Fall 2015

Techniques for Estimating the Variance of Specific Estimators within Complex Surveys

Laura Marie Galiardi

TECHNIQUES FOR ESTIMATING THE VARIANCE OF SPECIFIC ESTIMATORS WITHIN COMPLEX SURVEYS

A Thesis

Submitted to the McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Master of Science

By

Laura Galiardi

December 2015

TECHNIQUES FOR ESTIMATING THE VARIANCE OF SPECIFIC ESTIMATORS WITHIN COMPLEX SURVEYS

By

Laura Galiardi

Approved July29, 2015

_____	_____
Frank D'Amico, Ph.D. Professor of Statistics (Committee Chair)	John Kern, Ph.D. Associate Professor of Statistics, Chair (Committee Member)
_____	_____
Jeffery Jackson, Ph.D. Professor of Computer Science Director of Graduate Studies (Committee Member)	James C. Swindal, Ph.D. Dean, McAnulty College

ABSTRACT

TECHNIQUES FOR ESTIMATING THE VARIANCE OF SPECIFIC ESTIMATORS WITHIN COMPLEX SURVEYS

By

Laura Galiardi

December 2015

Thesis supervised by Dr. Frank D'Amico Professor of Statistics

The objective of this thesis is to present the various procedures for estimating the variance of specific statistics obtained from different types of survey designs, leading up to more advanced designs such complex surveys. The thesis starts with defining the various sampling designs that are to be used for illustrations (Ch 2). Chapter two gives further descriptions of how various sampling designs are performed (from simple designs to not so simple) and shows the sophistication in calculating the estimates and variances. Chapter three cites the actual equations necessary for estimating the variances of the statistics for each design and demonstrates the potential difficulty especially in estimating the variance of the statistics, as the designs get more complex. Each design is illustrated with numerical examples. Chapter four defines current methods for estimating the variance and introduces the Bootstrap and Jackknife approaches. In Chapter 5 the ideas behind what is considered to be a "complex survey" are described and two

nationally known complex surveys (NHANES and NHIS) currently being done in the U.S. are explained as examples. Chapter six reports the main statistical results, comparing the variances, etc., for all the designs and finally a summary conclusion is in chapter 7.

DEDICATION

This thesis is dedicated to my family for supporting me through the long, challenging but rewarding journey of my master's degree, especially my son Jackson who pushed me to complete this thesis.

In addition, I would like to thank my thesis advisor Dr. D'Amico for working so hard with his busy schedule to make sure I finished.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Obtaining information and knowledge about human beings has been an issue since civilizations began. Wanting to study human behaviors, emotions, and other pertinent characteristics, man has come up with meaningful ways to "survey" the population. Survey as used in this context implies the simple act of obtaining information from a single individual or collection of individuals. Usually surveys come in the form of a questionnaire that is filled out by someone but it can also be obtained in other ways such as a face-to-face interview, telephone, etc. However, the information obtained is simply accurate to what is collected at that point in time. In the case of health questionnaires, the surveys are called "prevalence" surveys because the statistics calculated from the surveys are prevalence's of diseases.

Survey design (or methodology) studies the sampling of individual units from a population. Most simple surveys collect data in the form of a questionnaire that is administered to only a random portion of the population, the sample. The answers to the questionnaire are in turn the data that will be statistically analyzed.To ensure the most accurate results the sample should mirror the target population being studied and the survey design(which is how the sampling is to be performed) should not allow for any bias.

In order for a survey to have the most accurate results the sampling method and survey design must follow specific guidelines. Kish (1965)explains," the sample design has two aspects: a selection process (that is, the rules and operations by which some members of the population are included in the sample); and an estimation process/estimator for computing the sample statistics (which yield the sample estimates of the population values). Different sampling designs would result in different estimates and variances. Choosing the design with the smallest error is the principle aim of sampling design" (Kish, 1965).

The true values of the population are known as the parameters. Although for the U.S. Population, they are usually unknown. The estimator or statistic is an estimate computed from the ***n*** elements in the sample. For example, the sample mean ($\bar{x}$) is the computed mean of the***n observations***. This particular estimate is only one among the many possible estimates that could have been obtained from the sampling design.Because the parameters of the population are unknown it is important to understand the variability of the sample estimates. Knowing how the sample estimates can vary helps us to draw conclusions about the parameters. A way of estimating the sampling variability would be to take another sample from the population and compare results. If this was repeated many, many times you would have what is known as the sampling distribution for the statistic. From the distribution of these samples you could determine approximate estimates of the parameters and the variance. This is the asymptotic theory that traditional statistics is based on. However with the recent advances in technology there are other methods (such as "resampling") being employed for obtaining estimates of the various parameters. Resampling methods treat the sample as if it were itself a population; then we take different samples from this new population and use the subsamples to estimate the variance and other statistics. Bootstrap and Jackknife (Lohr, 1999)are two resampling methods that are currently seen in the literature and there are various software packages that can apply these methods.

Still there are various designs for actually taking the sample. Simple random sample, systematic random sample, stratified, cluster and multi-stage sampling are just some names ofthe methods. Each design has its specific steps depending on what the population looks like. It is easy to assume that all we have to do is reach in and take a suitable sample ofunits from the population. But the units (could be people) are not necessarily nicely defined, even when the

population appears to be so. There may be several ways of listing the units, and the size we choose may very well contain smaller subunits. For example suppose we want to find out how many bicycles are owned by residents in a community of 10,000 households. We could just list the households and then take a simple random sample (SRS) of say 400 households or as many as we have estimated we needed. However, maybe the community can be divided into blocks (say 20 households each). The blocks may be grouped geographically for convenience and then randomly sample a certain number of blocks from the listing. Once a block has been selected, then we may either sample all the households within that block or take another random sample of households within each block. This latter plan is an example of cluster sampling. The blocks are the primary sampling units (psus) and the households are the secondary sampling units (ssus).

Bootstrapping is a resampling technique that calculates the variance estimates for a sample in which the Probability Sampling Unitsare sampled with replacement. Another resampling technique that calculates the variance estimates is the Jackknife method, however the Psus are not sampled with replacement in this method.For some clusters such as classrooms, medical practices, or workplaces, it may be just more convenient to sample the entire cluster than to attempt to subsample within the cluster. It is important to know that each method of sampling yields different estimates of the statistics and their variances. Depending on the structure of the sampling some methods may overestimate the variance and result in conservative confidence intervals (Lohr, 1999).

In simple random sampling surveys each observation has the same probability of being chosen. However, there are some cases in which each observation is disproportionately chosen making analysis difficult. Sometimes certain areas of a population may be oversampled, such as what might happen in trying to estimate the health quality of Hispanic or some other group that

may not be easily identified for sampling. In which case, we might oversample a particular neighborhood where we know many Hispanics migrate, otherwise in simple random sampling, these individuals would never be selected and our health estimates of the population would be skewed. These cases make up what are known as complex surveys and generally require quite a bit of mathematical adjustment in order to ultimately get unbiased estimates.

