information as the KL divergence, namely that the crossproduct matrix is quite a bit more different from the theoretical and empirical matrices than the latter two are from each other. With the KL result this is also a good affirmation of our SYG assumption as it suggests that $\pi_{ij}$ and $\pi_i\pi_j$ are in fact different from one another, but not as a uniform inequality. Another interesting result is the comparison of the three distributions against a uniform distribution (Table 5.3). These values are also particularly small. It is also encouraging that the absolute error between the theoretical matrix and the crossproduct is larger in magnitude than the distance between the crossproduct and uniform.

| | Theoretical $\pi_{ij}$ | Empirical $\pi_{ij}$ | $\pi_i\pi_j$ |
|---|---:|---:|---:|
| Uniform Distribution | 2376 | 2376 | 2365 |

Table 5.3: Integrated Absolute Error, Columns are F-hat

Another traditional measure is the Pearson Goodness of Fit, which compares an observed distribution $O$ with an expected distribution $E$, producing a value which follows a $\chi^2$ distribution.

$$\chi^2_{df} = \sum \left( \frac{O_i - E_i}{E_i} \right)^2$$

With knowledge of the degrees of freedom, this allows for hypothesis tests of the difference of two distributions, along with confidence intervals and other tests.

| | Theoretical $\pi_{ij}$ | Empirical $\pi_{ij}$ | $\pi_i\pi_j$ |
|---:|---:|---:|---:|
| Theoretical $\pi_{ij}$ | 0.0000 | 0.0048 | 33.4502 |
| Empirical $\pi_{ij}$ | 0.0048 | 0.0000 | 33.4487 |

Table 5.4: Pearson Goodness of Fit $\chi^2$, Rows are Expected, Columns Observed

Here too, it can be seen that the theoretical and empirical distributions are quite close to each other, and the crossproduct distribution much less so. Although the goodness-of-fit can be computed on this value, the degrees of freedom necessary at $N^2 - 1$ is such a tremendously large number that significance may not be detected if it is in fact there. Indeed, it may not make sense at all in the conventional sense given the input data. Recall that each element $i$ in the inclusion probability vector refers to a single US county, and that in contrast to an experiment or other scenario in which different factors are mutually orthogonal, it is not outrageous to suggest that some or many of the counties may be grouped together and understood as subject to the same effects. Therefore, a sort of "empirical" degree of freedom may be calculated to attempt to discern this true dimensionality.

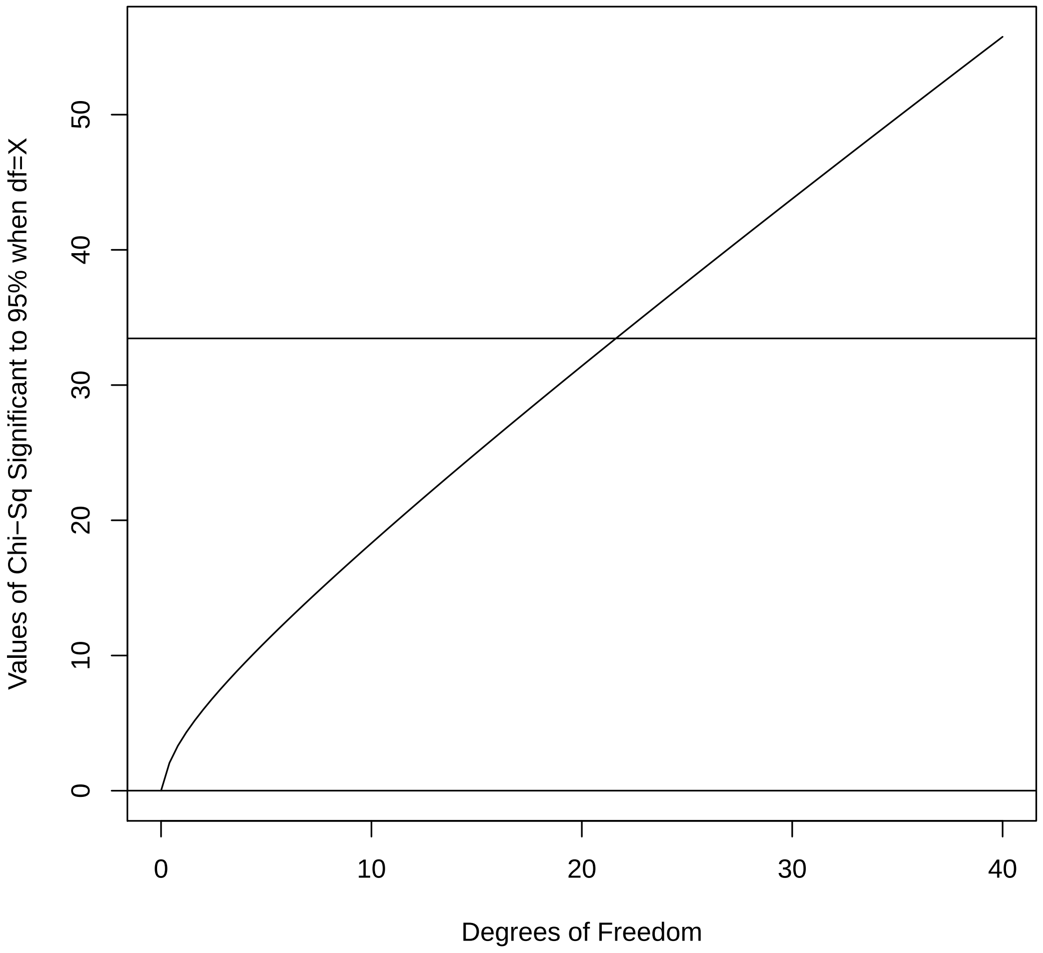


Figure 5.1: Largest values of $\chi^2$ significant at a 95% level for varying values of X

Consider Figure 5.1. This plot traces the critical value at which $\chi^2$ is significant at 95% confidence for degrees of freedom between 0 and 40, with a line drawn at the value of $\chi^2$ produced by the goodness-of-fit test of the crossproduct matrix against the theoretical joint inclusion distribution. Interpreting the KL Divergence as indicating a significant difference, we can place the "true" degrees of freedom at no more than approximately 22. As a goodness-of-fit's degrees of freedom represents the number of cells less one, and acknowledging that both rows and columns of the joint inclusion matrix are represented by the same set of counties, it can be suggested that the 3113 counties observed represent no more than $\sqrt{22}$, or between 4 and 5 distinct groupings though it does not indicate what these groupings may be.

Finally, the difference in matrices will be characterized in terms of their eigenvalues. All three distributions are chiefly characterized by one principle component, and then much less so by the subsequent components as demonstrated in Figure 5.2.