Because statistical computing software is now so readily available, resampling techniques have become the standard for computing the variability of various estimators in survey analysis. R is a language and environment for statistical computing and graphics. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems including Windows and MacOS. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Through the online data downloading there are macros written in R that can be used for analyzing survey data for SRS designs thru complex survey designs. And it is free where other computer software for analyzing complex designs, such as SUDAAN, SAS Survey, STATA, and SPSS SVR, all cost a minimum of approximately $1000 or more per yearly license per machine.

Chapter 2 Types of Survey Sampling Designs

Simple Random Sampling

"Simple random sampling" (SRS) is a very popular form of probability sampling because of its simplicity in use and analysis. A simple random sample is a subset (a sample) taken from the population. Each item (unit, patient, etc.) in the sample has equal probability of being chosen from the population (which is $\frac{1}{N}$, *N* being the number of items in the population). Ideally in small populations items should not be sampled more than once. Therefore one usually takes a simple random sample without replacement. However the probability of selection in small populations or when a substantial sample size is taken from a larger population can affect the variance estimates. But for the most part, ***n*** (sample size) divided by ***N*** (population size) is small (less than 5%) and consequently much of the theory in estimating the variance is based on sampling with replacement even though it is not true.

An example may be sampling patients with high blood pressure from hospitals around the county, in order to estimate the rates of various drug medications being used. Suppose there are a total of 1,000 patients with high blood pressure hospitalized in the county. Each patient has a $\frac{1}{1{,}000}$ chance of being chosen if this sampling is done with replacement. If the sampling is done without replacement then the first patient has a $\frac{1}{1{,}000}$ chance. The second patient has a $\frac{1}{999}$ chance and the third has $\frac{1}{998}$ chance of being chosen and so on down the line, until reaching the last patient who has 100% chance of being sampled. This is impractical because the idea is not to have a sample the same size of the population. Note that usually you start with an estimation process to determine the size of the sample that is necessary to accomplish the analysis. This is based on other factors that are not covered or discussed in this thesis. But once ***n*** is determined

then ***n/N*** is called the "sampling fraction and abbreviated with the letter " *f.* " The sampling fraction represents what percent of the population needs to be picked. Then depending on the structure of the population, exactly how we obtain the ***n*** units is referred to as the "design". Additionally, one other definition that is important in estimating variances is the finite population correction (fpc) factor. It is defined simply as one minus the sampling fraction, in other words,

$$\text{FPC} \quad = \quad 1 - f \quad = \quad 1 - \frac{n}{N} = \quad \frac{N-n}{N}.$$

From this equation, it can be seen that when ***n*** is relatively small compared to ***N***, then the fpc is approximately 100%. Similarly when ***n*** is large relative to ***N***, ultimately meaning that a substantial size of the population is being selected then the fpc plays a larger role in the variance.

In Figure 2.1, the population size is N = 12 and if n = 4 is chosen a priori, then each person in the population has a $\frac{1}{12}$ chance being sampled if replacement occurs. Thus the sampling fraction is 25% and if replacement does not occur then person two (say the first casedrawn) has a $\frac{1}{12}$ chance of being sampled. Person five (the second case to be selected) has a $\frac{1}{11}$ probability of being chosen, while person eight (third selection) has a $\frac{1}{10}$ chance, and the fourthperson (number 10) in the sample had a $\frac{1}{9}$ chance of being selected (Kernler, 2014). If this were a true sampling situation then the fpc would certainly influence the variance.

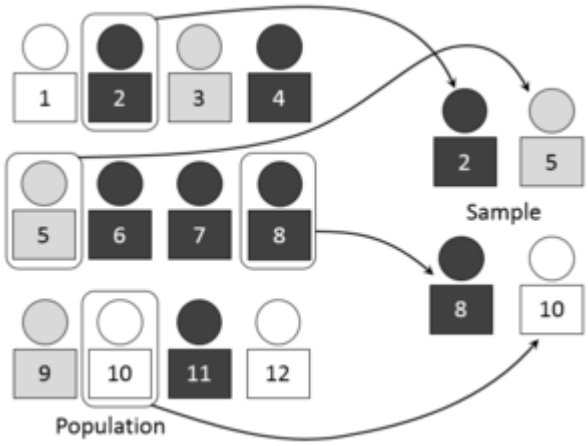

Figure 2.1. Source: Dan Kernler (2015b), licensed under CC BY SA 4.0

Systematic random sampling

Another form of sampling is systematic sampling; which is performed by arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start, known as the seed, and then proceeds with the selection of every k^{th} element from then onwards. In most cases, $k = \frac{N}{n}$. Usually systematic random sampling is done in multiple passes through the ordered list. At the outset, each element has an equal probability of being chosen in order to eliminate any bias. For example, suppose there were 1,000 patient visits last year at some clinic and we wish to chart audit a random sample of 50 of them (5%). One could simply obtain the listing of those visits and do a SRS. Or systematically, you could start with the listing and decide you'll make two passes through the listing. At the end of each pass, you would have chosen about 25 charts from each pass and you would have your 50 cases at the end of two passes. If you were only

doing one pass, you would randomly choose a number from 1 to 20 (20 because 1000/50 = k = 20) and then once having that number (say it is the 10th), then you choose the 10th, 30th, 50th, all the way through the last draw. So the system is choosing every 20th case throughout the listing. If you wanted to do this in two passes then you would randomly pick a number from 1-40, say that number is 6, then you sample the 6th, the 46th, the 86th up to the end. That would produce an approximate sample size of 25. You would then take another random number from 1-40 (now 6 is excluded) and repeat the process to obtain the 50 total cases to be audited.

Figure 2.2 shows a simple picture example of a systematic draw where there are twelve people in the population and we need a sample size of four. The seed randomly starts at person two and every third (12/4 = 3) person is sampled. This leaves four people sampled out of the twelve people in the population. As in the previous SRS example, if this were a real problem then the fpc would need to be incorporated in the estimation equations due to its influence on the variance.

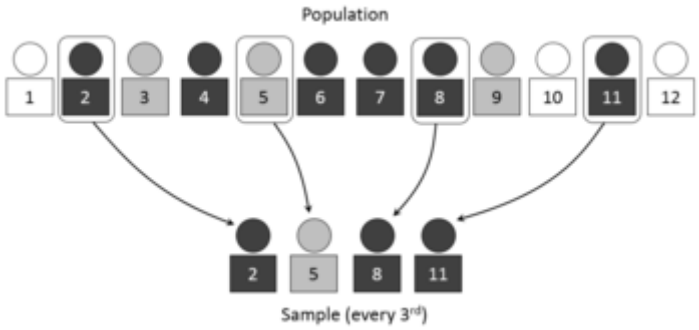

Figure 2.2 Source: Dan Kernler (2015d), licensed under CC BY SA 4.0

Stratified Sampling

Stratified sampling is also a type of survey sampling, where the population is organized into a number of distinct categories (called strata). Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. In order for each element to have an equal probability of being sampled, the stratums must not overlap.