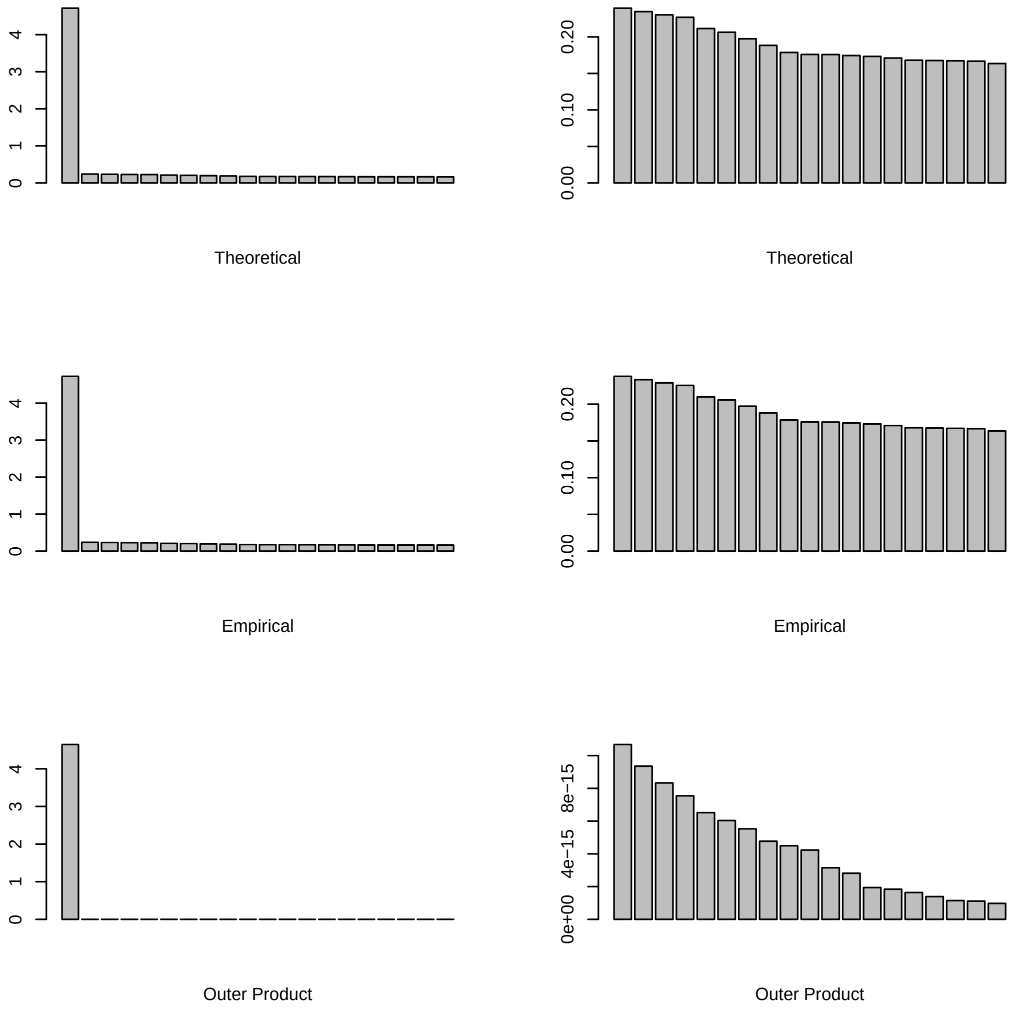


Figure 5.2: Magnitude of Eigenvalues 1:20 (left) and 2:20 (right)

Note that the joint inclusion matrices share an uneven decrease in eigenvalue magnitude whereas the crossproduct matrix has a smooth decline on a much much smaller vertical scale. This suggests the latter lacks the same degree of internal structure as the others, but more importantly it falls below the precision threshold of the computer used (approximately $10^{-14}$) which suggests that it is actually indistinguishable from zero. As Figure 5.2 implies, the theoretical distribution has

|  | $\pi_{ij}$ | $\pi_i\pi_j$ |
|---|---|---|
| $\frac{\sum_2^N \lambda^2}{\sum \lambda^2}$ | 0.0908 | 0 |

Table 5.5: Percent of Variation in Secondary Eigenvalues

variation across several orthogonal components suggesting structure, as where the crossproduct's variation drops quickly to zero with the removal of the first principle component. Table 5.5 makes clear the difference in the two values, demonstrating not only the zero, but that more than 9% of variation in the $\pi_{ij}$ approach cannot be explained by the outer product alone. As analyses thus far are all consistent that the theoretical and empirical joint inclusion matrices are nearly identical, only the theoretical distribution will be considered going forward.

In all these investigations, it can be seen that in comparing the theoretical joint inclusion matrix and the outer product of the first-order inclusion matrix with itself that not only do the two matrices demonstrate divergence from each other, but also that values of one are not strictly greater than the other, and also that their variance structures differ. In conclusion, in employing the CHR estimator, there need not be any concern that $\pi_{ij} - \pi_i\pi_j$ consistently equals zero.

## 5.3 Simulation Study

Several variables were analyzed, each representing a different combination of correlation with probability of selection and value of interest. Additionally, as the Horvitz-Thompson estimator is intended for estimating totals while the Hájek and CHR are intended for means (and SRS is agnostic), both means and totals for the values in question were computed.

The four estimators examined and their corresponding approximate measures of variance are as follows.

Horvitz-Thompson:

$$\hat{t} = \sum y_i/\pi_i, \qquad \bar{y} = 1/N \sum y_i/\pi_i$$
$$\hat{V}(\hat{t}) = N^2(1/n - 1/N)s^2, \qquad \hat{V}(\bar{y}) = -1/2 \sum \check{\Delta}_{ij}(\check{y}_i - \check{y}_j)^2$$

Hájek:

$$\begin{aligned}
\hat{t} &= \frac{N \sum y_i/\pi_i}{\sum 1/\pi_i}, \\
\hat{V}(\hat{t}) &= N^2 \left( \sum \frac{1}{\pi_i^2} \right) \sum \check{\Delta}_{ij} \left( \frac{y_i - \bar{y}}{\pi_i} \right) \left( \frac{y_j - \bar{y}}{\pi_j} \right), \\
\bar{y} &= \frac{\sum y_i/\pi_i}{\sum 1/\pi_i} \\
\hat{V}(\bar{y}) &= \left( \sum \frac{1}{\pi_i^2} \right) \sum \check{\Delta}_{ij} \left( \frac{y_i - \bar{y}}{\pi_i} \right) \left( \frac{y_j - \bar{y}}{\pi_j} \right)
\end{aligned}$$

Chaudhuri-Handcock-Rendall:

$$\hat{t} = N \frac{\frac{1}{2} \sum \sum_{i<j} (y_i + y_j)/\pi_{ij}}{1/\pi_{ij}}, \qquad \bar{y} = \frac{\frac{1}{2} \sum \sum_{i<j} (y_i + y_j)/\pi_{ij}}{1/\pi_{ij}}$$
$$\hat{V}(\hat{t}) = 4N^2 \frac{(N-n)}{N-1} \frac{\sum G_i^2}{\sum \nu_i^2}, \qquad \hat{V}(\bar{y}) = 4 \frac{(N-n)}{N-1} \frac{\sum G_i^2}{\sum \nu_i^2}$$

where

$$\begin{aligned}
G_i &= \sum_{i>j} \sum_{j}^{n} \frac{1}{\pi_{ij}^2} \left( \frac{y_i + y_j}{2} - \frac{\sum_{i>j} \frac{\frac{1}{2}(y_i + y_j)}{1/\pi_{ij}}}{\sum_{i>j} 1/\pi_{ij}} \right) \\
\nu_i &= \sum_{i>j} \sum_{j}^{n} \frac{1}{\pi_{ij}}
\end{aligned}$$

and lastly SRS:

$$\hat{t} = N \frac{\sum y_i}{n}, \qquad \bar{y} = \sum y_i/n$$
$$\hat{V}(\hat{t}) = N^2(1/n - 1/N)s^2 \qquad \hat{V}(\bar{y}) = (1/n - 1/N)s^2$$

Note that in most cases the variance for the total can be converted to the variance for the mean by multiplying by a constant $N^2$, and vice versa. Additionally, the structural similarity of the Hájek and CHR estimators is apparent here.

The HT performs best when the value of interest is highly correlated with the selection probability as it is designed to assume $y_i/\pi_i$ is close to constant. Hájek by contrast is designed as an average, as it contains a term $1/\sum \pi_i = \hat{N}$, effectively averaging over the population size implied by the observed weights. The CHR estimator shares this property as it is a variation on the Hájek with a more nuanced approach to weights. The SRS estimator, essentially a sample average, is equally suited to either estimate and is included in this analysis for comparison only as a "worst case scenario"; as sampling was conducted proportional to size rather than with equal probability the typically unbiased, high-variance estimator will show both bias and greater than typical variance under the PPS sampling scheme.

The results of the simluation study are documented in Tables 5.8 and 5.9, and show the rankings of the estimators and their observed standard errors, with 1 representing most accurate and 4 representing least.

| | $\rho(X, p)$ | HT | SRS | Hajek | CHR |
|---|---|---|---|---|---|
| Num. Dem. Votes | 0.978 | 1 | 4 | 2 | 3 |
| Num. Other Votes | 0.923 | 1 | 4 | 2 | 3 |
| Pct. Voted Dem. | 0.268 | 2 | 3 | 1 | 1 |
| Neg. Corr. | -0.860 | 1 | 4 | 2 | 3 |
| No Corr. | -0.020 | 3 | 2 | 1 | 1 |
| Dem. Win County | 0.236 | 3 | 4 | 2 | 1 |

Table 5.6: Accuracy Ranking for Estimators of 'Total'

The Hájek-based lineage of the CHR estimator is apparent in the results, as like the Hájek estimator it is out-performed by the Horvitz-Thompson in cases

| | $\rho(X,p)$ | HT | SRS | Hajek | CHR |
|---|---|---|---|---|---|
| Num. Dem. Votes | 0.978 | 1 | 4 | 2 | 3 |
| Num. Other Votes | 0.923 | 1 | 4 | 2 | 3 |
| Pct. Voted Dem. | 0.268 | 3 | 4 | 1 | 2 |
| Neg. Corr. | -0.860 | 1 | 4 | 2 | 3 |
| No Corr. | -0.020 | 3 | 2 | 1 | 1 |
| Dem. Win County | 0.236 | 3 | 4 | 2 | 1 |

Table 5.7: Accuracy Ranking for Estimators of 'Mean'

where correlation is either strongly positive or strongly negative, such as a strong correlation case like Figure 5.3. Also similarly, when correlation is low CHR and Hájek come out ahead. In particular, the binary variable indicating a Democratic win in a county (and thus either the expectation of a Democratic win in the mean case or the count of Democratic-won counties in the total case) is predicted best by CHR (Figure 5.4). "Democrats Win" has one of the lower correlations with selection probability at $\rho = 0.23$, and the mean square error of the estimator is superior to both the HT and Hájek cases. Although the MSE is lower in the SRS case, the corresponding estimator is strongly biased and least accurate among all others.