In Figure 2.3, there are three strata of unequal size. Each strata is organized according to color: one white stratum, one black stratum, and one grey stratum. The essential idea is that each strata is considered a subpopulation, usually because of some common attribute; such as race, gender, income, etc. Then within each strata a probability sample is taken proportionate to the size of the strata relative to the total population. In the example below, we still want $\boldsymbol{n} = 4$ and since the black strata has 50% of the population within it, then half of the sample will be randomly selected from that strata. Thus person ten and five have a $\frac{1}{3}$ chance of being sampled out of the first and last stratum respectively. Person two has a$\frac{1}{6}$ chance while person eight had a $\frac{1}{5}$ probability of being picked, if we are sampling without replacement.

If all the strata were the same size then a SRS would yield on average a sample, which would be representative of the population. However, when strata have varying sizes then independent probability samples must be taken from each strata. Then the overall estimates of the parameters are pooled using one of the various weighting schemes. An example is shown in the next chapter.

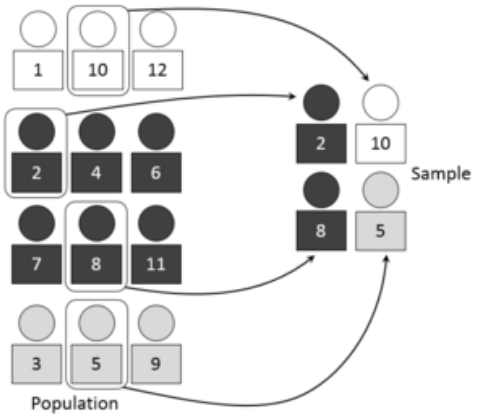

Figure 2.3. Source: Dan Kernler (2014c), licensed under CC BY SA 4.0

Cluster Sampling

Another form of sampling by groups is called cluster sampling. Stratified and cluster sampling have similarities. They are efficient to use because sometimes it is more cost-effective to select respondents in groups (strata or clusters). But quite often the sampling units are "clustered" by geography, location, density, etc. For instance in surveying households within a city, we might choose to select a certain number of city blocks (for convenience) and then interview every household within the selected blocks. Here the city blocks are the clusters of households. These city blocks can be thought of as strata, but strata usually are different with respect to some other characteristic of the population, such as ethnicity. A certain area of the city comprising different city blocks may be predominantly Hispanic, where other areas may be predominantly white. In order to have representation, the sampling design might first stratify the city by ethnicity and then within each strata would be clusters of city blocks. Then the sampling would take on various stages (multistage sampling).

Cluster sampling is commonly implemented as a multistage sampling process. Frongillo (1996) describes a more complex form of cluster sampling. The first stage consists of constructing the clusters that will be sampled (in some cases the clusters to be sampled are chosen for convenience and not selected using probability sampling). In the second stage, a sample of units is randomly selected within each cluster (rather than using all units contained in all selected clusters). In the following stage(s), in each of those selected clusters, random samples of units are selected. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed. This technique is essentially the process of taking random subsamples of preceding random samples. Various national government surveys have this multistage process of selecting individuals from across the U.S. For example, the country is clustered first by density or some other characteristic (such as whether or not that cluster has voted Democratic, etc. in the past) and within each cluster there may be certain strata (defined maybe by median household income). Within these strata, there may be further clusters of the city blocks, etc.

Figure 2.4 shows a simple population of 12 people arranged in six different clusters. Now each of these clusters contains the same number of people, however this is not always the case. Some clusters just like strata can have unequal sizes. The sample in Figure 2.4 contains six clusters and in this design, once a cluster is taken, then every unit within that cluster is in the final sample.

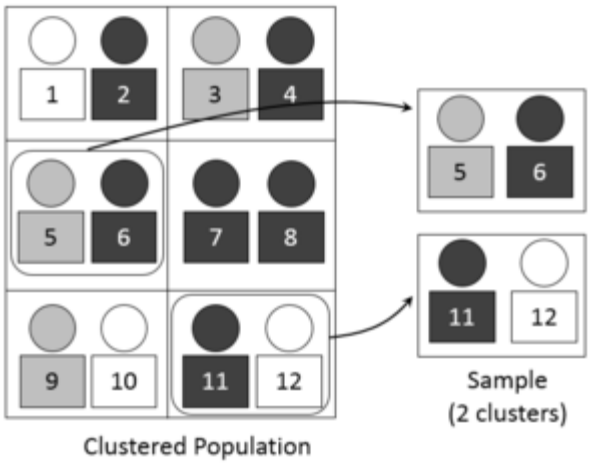

Figure 2.4. Source: Dan Kernler (2014a), licensed under CC BY SA 4.0

As mentioned previously, in stratified and cluster the individual sampling elements may not have an equal probability of being sampled. In this case the estimates are usually weighted, if the sample design does not give each individual an equal chance of being selected. For instance, when households have equal selection probabilities but one person is interviewed from within each household, this gives people from large households a smaller chance of being interviewed. This can be accounted for using survey weights. Similarly, households with more than one telephone line have a greater chance of being selected in a random digit dialing sample, and weights can adjust for this. Weights can also serve other purposes, such as helping to correct for non-response.

Complex Survey sampling

In simple surveys each observation has the same probability of being chosen. However, there are some cases in which each observation is disproportionately chosen. In order to have the

sample represent the population, information on sampling proportions (called sample weight) is required to properly analyze the survey. In large government national surveys, the sampling process consists of multiple stages of sampling; and at every stage except the last stage, clusters of observations are sampled. At the final stage, the individual observations are sampled. These types of designs where the selection is through a multistage process and in some cases may include "oversampling" in order to obtain reliable estimates are known as "complex surveys".

A simple example of a complex survey could be where school children are to be sampled. Then the first sampling stage would be to randomly select the schools; then next the classrooms within the schools, and finally the children within classrooms. The first stage of sampling produces what are called the "primary sampling units (PSUs)." This type of sampling is often required because it is logistically impossible, difficult, or expensive to simply try and design a SRS of children directly. The use of multi-stage cluster sampling means that observations cannot be assumed to be independent as is commonly done for a SRS. Observations that are from the same cluster will likely be more similar to each other than to observations from a different cluster (Frongillo, 1996).

Chapter 3 Numerical Examples Showing the Variance Estimation

The major goal of sampling is to obtain estimates of population parameters that are as precise as possible, correct and free of bias. We would want to obtain these estimates without actually sampling the entire population and from only sampling small proportions of the population. Since the estimates are based on random samples, they themselves are considered random variables and thus can be described using probability functions. The precision of the estimates while certainly is a function of the sample size, it is also related to the sampling design. As the sample design becomes more intricate, the estimation of the variance also becomes more detailed. This chapter shows examples (Kish, 1965) of how to calculate the variance of the sample mean for each of the designs. In general, the precision of an estimate is often described using 95% confidence intervals (95% CI). The width of the confidence intervals is related to the variance of the estimate. Intervals that are wider show less precision, hence it is important to properly estimate the variance.

Simple Random Sampling

The simple mean of the sample of a SRS selection is the SRS mean, and we distinguish it with the subscript 0;

$$\overline{y_0} = \frac{y}{n} = \frac{1}{n}\sum_j^n y_j$$

Simple random sampling is a sample design specifying both the srs selection and the simple mean estimate. The variance of the srs mean $\overline{y_0}$ is computed as

$$\text{Var}\,(\overline{y_0}) = (1 - f)\ \frac{s^2}{n},$$

$$\text{Where} s^2 = \frac{1}{n-1}\sum_j^n (y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_j^n y_j{}^2 - \frac{y^2}{n}\right] = \frac{n\sum^n y_j{}^2 - y^2}{n\,(n-1)}$$

The standard error of $\overline{y_0}$ is the square root of its variance;

$$\text{se}\,(\overline{y_0}) = \sqrt{\text{Var}\,(\overline{y_0})} = \sqrt{(1-f)}\,\frac{s}{\sqrt{n}}$$

The following SRS (Table 3.1) is from a list of city blocks in Ward 1 in Massachusetts (1950 U.S. Census). The purpose of the survey was to estimate the proportion of dwellings actually owned by the dweller. The actual city blocks are manually arranged in 27 groups of 10 blocks each. The x_i represents the number of dwellings, and the y_i denote dwellings occupied by renters. The population of N = 270 blocks was numbered in the actual problem from 232 to 772 (with block listing numbers that are close together actually representing close proximity to one another within the city). The arranging of the blocks into groups was of 10 was done for convenience. It was estimated that each interviewer could cover about 10-city blocks (or one group) in a day. In this example, with three-digit random numbers, the following SRS of n = 20 was selected. Thus the unit of sampling is the block and in a SRS, every block has an equal chance of being selected. So n /N $= f$ = 20/270 = .074. This example will be used again (Ch 6) to show the process of estimating the variance for a ratio, such as (y_i / x_i) but this example shows how to calculate the variance of both X and Y individually. An issue with this design is the number of dwellings within each group is not the same. But according to the sampling design every dwelling within each block is sampled once the group was selected.

j Sample No.	Block Listing Numbers	x_i	y_i
1	689	5	3
2	537	9	5
3	545	18	5
4	420	68	52
5	436	32	21
6	385	48	34
7	575	11	3
8	727	1	0
9	753	1	0
10	451	4	0
11	701	29	17
12	566	31	14
13	680	5	0
14	735	2	0
15	528	4	2
16	541	102	54
17	564	20	11
18	380	15	11
19	730	1	0
20	376	29	23

Table 3.1 Sample data from an SRS

The summations of *x* and *y* are

$\sum x_j$ = 435 $\sum^n x_j{}^2$ = 22,239 $\sum y_j$= 255 $\sum^n y_j{}^2$ = 8545

The sample means for the two variances are respectively,

$\overline{y_0} = \frac{y}{n} = \frac{255}{20}$ = 12.75 and $\overline{x_0} = \frac{x}{n} = \frac{435}{20}$ = 21.75.

The sample variances of the two means are

Var ($\overline{y_0}$) = (1 - $\frac{20}{270}$) $\frac{278.62}{20}$ = 12.90 , where $s_y{}^2 = \frac{1}{19}$ [8545 - $\frac{255^2}{20}$] = 278.62 ,

and

Var ($\overline{x_0}$) = (1 - $\frac{20}{270}$) $\frac{672.51}{20}$ = 31.14 , where $s_x{}^2 = \frac{1}{19}$ [22,239 - $\frac{435^2}{20}$] = 672.51 ,

and the standard deviations are

$s_y = \sqrt{278.62}$ = 16.7 and $s_x = \sqrt{672.51}$ = 25.9.

The standard errors of the means are

Se ($\overline{y_0}$) = $\sqrt{12.90}$ = 3.59 and Se ($\overline{x_0}$) = $\sqrt{31.14}$ = 5.58

SYSTEMATIC SAMPLING

There is no design-unbiased variance estimator in systematic sampling(Were, 2015). If we are interested in the true error variance, the only way is to repeat the systematic sample and

calculate the variance of all the estimations produced. This is not a viable approach for practical implementation. What is most frequently done for variance estimation of a systematic sample is treating the sample as a simple random sample. However this form of variance estimation consistently over-estimates the true error variance. Numerous approximations have been developed to better estimate the variance of a systematic. One of these methods (successive difference model) is described below. This algorithm was selected due to the relative ease of implementation(Aune-Lundberg, 2014).

Successive Difference Model (Ordering of Collection Important)

A systematic sample selected with the interval ***k*** after a random start yields an equal probability of being selected because each element has a probability of 1/k being selected. Using this sampling, the sample mean is considered an unbiased estimate of the population mean if the sample size was fixed at ***n***. If ***n*** is not fixed, then the mean is not technically unbiased, but it will usually be a good estimate. Then the variance equation for this, where g indexes the units of the ordered sample, is given by

$$\text{Var}(\bar{y}) = \frac{1-f}{2n(n-1)} \sum_{g=1}^{n-1} (y_g - y_{g+1})^2.$$

Suppose a systematic sample of n = 40 city blocks out of a population of N = 4000 city blocks results in the following sample, each number represents the number of houses on sampled block, **<u>presented in the order they were drawn</u>**:

10, 8, 6, 5, 9, 8, 8, 5, 9, 9, 9, 10, 4, 3, 1, 2, 3, 4, 0, 6,

3, 5, 0, 0, 0, 0, 4, 0, 8, 0, 10, 5, 6, 1, 3, 3, 1, 5, 5, 4.

The mean of the sample is $\bar{y} = \frac{y}{n}$ = 185/40 = 4.625. The variance is given by

$$\text{Var}(\bar{y}) = \frac{0.99}{2 \; x \; 40 \; x \; 39} 540 = 0.171,$$

where $\sum_{g}^{n-1} (y_g - y_{g+1})^2 = (10-8)^2 + (8-6)^2 + (6-5)^2 + (5-9)^2 + \ldots\ldots\ldots + (1-5)^2 + (5-5)^2 + (5-4)^2 = 540.$

CLUSTER SAMPLING

Clusters of Equal Size

Suppose that from a population of *A* clusters, *a* sample clusters are selected with equal probability. In the selected clusters, all *B* elements are included in the sample which consists of *a X B* = *n* elements.

The sample mean of the ***n*** elements in the sample typically serves to estimate the population mean. It is also the mean of the a cluster means:

$$\bar{y} = \frac{1}{n} \sum_{j}^{n} y_j.$$

The sample size here is fixed at n = a X B. Again the selection of the sample clusters may be systematic, stratified, or even a SRS. For this clustering design, the variance is

$$\text{Var}(\bar{y}) = (1 - f)\frac{S_a^2}{a}, \quad \text{where} S_a^2 = \frac{1}{a-1} \sum_{\alpha}^{a} (\bar{y}_\alpha - \bar{y})^2.$$

This formula resembles the variance of simple random sampling. In both cases, the variance of the mean is directly proportional to the variance between sampling units and

inversely proportional to the number of sampling units. The unit variances are respectively $s^2$and s_a^2 ; and the sample sizes are, respectively, *n* and *a*. Similarly, the cluster sample mean is the mean of *a* sampling units selected at random. The variance of the sample mean arises entirely from variance between the cluster means.