In other cases the CHR estimator was not clearly the best. The percent of a county's votes won by Democrats, another low-correlation variable at $\rho = 0.26$, was predicted equally well by the CHR and Hájek estimators though the CHR had a superior variance for both the total and mean cases. Though nearly identical in correlation to the previous case, this can potentially be explained by the nature of the data; county vote percentage was stored as a value between 0 and 100, while democratic win in a county was stored soley as a 0 or 1. This informs the $y_i + y_j$ term of the CHR estimator. Wheras the binary variable, with a possible value for the term in the set $\{0, 1, 2\}$ can easily acheive a value of 0 the continuous

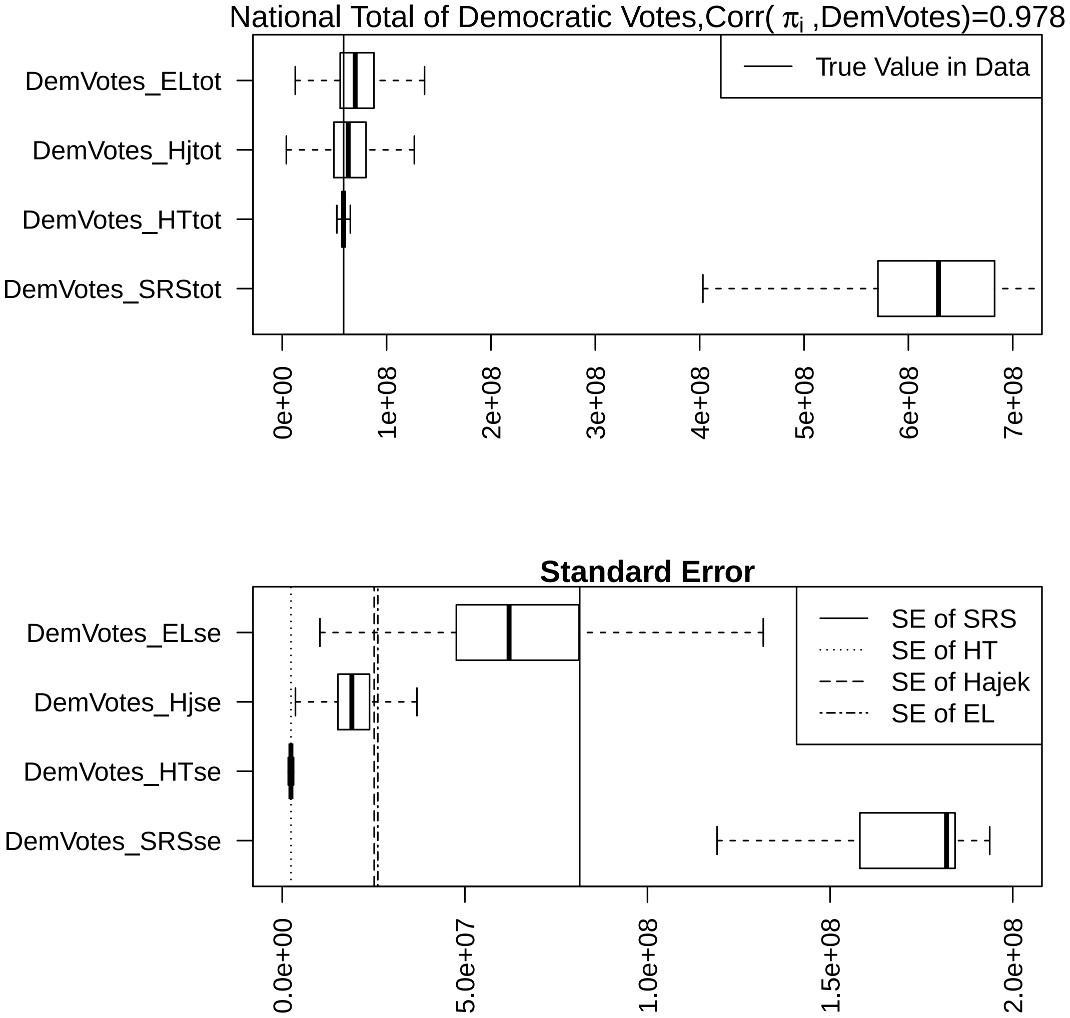


Figure 5.3: National Total of Democratic Votes, y = Democratic Votes by County

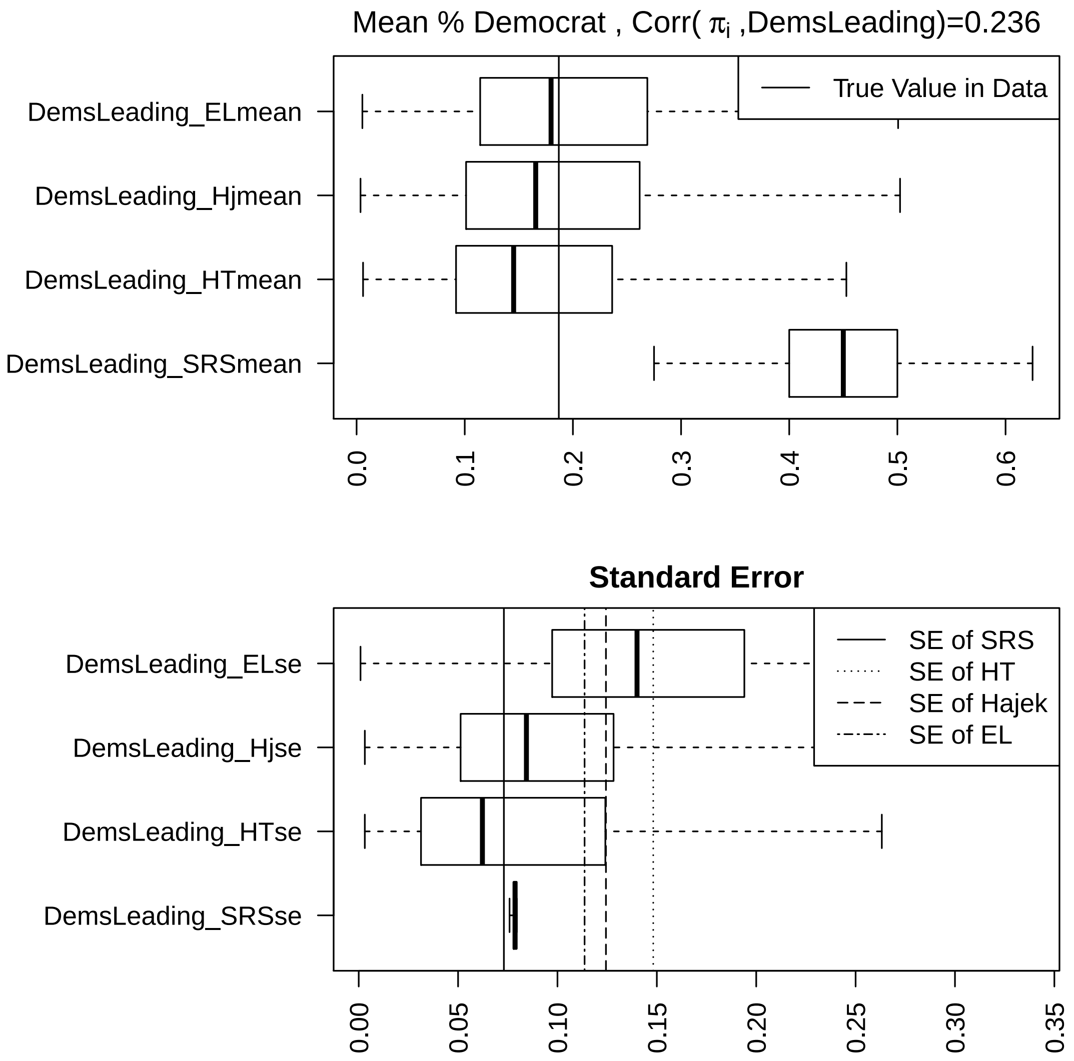


Figure 5.4: Mean Number of Counties Won by Democrats, y=Proportion of Counties Won by Democrats

case makes a value of 0 very unlikely, as it requires two counties with literally no Democratic votes being sampled at once. This virtual inability to cover the entire domain of the term may result in an inflation of the estimator, reducing its efficiency. Finally, in high correlation cases won by the HT estimator, CHR and Hájek performed similarly to achieve second place with very similar estimates and MSEs. Graphical summaries of all estimators and their mean square errors are produced in Appendix A.