The form S_a^2 denotes unit variance between the cluster means $\bar{y}_\alpha$. The factor (1-f) becomes negligible for small f, and it disappears for selection with replacement, when clusters are permitted to appear more than once in the sample. But this a theoretical result and sampling with replacement is not usually done in practice.

An example taken from Kish is about a newspaper that has 39,800 subscribers served by carrier routes. There is a card for each subscriber; in a file the cards of each carrier's route are kept together in geographical order and neighboring routs follow each other. The number of cards per carrier varies between 50 and 200. The chief purpose of the survey is to find out how many of the subscribers own their homes. An interview survey of about 400 subscribers is pre-determined, in small clusters of 10 subscribers each. These save travel time, because an interviewer can generally obtain the 10 interviews in one neighborhood in a short period of time.

The sampler regards the N = 39,800 cards as a frame of A = 3980 clusters of B = 10 each. A few of the clusters will be split between two routes. He selects a = 40 different random numbers, from 1 to 3980. Each random number r denotes the selection of 10 cards numbered from (10r – 9) to 10r; e.g., number 179 will select cards 1781-1790.

The results of the 40 clusters follow (here each cluster has 10 homes within). In terms of y_α , the data represents the number of homeowners in each cluster of 10 households:

10, 8, 6, 5, 9, 8, 8, 5, 9, 9, 9, 10, 4, 3, 1, 2, 3, 4, 0, 6,

3, 5, 0, 3, 0, 0, 4, 0, 8, 0, 10, 5, 6, 1, 3, 3, 1, 5, 5, 4.

So in the first cluster all 10 of the subscribers owned their own home, similarly, in the second cluster 8 out of 10 subscribers owned their own home, and so.

We need to compute only two numbers; y = $\sum^{a} y_{\alpha}$ = 185; and$\sum^{a} {y_{\alpha}}^{2}$ = 1263. The sample mean is simply:

$$\bar{y}=\frac{y}{n}=\frac{185}{400}= 0.4625 = 46.2 \text{ percent.}$$

The estimated variance of the sample mean is;

$$\text{Var } (\bar{y}) =\frac{1-f}{a} S_a^2 = \frac{1-f}{a} \left[\frac{1}{(a-1)B^2}\left(\sum_{a=1}^{a} {y_{\alpha}}^{2} - \frac{y^2}{a}\right)\right]$$

$$= \frac{0.99}{40}\left[\frac{1}{3900} (1263 - \frac{185^2}{40})\right]$$

$$= \frac{0.99}{40} \frac{(1263-855.6)}{3900}$$

$$= 0.99 (.002603) = 0.002585.$$

The standard error is $\sqrt{0.002585}$ = 0.05084 = 5.1 percent. The total number of subscribers who own their own home is estimated as N ($\bar{y}$) 39,000 x 0.4625 = 18,408 = 18,400, with a standard error of 39,800 x 0.05084 = 2023 = 2000. The 95% confidence interval can be obtained by subtracting and adding two standard errors from the estimated population mean. (18,400 – 4,000 , 18,400 + 4,000) = (14,400 , 22,400)

Clusters of Unequal Sizes

Notation for cluster sampling is defined as:

$$y_{ij} = \text{measurement for jth element in the } i^{\text{th}} \text{ psu}$$

N = number of psus (primary sampling unit sample) in the population

$$M_i = \text{number of ssus (secondar sampling unit sample) in psu } i$$

$M_O = \sum_{i=1}^{N} M_i$ = total number of ssus in the population

$t_i = \sum_{j=1}^{M_i} y_{ij}$= total in psu*i*

t = population total

n = number of psus in the sample

m_i = number of ssus in the sample from psu*i*

$\bar{y}_i = \sum_j \frac{y_{ij}}{m_i}$ = sample mean (per psu) for psu*i*

$\hat{t}_i = \sum_j \frac{M_i}{m_i} y_{ij}$ = estimated total for psu*i*

$\widehat{t_{unb}} = \sum_i \frac{N}{n} \hat{t}_i$ = unbiased estimator of population total

$S_t^{\,2} = \frac{1}{n-1} \sum_i (\hat{t}_i - \frac{\widehat{t_{unb}}}{N})^2$

$S_i^{\,2} = \sum_j \frac{(y_{ij} - \bar{y}_i)^{\,2}}{m_i - 1}$ = sample variance within psu*i*

w_{ij} = sampling weight for ssu*j* in psu*i*

Clusters are rarely of equal size in social surveys. The difference between unequal and equal-sized clusters is that the variation among the individual cluster totals t_i is likely to be large when the clusters have different sizes. The investigators conducting the Enumerative Check Census of 1937 were interested in the total number of unemployed persons, and t_i would be the number of unemployed persons in postal route i. One would expect to find more persons, and hence more unemployed persons, on a postal route with a large number of households than on a postal route with a small number of households. So we would expect that t_i would be large when the psu size M_i is large, and small when M_i is small. Often, then ${S_i}^2$ is larger in a cluster sample when the psus have unequal sizes than when the psus all have the same number of ssus.

The probability that a psu is in the sample is n/N, as a srs of ***n*** of N psus is taken. Since one-stage cluster sampling is used, an ssu is included in the sample when its psu is included in the sample. Thus

$$w_{ij} = \frac{1}{P(ssu\,j\,of\,psu\,i\,is\,in\,sample)} = \frac{N}{n}.$$

One-stage cluster sampling produces a self-weighting sample when the psus are selected with equal probabilities. Using the weights gives us

$$\widehat{t_{unb}} = \sum_i \sum_j w_{ij}\, y_{ij}.$$

We can use the two formulas above to derive an unbiased estimator for $\overline{y_U}$ and to find its standard error. Defined

$$M_O = \sum_{i=1}^{N} M_i.$$

As the total number of ssus in the population; then $\widehat{\overline{Y_{unb}}} = \widehat{t_{unb}}/ M_O$ and SE ($\widehat{\overline{Y_{unb}}}$) = SE ($\widehat{t_{unb}}$)/ M_O. The unbiased estimator of the mean $\widehat{\overline{Y_{unb}}}$ can be inefficient when the values of M_i are unequal since it, like $\widehat{t_{unb}}$, depends on the variability of the cluster totals t_i. It also requires the$M_O$be known; however, we often know M_i only for the sampled clusters. In the Enumerative Check Census, for example, the number of households on a postal route would only be ascertained for the postal routes actually chosen to be in the sample.

STRATIFIED SAMPLING

The weighted mean and its variance

First we examine the properties of a weighted mean as a general concept. This will allow us to develop later in detail the special formulas appropriate to diverse methods of stratified sampling. We want to develop a sample estimate for a weighted population mean, $\overline{y_w}$;

$$\overline{y_w}=\sum w_h \overline{y_h} = w_1\overline{y_1} + w_2\overline{y_2} + \ldots + w_h\overline{y_h} + w_H\overline{y_H}.$$

That is, the population mean is equal to the sum of the H strata means $\overline{y_h}$, each multiplied by its proper weight w_h , where $\sum w_h$ = 1. The weighted sample mean is

$$\overline{y_w} = \sum w_h \overline{y_h}.$$

The sample mean is obtained separately and independently for each stratum, and it is then multiplied by the weight of the stratum. These products are summed over the H strata to obtain the weighted sample mean. The variance of this weighted mean is obtained by combining the

separate variances of the stratum means; The variance of each stratum mean is multiplied by the square of the stratum weight and the products are added over the H strata;

$$\text{var}(\overline{y_w}) = \sum w_h^2 \text{var}(\overline{y_h}).$$

A sample has to be taken in each stratum to estimate its stratum mean, and the sample from each stratum must contain at least two sampling units to permit the computation of the variance in the stratum. The processes of selection and estimation are performed separately and independently within each stratum. Note that nothing was said in this discussion about the sample design within the strata. Thus, within the several strata, different sampling fractions may be used, as well as different methods of selection, estimation and observation.

The sum of the elements contained in all H strata equals the totality of the n elements in the entire population, because each element of the population occurs in one, and only one, stratum;

$$N = \sum N_h = N_1 + N_2 + \ldots + N_h + \ldots + N_H.$$

The number of elements selected into the sample from the h^{th} stratum is denoted by n_h. The number of elements in the entire sample is

$$n = \sum n_h = n_1 + n_2 + \ldots + n_h + \ldots + n_H.$$

Typically, $\overline{y_h}$ is the mean of the N_h elements in the h^{th} stratum; that is,

$$\overline{y_h} = \frac{1}{N_h} \sum_i^{N_h} Y_{hi} = \frac{Y_h}{N_h},$$

Where Y_{hi} is the value of the i^{th} element in the h^{th} stratum, and Y_h is their sum in the h^{th} stratum. The weights frequently, but not always, represent the proportions of the population elements in

the strata and $W_h = N_h$/N. Then the weighted mean is equal to the ordinary mean per element of the population:

$$\overline{Y_W} = \sum \frac{N_h}{N} = \overline{Y_h} = \frac{1}{N} \sum Y_h = \frac{Y}{N} = \overline{Y},$$

And then

$$\sum W_h = \sum \frac{N_h}{N} = \frac{1}{N} \sum N_h = \frac{N}{N} = 1.$$

The weight W_h of the stratum is generally, the proportion of the population contained in the stratum, and so $\sum W_h = 1$. This can be reinforced now by permitting N_h to be any arbitrary measure of the size of the stratum, no longer restricting it to a count of elements. If we denote with N = $\sum N_h$ the sum of these arbitrary measures, $W_h = N_h$/ N we obtain

$$\overline{Y_W} = \frac{1}{N} \sum N_h \overline{Y_h}.$$

and

$$\text{var} (\overline{Y_W}) = \frac{1}{N^2} N_h{}^2 \text{var} (\overline{Y_h}).$$

We now treat a basic class of stratified samples: those with random selections of elements within each stratum. They are a sample of elements because the elements are selected individually and separately, rather than in clusters. They are stratified because the selection is carried on separate and independently within each stratum. They are random because the n_h sample elements are selected with simple random sampling. In this section we present general fundamentals and formulas which can be used for any stratified random sample of elements. They can be used for both disproportionate designs and for proportionate samples. The simple mean of the elements in the h^{th} stratum is

$$\overline{Y_{h0}} = \frac{1}{n_h}\sum_i^{nh} Y_{hi}$$

This is the mean of theh^{th} stratum, and selected with simple random sample, as the subscript 0 denotes. For combining the different strata we use the previous mean formula and obtain for the mean of any stratified random sample of elements;

$$\overline{Y_{w0}} = \sum_h^H W_h \overline{Y_{h0}} = \sum_h^H W_h \frac{1}{n_h}\sum_i^{nh} Y_{hi}$$

The variance of the simple random sample of n_h elements in the h^{th} stratum is

$$\text{Var}\,(\overline{Y_{h0}}) = (1 - f_h)\frac{{S_h}^2}{n_h}, \text{ where } {S_h}^2 = \frac{1}{n_{h-1}}\left(\sum_i^{n_h} {Y_{hi}}^2 - \frac{{Y_h}^2}{n_h}\right).$$

We combine the variances of the stratum means and obtain the variance of the sample mean $\overline{Y_{w0}}$ as

$$\text{Var}\,(\overline{Y_{w0}}) = \sum {W_h}^2 (1 - f_h)\frac{{S_h}^2}{n_h}$$

If the weights are based on the proportions of population elements in the strata, then (since W_h = N_h/N and $n_h = f_h N_h$) the above equations can also be expressed, respectively, as

$$\overline{Y_{w0}} = \frac{1}{N}\sum_h^H N_h \frac{1}{n_h}\sum_i^{n_h} Y_{hi} = \frac{1}{N}\sum_h^H \frac{1}{f_h}\sum_i^{nh} Y_{hi}$$

$$\text{Var}\,(\overline{Y_{w0}}) = \frac{1}{N^2}\sum_h^H (1 - f_h)\frac{{N_h}^2}{n_h}{S_h}^2, \text{ where } {S_h}^2 = \frac{1}{N^2}\sum_h^H \frac{1-f_h}{f_h} N_h {S_h}^2.$$

If the $\overline{Y_{w0}}$ denotes a proportion $P_{w0} = \sum W_h P_h$its variance may be written in an equivalent form which is easier to compute;

$$\text{Var}(P_{w0}) = \sum W_h^2 \left(1 - f_h\right) \frac{P_h(1-P_h)}{n_h - 1}.$$

For a numerical illustration and using the data from Table 3.2,suppose we select a sample of n = 400 employees out of N = 10,000 employed in a factory. We use the several departments of the factory as strata, because it can be done easily, and because we suspect that there may be large differences in the employee's responses among the departments for the survey variables.