In addition to the accuracy of the estimators, it is worth noting the role of standard error in evaluation. The quantities at issue, including percentages, totals and binary variables all show variances on different scales, and so the normal computation for standard error is here swapped for root-normalized mean square error (RNMSE),

$$\frac{\sqrt{MSE(\hat{\theta})}}{x_{max} - x_{min}}.$$

Predictably, when the SRS estimation is a bad estimator for a given quantity its RNMSE is also the worst. More interesting, among the most accurate estimators for a quantity (highlighted in blue in tables 5.8 and 5.9) the most accurate estimator and smallest error never coincide, although they do come close. In the no-correlation case, the CHR estimator even provides the best estimation, but with the worst error!

|  | True Value | $\rho(X,p)$ | SRS | (MSE) | HT | (MSE) | Hj | (MSE) | CHR | (MSE) |
|---|---|---|---|---|---|---|---|---|---|---|
|  | True Value | $\rho(X,p)$ | SRS | (MSE) | HT | (MSE) | Hj | (MSE) | CHR | (MSE) |
| Num. Dem. Votes | 5.9E+07 | 0.98 | 6.3E+08 | 0.73 | 5.9E+07 | 0.1 | 6.3E+07 | 0.069 | 7E+07 | 0.079 |
| Num. Other Votes | 1.2E+06 | 0.92 | 1.1E+07 | 0.66 | 1.2E+06 | 0.083 | 1.3E+06 | 0.076 | 1.4E+06 | 0.086 |
| Pct. Voted Dem. | 1.2E+05 | 0.27 | 1.5E+05 | 0.49 | 1.1E+05 | 0.037 | 1.2E+05 | 0.083 | 1.2E+05 | 0.086 |
| Neg. Corr. | -6.1E+11 | -0.86 | -4.9E+12 | 0.64 | -6.1E+11 | 0.1 | -6.5E+11 | 0.069 | -7.2E+11 | 0.078 |
| No Corr. | 1.5E+05 | -0.02 | 1.5E+05 | 0.1 | 1.4E+05 | 0.033 | 1.5E+05 | 0.11 | 1.5E+05 | 0.12 |
| Dem. Win County | 5.8E+02 | 0.24 | 1.4E+03 | 0.39 | 4.5E+02 | 0.022 | 5.2E+02 | 0.13 | 5.6E+02 | 0.13 |

Table 5.8: Median Sample Estimate and Estimator Normalized Root MSE for Estimators of Totals

|  | True Value | $\rho(X,p)$ | SRS | (MSE) | HT | (MSE) | Hj | (MSE) | CHR | (MSE) |
|---|---|---|---|---|---|---|---|---|---|---|
|  | True Value | $\rho(X,p)$ | SRS | (MSE) | HT | (MSE) | Hj | (MSE) | CHR | (MSE) |
| Num. Dem. Votes | 1.9E+04 | 0.98 | 2E+05 | 0.73 | 1.9E+04 | 0.1 | 2E+04 | 0.069 | 2.2E+04 | 0.079 |
| Num. Other Votes | 3.9E+02 | 0.92 | 3.6E+03 | 0.66 | 3.9E+02 | 0.083 | 4.2E+02 | 0.076 | 4.6E+02 | 0.086 |
| Pct. Voted Dem. | 39 | 0.27 | 48 | 0.49 | 37 | 0.037 | 40 | 0.083 | 40 | 0.086 |
| Neg. Corr. | -2E+08 | -0.86 | -1.6E+09 | 0.64 | -2E+08 | 0.1 | -2.1E+08 | 0.069 | -2.3E+08 | 0.078 |
| No Corr. | 50 | -0.02 | 48 | 0.1 | 45 | 0.033 | 50 | 0.11 | 50 | 0.12 |
| Dem. Win County | 0.19 | 0.24 | 0.45 | 0.39 | 0.15 | 0.022 | 0.17 | 0.13 | 0.18 | 0.13 |

Table 5.9: Median Sample Estimate and Estimator Normalized Root MSE for Estimators of Means

# CHAPTER 6

# Discussion

The intent of this study was to evalute the estimator put forth by Chaudhuri, Handcock and Rendall, and in doing so to review much of the elementary theory that supports it. Survey sampling with and without auxiliary information was discussed, and also the empirical likelihood of Owen was discussed. Finally, the CHR estimator itself was described and its performance evaluated in the context of a dataset describing voting patterns in the 2004 US Presidential Election. In addition to comparing sample estimates of voting percentages and proportions against a known population result, the data itself was examined for sources of variation, chiefly by attempting to derive its degrees of freedom.

At the outset of this study it was not at all clear how the CHR estimator would perform when held up against more traditional methods. This comes mostly as a result of its novel application of empirical likelihood, particularly because weights are typically treated as a direct inversion or other function of the observed sampling probabilities instead of as random variables. Although the theoretical underpinnings of the strengths and weaknesses of the estimator are detailed in Chaudhuri, Handcock and Rendall's own papers, the simulation conducted here demonstrates that it closely follows the trend of the Hájek estimator upon which it was based, always equating its accuracy and in cases of low correlation surpassing it. In cases of high correlation, other estimators such as the Horvitz-Thompson provide more efficient estimates. The degrees of freedom in the data were estimated to be no greater than $\sqrt{22}$, suggesting that among all of the 3113 counties represented as sampling units in the data there are only three or four distinct categories providing the observed variation.

The study conducted here can be extended in several ways. First, the estimates of the superpopulation parameters might be put to work in a predictive estimation such as a regression, attempting to determine the nature of the superpopulation and using it to reproduce the variable of interest across the finite population. Another extension could be an in-depth study of the role of auxiliary information. In addition to the simplex investigated above to determine weights, Chaudhuri, Handcock and Rendall specified a second simplex in which the weights are optimized against an auxiliary variable $A$. Finally, the estimator may be recast in terms of estimators other than the Hájek to attempt to improve upon specialized domain-specific estimators.

# Appendix A

# Estimator Comparisons

The following are plots of variables of interest for each of the discussed estimators. They reflect estimations of samples selected proportional to size, as drawn via Tillé's method Tillé (1996).

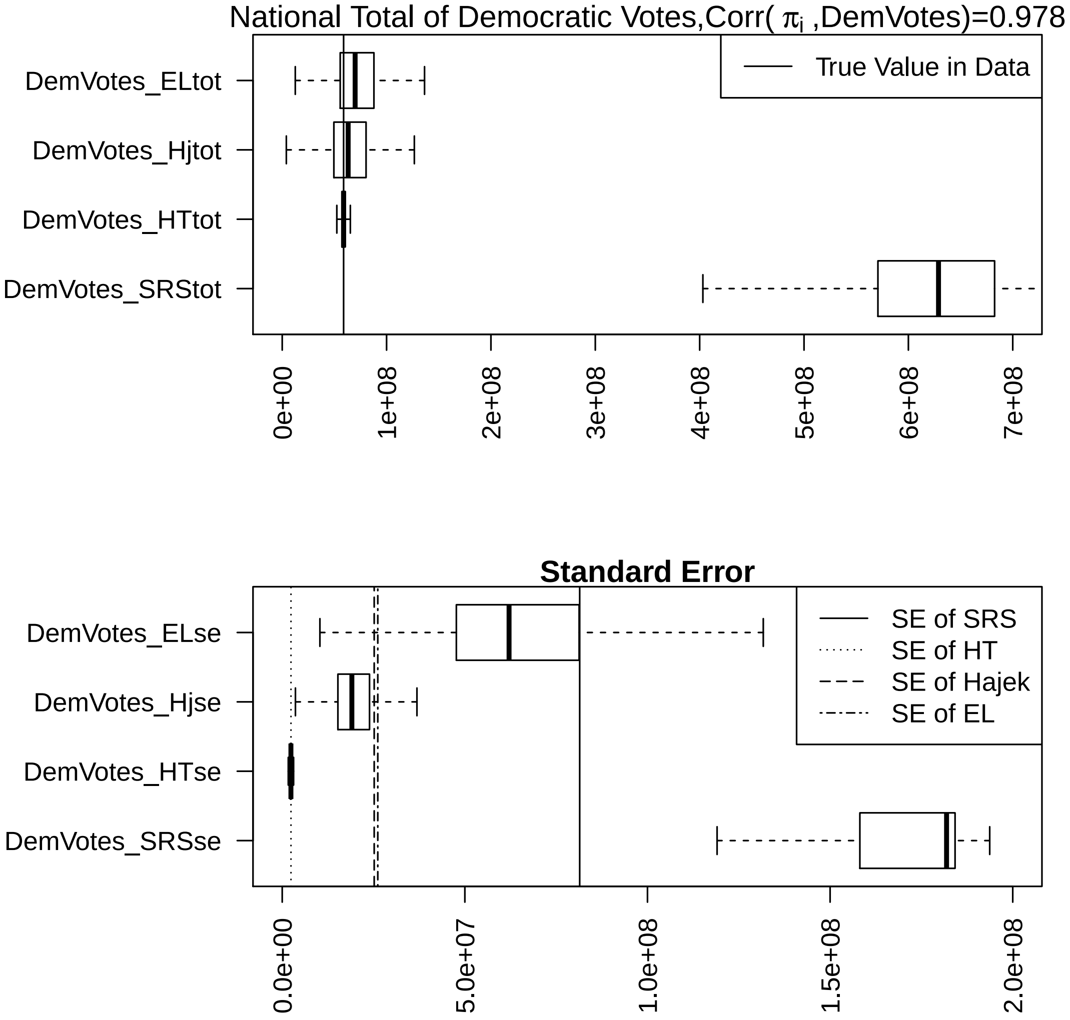


Figure A.1: National Total of Democratic Votes, y = Democratic Votes by County

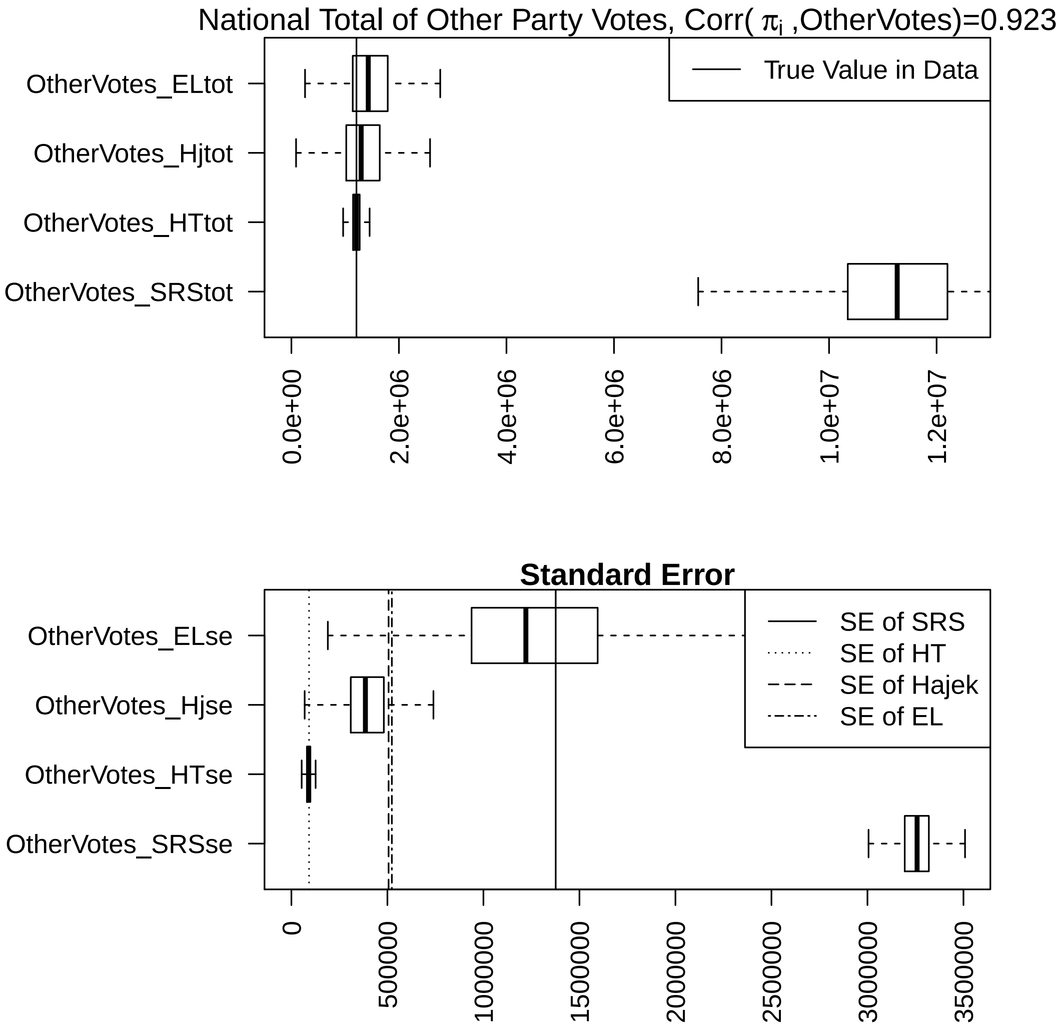


Figure A.2: National Total of Other Party Votes,y = Other Votes by County

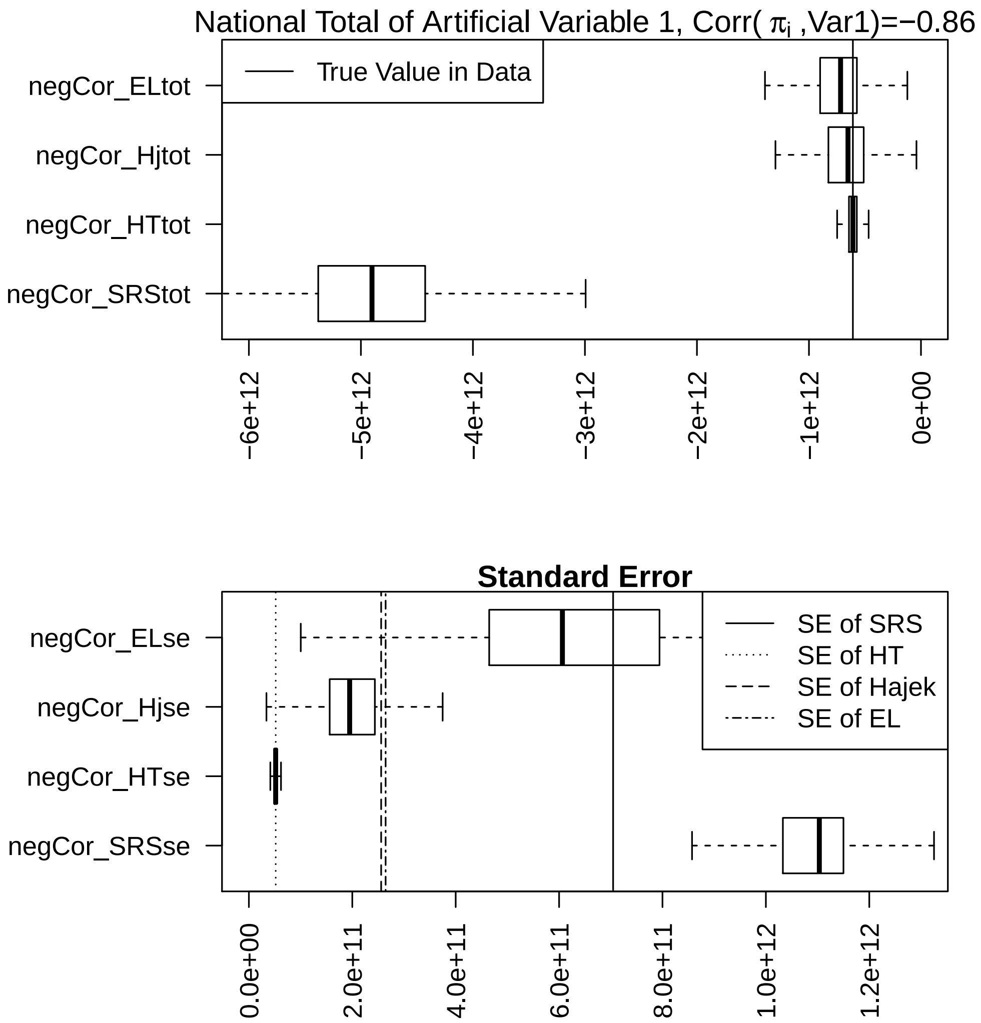
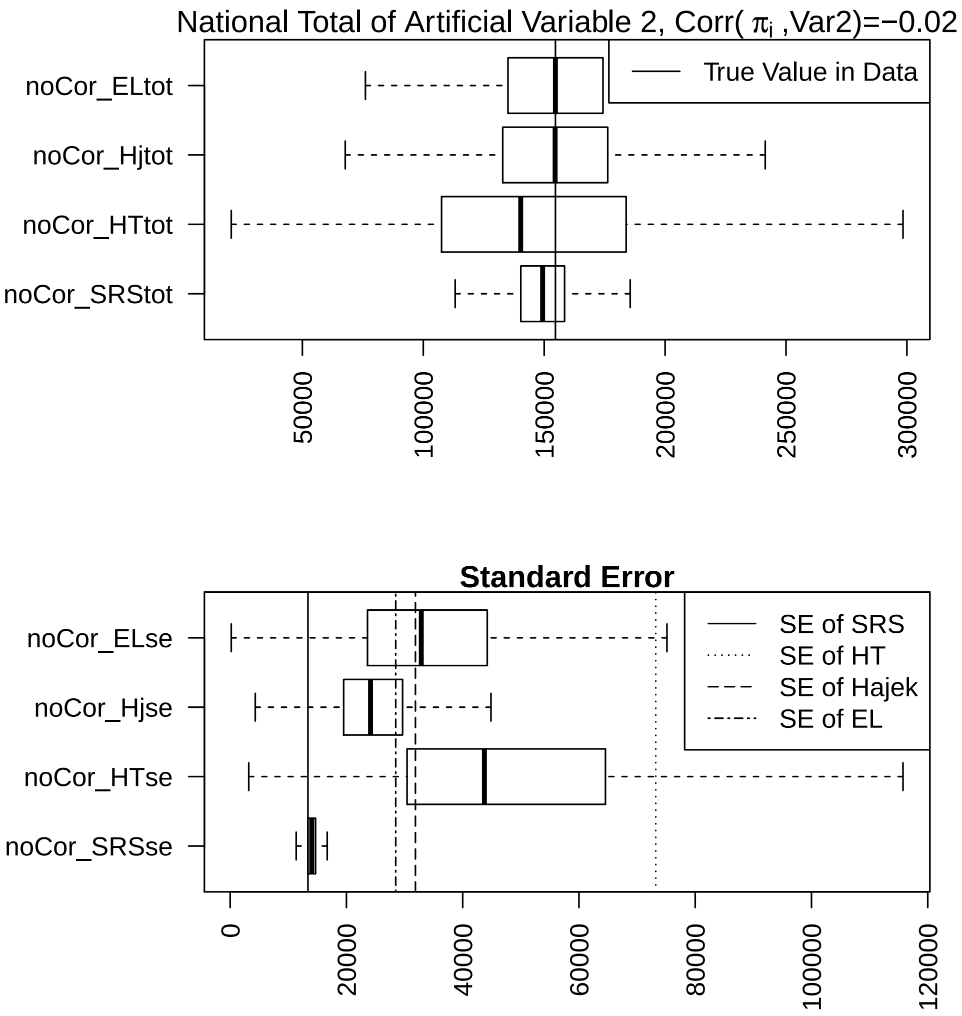