	Symbol	Assembly	Foundry and Machine	Office and Miscellaneous	Entire Factory
Stratum Number	h	1	2	3	Total
Population Size	N_h	5000	3010	1990	10,000
Stratum Weight	W_h	0.50	0.30	0.20	1.00
Sample Size	n_h	200	120	80	400
Number of "yes" answers	y_h	14	18	48	80
The percentage of "yes" answers	$\overline{y_h}$	7%	15%	60%	20

Table 3.2

$$P_{w0} = \sum W_h \overline{Y_{ho}} = 0.50(7) + 0.30(15) + 0.20(60) = 20 \text{ percent}$$

And the variance of that mean is

$$\text{Var}(P_{w0}) = [0.50^2 \frac{24}{25}\frac{7\,(93)}{199} + 0.30^2 \frac{24}{25}\frac{15(85)}{119} + 0.20^2 \frac{24}{25}\frac{60(40)}{79}\,] \text{ x } 0.001$$

$$= 0.000288$$

The standard error of the $P_w = 0.20$ is

$$\sqrt{0.000288} = 0.017 = 1.7 \text{ percent}$$

Chapter 4 Description of current methods used in Variance Estimation

For the purpose of this thesis Jackknife and Bootstrap are two resampling techniques that were used to verify the precision of the sample statistics (mean, median, variance, and percentile) computed for data sets. Later a comparison of the two techniques will examine if Bootstrap or Jackknife was more accurate at estimating such statistics. Bootstrap and Jackknife execute two different methods to establish estimates but both techniques develop large amounts of subsamples from the sample. Lohr explains how, resampling methods treat the sample as if it were itself a population: they take different samples from this new "population" and use the subsamples to estimate the variance." There are several resampling techniques available but his paper will focus on the Bootstrapping and Jackknife methods.

The use of high powered computer software has made the Jackknife and Bootstrap procedures very popular. Shao and Tu further describe the frequent use of both methods, "The jackknife and bootstrap are the most popular data-resampling methods used in statistical analysis. These resampling methods replace theoretical derivations required in applying traditional methods (such as substitution and linearization) in statistical analysis by repeatedly resampling the original data and making inferences from the resamples. Because of the availability of inexpensive and fast computing software, these computer-intensive methods have caught on very rapidly in recent years and are particularly appreciated by applied statisticians."

The Jackknife method was introduced by Quenouille in 1956. Quenouille's original purpose for the Jackknife was to reduce the bias of sample estimates. It was not until 1958 when Tukey proposed using it to estimate variances and calculate confidence intervals. The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset

and calculating the estimate based on the rest of the observations and then finding the average of these calculations. Each data point goes through the process of being deleted. If one observation is deleted at a time then a sample of 1,000 will have 1,000 subsample of the sample. Therefore every time Jackknife is run on a data set the exact same result will occur.

Some advantages of the Jackknife include working for multistage samples in which other resampling techniques do not. As mentioned previously, it provides a consistent estimator of the mean and variance and tends to be asymptotically true (Rao and Shao, 1992). However the Jackknife method requires intensive computation that must be done with computer software. Most data sets are too large and time consuming for hand calculation and require running computer programs. Furthermore, this method assumes independence between the random variables (and identically distributed data points), and if that assumption is violated, the results will be of no use. Thus the variables must be independent from each other for the Jackknife estimates to be accurate.

Using the cluster example from chapter 3, I will demonstrate how R can generate Jackknife estimates of the mean and variance of the 39,000 newspaper subscribers. Table 4.1 contains the R code for the first few steps of the Jackknife example. The bootstrap package in R was installed for analysis. The first step is to input the data named *ownhome* and call the package bootstrap which contains the function jackknife.

INPUT
ownhome<-c(10, 8, 6, 5, 9, 8, 8, 5, 9, 9, 9, 10, 4, 3, 1, 2, 3, 4, 0, 6, 3, 5, 0, 3, 0, 0, 4, 0, 8, 0, 10, 5, 6, 1, 3, 3, 1, 5, 5, 4)
library(bootstrap)

Table 4.1

The following code in table 4.2 jackknifes the sum of the forty clusters. Forty subsamples were generated because one cluster was deleted at a time and then the rest of the data was totaled. This process was written as a function called theta, the results of the function are jackknifed and the vector of subsamples are called results$jack.values. The mean of results$jack.values is 180.375.

INPUT	OUTPUT
theta<- function(ownhome){sum(ownhome)}	
results<-jackknife(ownhome,theta)	
results$jack.values	[1] 175 177 179 180 176 177 177 180 176 176 175 181 182 184 183 182 181 185 179 [20] 182 180 185 182 185 185 181 185 177 185 175 180 179 184 182 182 184 180 180 [39] 181 176
mean(results$jack.values)	[1] 180.375

Table 4.2

Finding the sample mean is done by taking

$$\bar{y} = \frac{y}{n} = \frac{180.375}{400} = 0.4509 = 45 \text{ percent.}$$

The estimated variance of the sample mean

$$\text{Var}\,(\bar{y}) = \frac{1-f}{a} S_a^2 = \frac{1-f}{a}\left[\frac{1}{(a-1)B^2}\left(\sum^a {y_\alpha}^2 - \frac{y^2}{a}\right)\right]$$

$$= \frac{0.99}{40}\left[\frac{1}{3900}\left(1263 - \frac{180.375^2}{40}\right)\right]$$

$$= \frac{0.99}{40}\frac{(1263-813.37)}{3900}$$

$$= 0.02475\ (.11539) = 0.002855.$$

The standard error is $\sqrt{0.002855}$ = 0.05344 = 5.3 percent. The total number of subscribers who own their own home is estimated as N($\bar{y}$) 39,800 x 0.4509 = 17,945 with a standard error of 39,800 x 0.05344 = 2127. The 95% confidence interval can be obtained by subtracting and adding two standard errors from the estimated population mean. (17,945 – 4,254 , 17,945 + 4,254) = (13,691 , 22,199).

Bootstrap is the other resampling technique utilized in this research. The Bootstrap is a method of resampling with replacement where as the Jackknife is a resampling method without replacement. It was introduced as a spinoff of the Jackknife by Bradley Efron in 1979. The bootstrap subsamples are pulled from the sample of size***n***. Assuming that ***n*** is sufficiently large there is zero probability the resamples are identical to the original sample. The sampling processes is repeated a large number of times e.g. 10,000. For each of the bootstrap samples the mean is calculated (or any estimator) and called the bootstrap estimates. We now have 10,000

subsamples of the sample mean and can make a histogram of bootstrap means. This provides an estimate of the shape of the distribution of the mean from which we can answer questions about how much the mean varies. The method here, described for the mean, can be applied to almost any other statistic or estimator (median, percentile, or variance).

The popularity of the bootstrap method is attributed to its simplicity to derive estimates for complex data sets. Bootstrap is also an appropriate way to control and check the stability of the results. Histograms and other plots provide much information as to the variability of each statistic. For most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance. In some conditions bootstrapping is asymptotically consistent, but it does not provide general finite-sample guarantees. The apparent simplicity may conceal the fact that important assumptions are being made when undertaking the bootstrap analysis. Similar to the Jackknife method the Bootstrap assumes data independence.

Once again using the newspaper subscriber example from Chapter 3, R will be used to compile bootstrap estimates to find the variance of the sample mean. Table 4.3 shows the steps needed for bootstrap analysis and is similar to the Jackknife code. The first step is to input the data named ownhome. Then a function called *samplesum* was written that returns the sum of the *ownhome* data. This function is bootstrapped 1,000 times using the command boot from R.