Figure A.3: National Total of Artificial Variable 1, y = Variable 1


National Total of Artificial Variable 2, Corr( $\pi_i$ ,Var2)=−0.02

Standard Error

Figure A.4: National Total of Artificial Variable 2, y = Variable 2

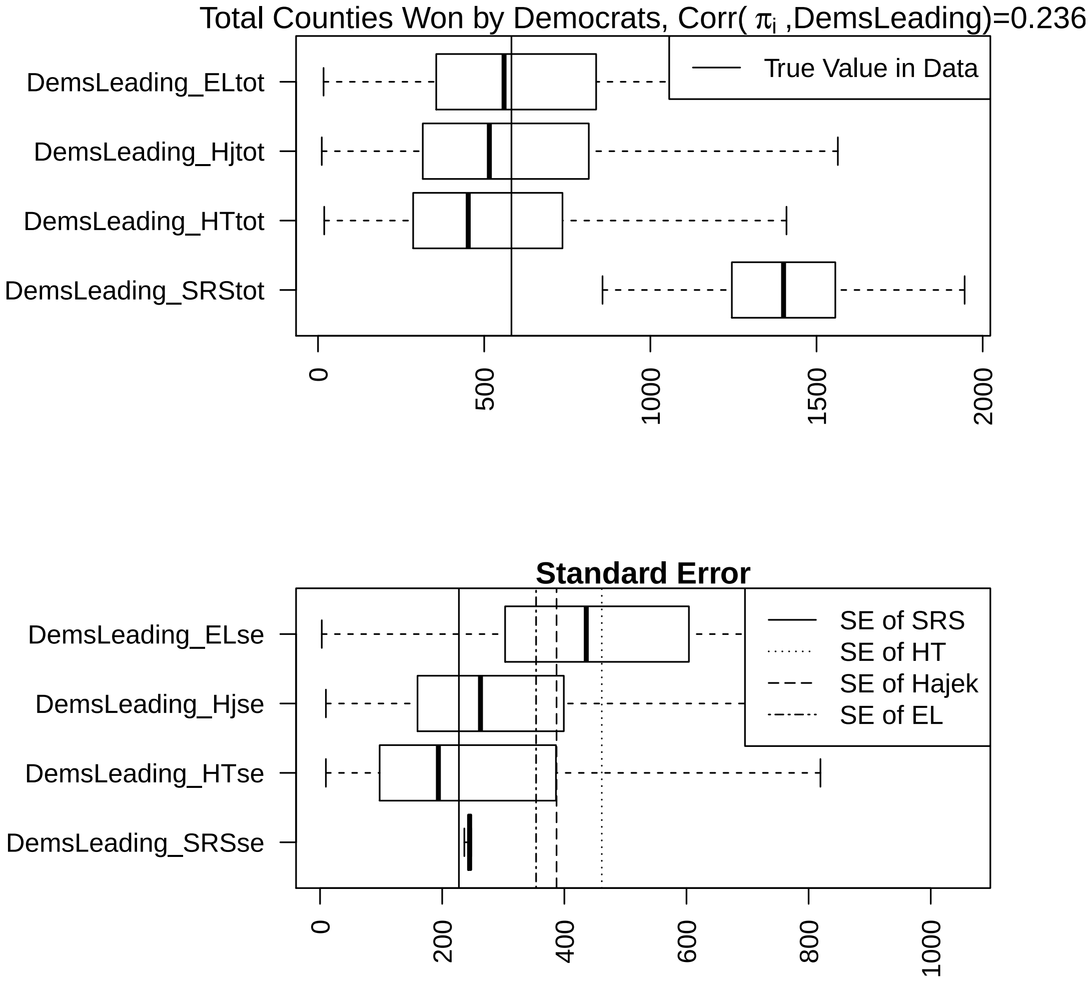


Figure A.5: Total Number of Counties Won by Democrats, y = Binary, Democrats Win County

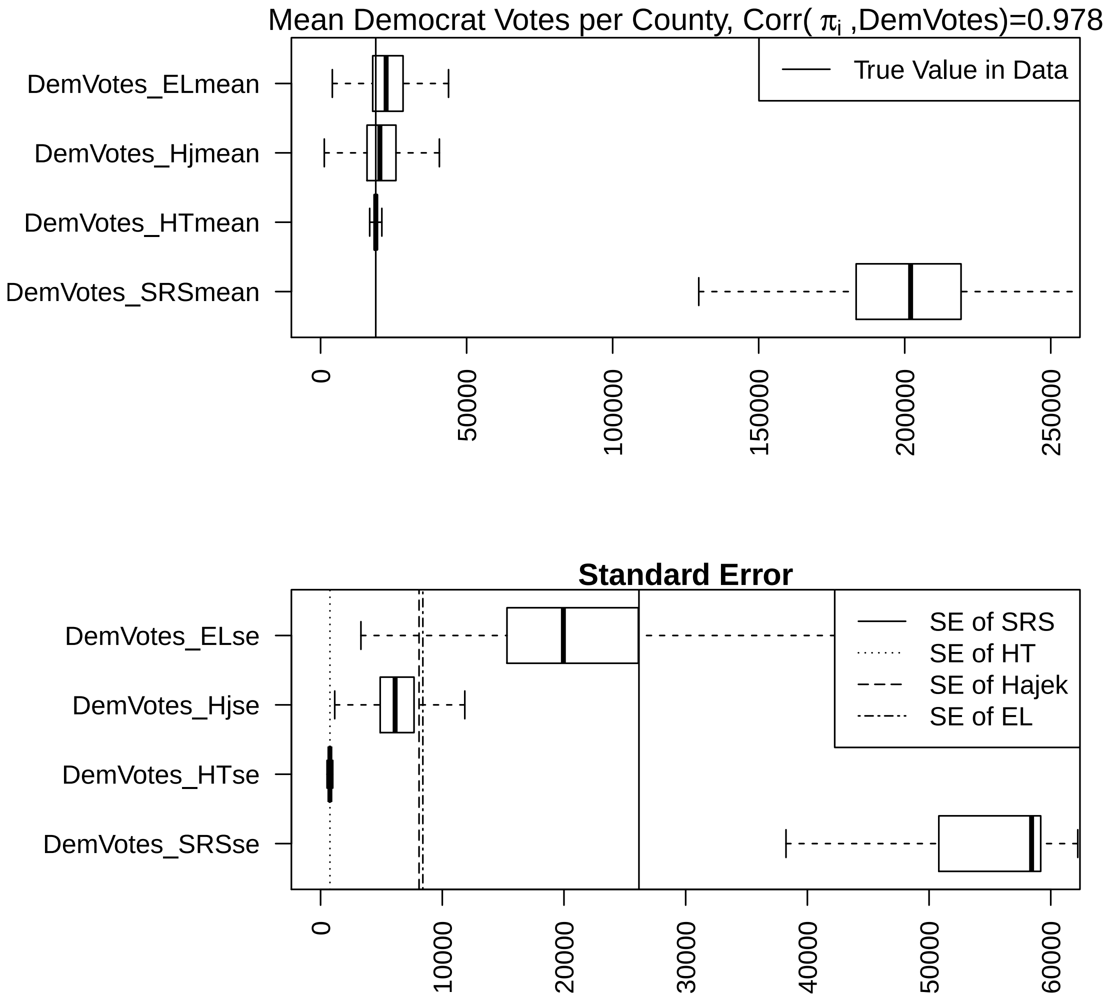


Figure A.6: Number of Democratic Votes, y = Democratic Votes per County

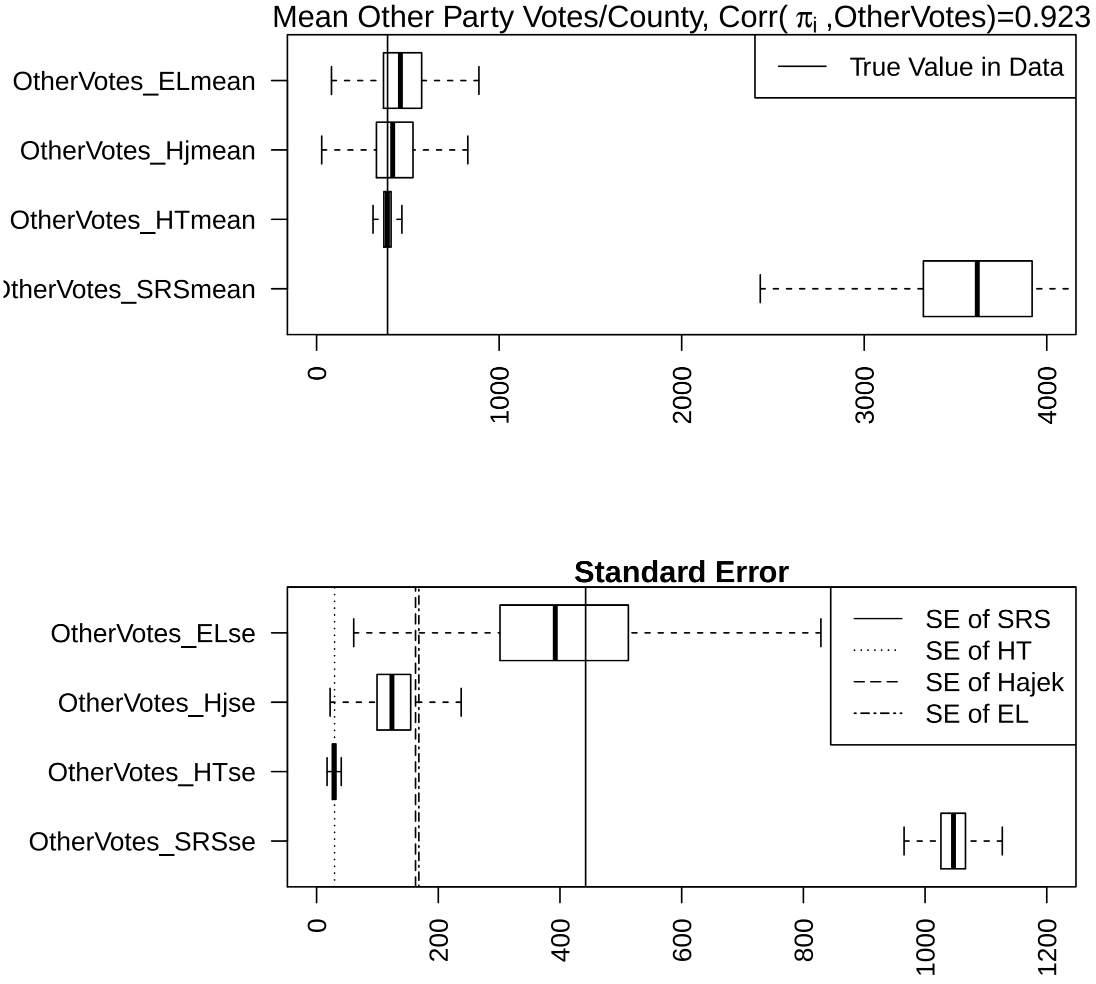


Figure A.7: Number of Other Votes, y = Other Votes per County

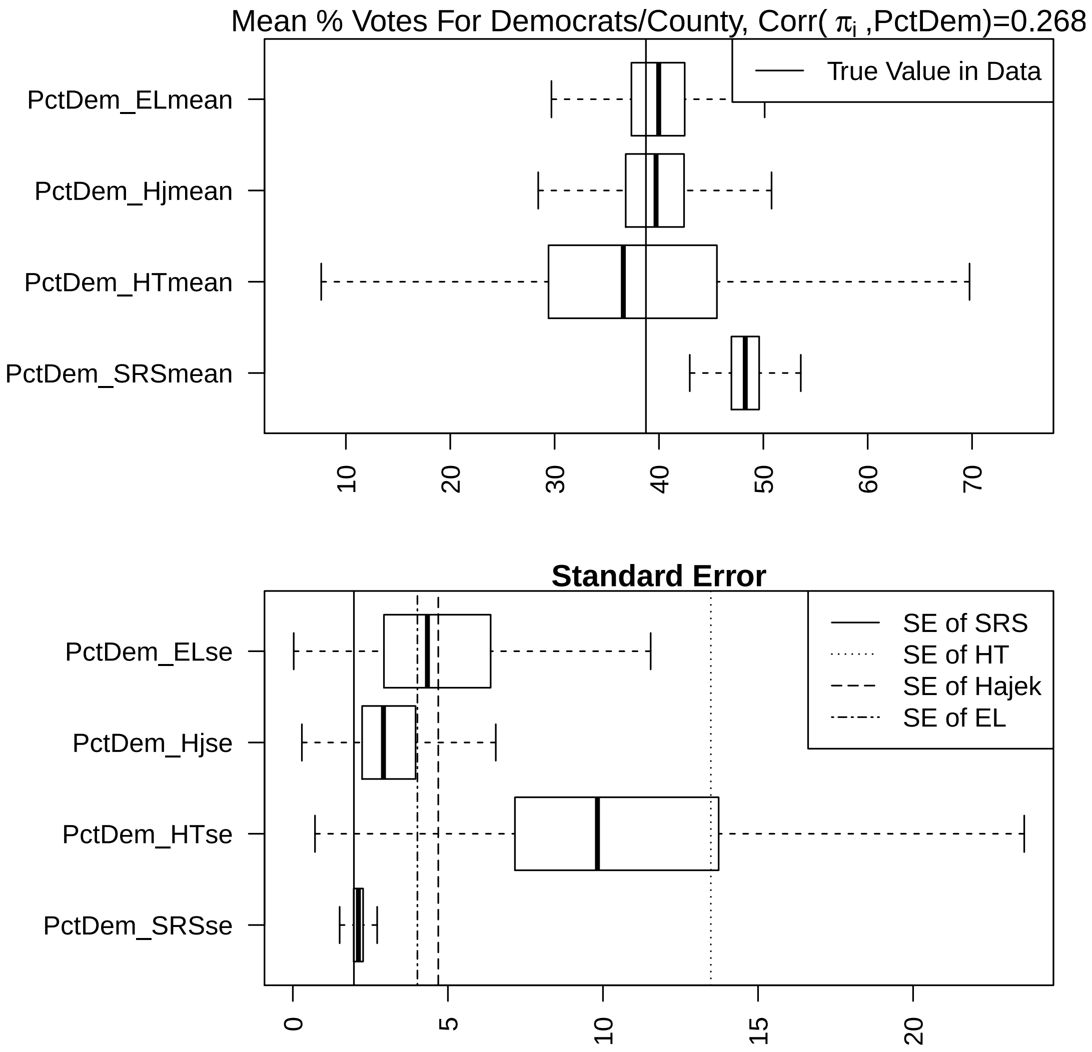
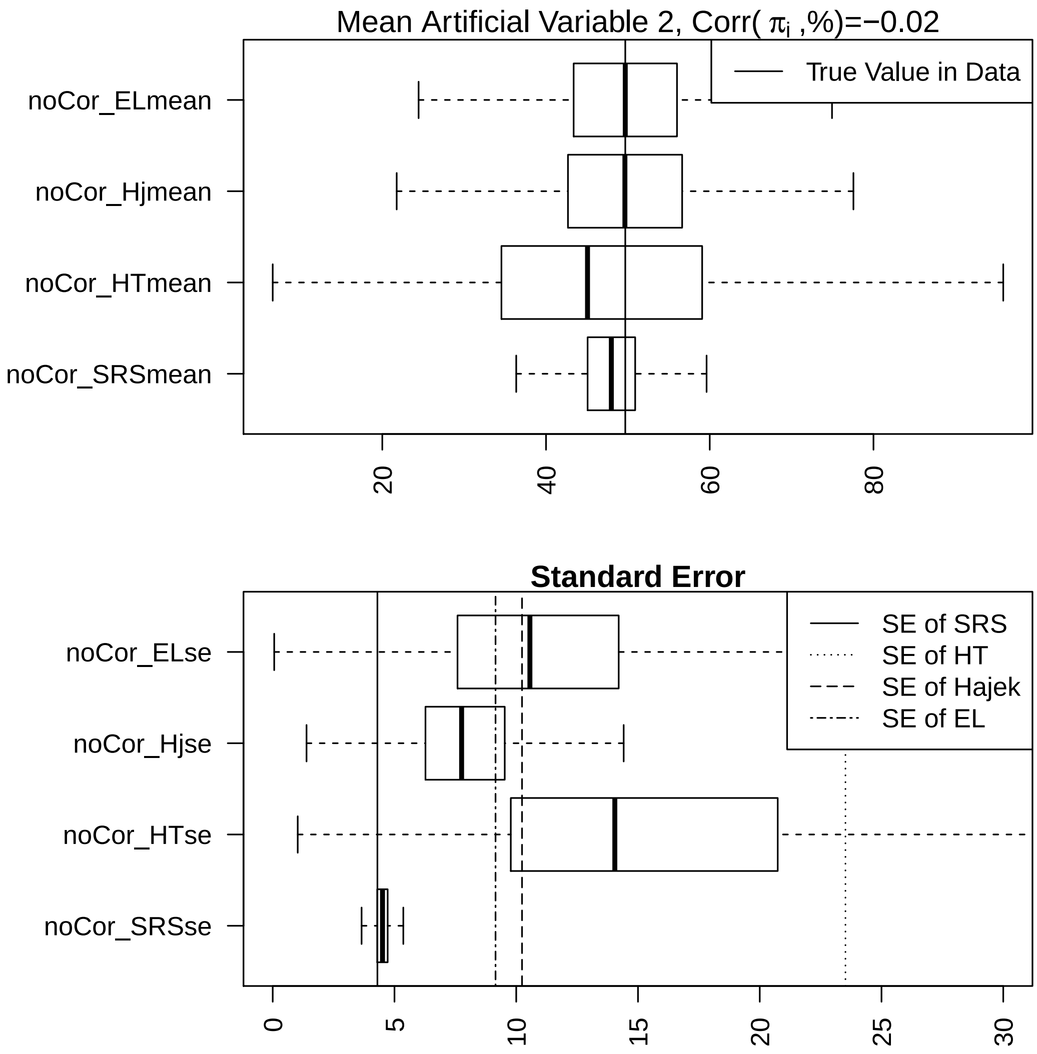


Figure A.8: Mean Percentage of County Voting Democrat, y = Percentage Voting Democrat by County


Mean Artificial Variable 2, Corr( $\pi_i$ ,%)=−0.02

**Standard Error**

Figure A.9: Mean Artificial Variable 2, y=Variable 2

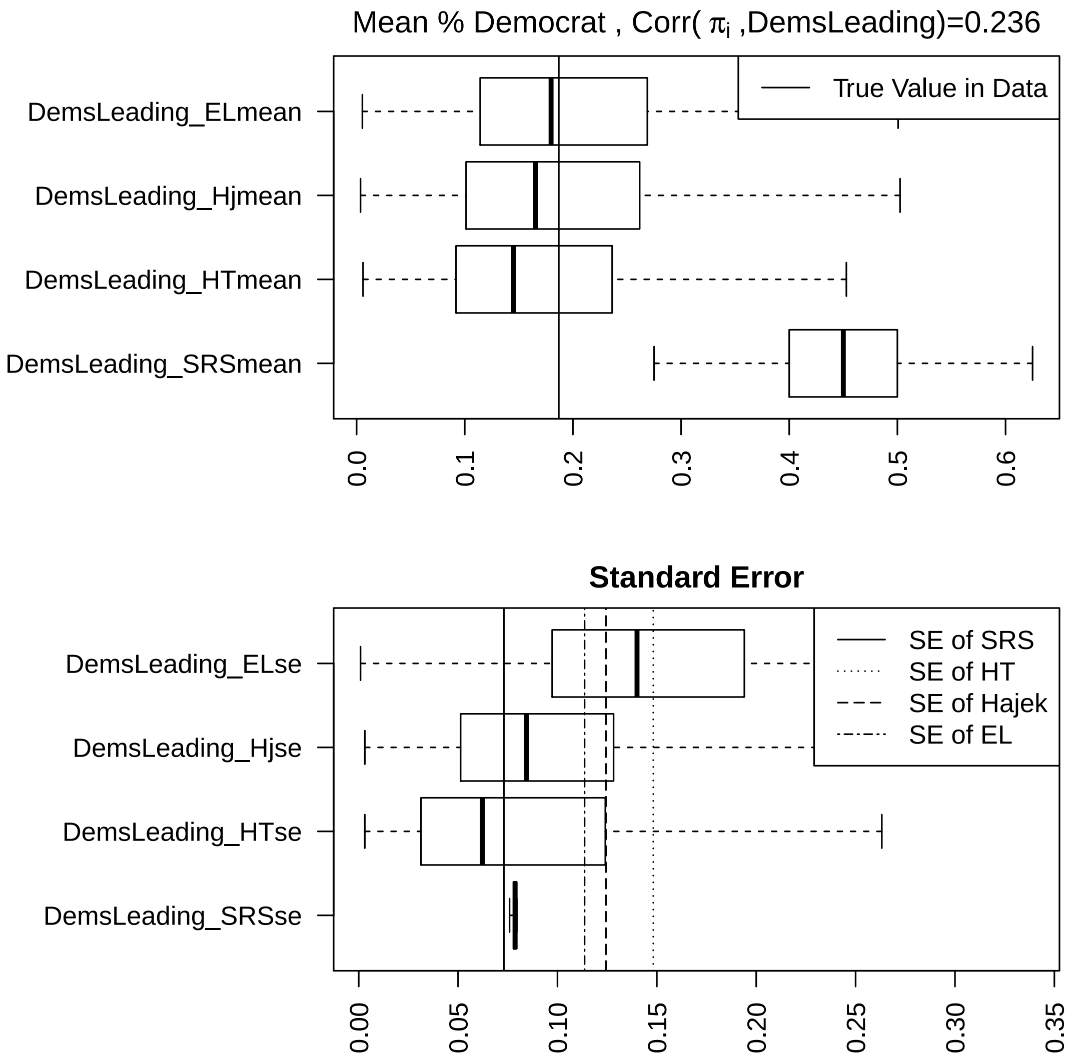


Figure A.10: Mean Number of Counties Won by Democrats, y=Proportion of Counties Won by Democrats

# Bibliography

(2013). Online etymology dictionary .

Brewer, K. (1983). W & hanif, m.(1983). sampling with unequal probabilities. *Lecture Notes in Statistics* **15**.

Bureau, U. C. (2009). Usa counties.

Chaudhuri, S., Handcock, M. S. & Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 311–328.

Chaudhuri, S., Handcock, M. S. & Rendall, M. S. (2013). A conditional empirical likelihood approach for combining sampling design and population level information. Unpublished manuscript.

Commission, U. E. A. (2006). Election administration voting survey 2004.

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical science* , 297–310.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Lumley, T. (2011). *Complex surveys: A guide to analysis using R*, vol. 565. Wiley.

Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall, CRC, Boca Raton.

Patil, G. P. & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, pp. 179–189.

PAWITAN, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford Science Publications. OUP Oxford.

PFEFFERMANN, D. & SVERCHKOV, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B* , 166–186.

PFEFFERMANN, D. & SVERCHKOV, M. (2003). Fitting generalized linear models under informative sampling. *Analysis of Survey Data* , 175–195.

QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* , 300–325.

SÄRNDAL, S. & WRETMAN (1992). *Model assisted survey sampling*. Springder-Verlag: New York, NY.

SEN, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

TILLÉ, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika* **83**, 238–241.

TILLÉ, Y. & MATEI, A. (2012). *sampling: Survey Sampling*. R package version 2.5.

YATES, F. & GRUNDY, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)* , 253–261.