INPUT
ownhome<-c(10, 8, 6, 5, 9, 8, 8, 5, 9, 9, 10, 4, 3, 1, 2, 3, 4, 0, 6, 3, 5, 0, 3, 0, 0, 4, 0, 8, 0, 10, 5, 6, 1, 3, 3, 1, 5, 5, 4, 9)
samplesum<- function(ownhome,i) {
d2 <- ownhome[i]
return(sum(d2)) }
b = boot(ownhome, samplesum, R=1000)

Table 4.3

The mean of the 1,000 subsamples is found by calling b, which is shown in table 4.4 as 185.46

INPUT	OUTPUT
b	Bootstrap Statistics : original bias std. error t1* 185 0.462 20.90807
mean(b$t)	[1] 185.462

Table 4.4

Below in figure 4.1 and 4.2 is a histogram and normal quantile plot of the 1,000 bootstrap samples. The dotted line at 185.46 in the center of the histogram is the mean of the 1,000 bootstrap samples.

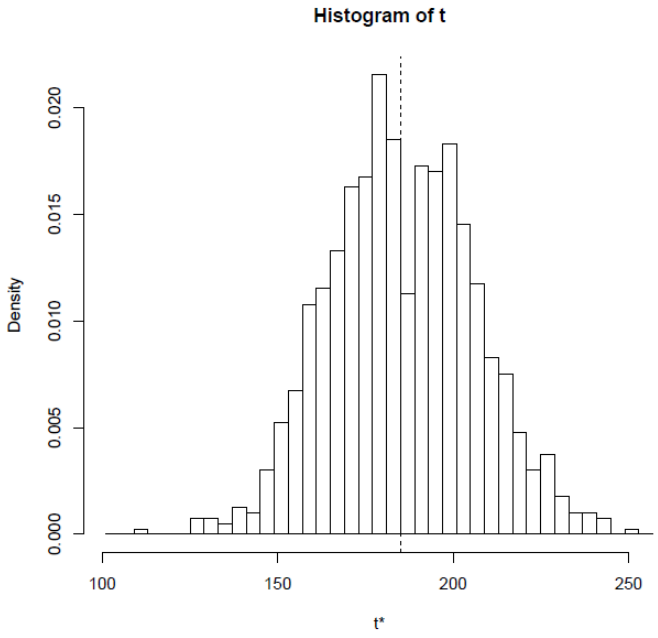

Figure4.1

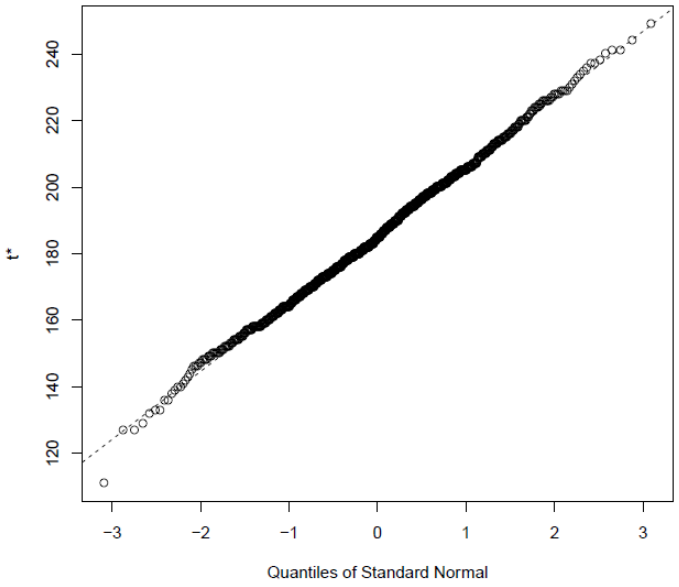

Figure 4.2

Finding the sample mean is done by taking

$$\bar{y} = \frac{y}{n} = \frac{185.462}{400} = 0.4635 = 46.35 \text{ percent.}$$

The estimated variance of the sample mean is:

$$\text{Var}(\bar{y}) = \frac{1-f}{a} S_a^2 = \frac{1-f}{a} \left[\frac{1}{(a-1)B^2} \left(\sum^a {y_\alpha}^2 - \frac{y^2}{a} \right) \right]$$

$$= \frac{0.99}{40} \left[\frac{1}{3900} \left(1263 - \frac{185.462^2}{40}\right) \right]$$

$$= \frac{0.99}{40} \frac{(1263 - 859.9038)}{3900}$$

$$= 0.02475\ (.103358) = 0.002558.$$

The standard error is $\sqrt{0.002558}$ = 0.05057 = 5.05 percent. The total number of subscribers who own their own home is estimated as N($\bar{y}$) 39,800 x 0.4635 = 18,560 with a standard error of 39,800 x 0.05057 = 2012. The 95% confidence interval can be obtained bysubtracting and adding two standard errors from the estimated population mean. (18,560 – 4,025, 18,560 + 4,025) = (14,535 , 22,585).

Chapter5 Complex Surveys

Most large surveys involve several of the concepts previously mentioned. The survey may be stratified with several stages of clustering and quite often employs multiple probability schemes in order to obtain samples that are "rich" in information. Further information on sampling proportions called sample weights is required in order to properly analyze a survey of this nature. Often the analysis relies on various levels of regression estimations. Surveys combining the components of random sampling, stratification, clustering, and sophisticated regression type analyses are referred to as "complex surveys".

Complex surveys have multiple stages of sampling. At every stage except the lowest stage, clusters of observations are sampled. At the lowest stage, the individual observations are sampled. Frongillo gives an example: a survey of school children may be collected by first sampling the schools within some region; then sample the classrooms within each school and finally sample the children within those classrooms. This type of sampling is often required because it is logistically impossible, difficult, or expensive to sample the students directly. The formulas for estimating the standard errors are complicated and require specialized computer intensive software packages.

There are many datasets obtained from various complex surveys that are available online. All contain various de-identified information on the subjects. In order that subjects cannot be identified, variables such as age are recoded into categories; similarly rather than give a subject's address, their city or county may be recoded and only those codes are given in the survey. Depending on the purpose of the survey, sensitive information is not downloadable but may be obtained from the survey collection agencies with specific IRB approval.

The U.S. government spends millions of dollars each year collecting survey data from the U.S. population. These datasets are organized coded files that can be obtained with specific statistical software. They are freely available and can be downloaded from the web. While each survey has its objectives, the government collects this data with really no primary purpose. Their intention is for researchers to have access to population data for getting answers to health related questions that the investigator has come up with. Consequently these datasets are referred to as "secondary data files." To see the multitude of data and listings of these files, all one has to do is to Google www.cdc.gov/nchs. All files are identified by their acronym name such as NHIS (National Health Interview Survey), NSFG (National Survey of Family Growth), NHANES (National Health and Nutritional Exam Survey), NHDS (National Hospital Discharge Survey), etc. Many of these surveys are performed yearly. The NHIS is the major survey that the government uses to obtain information on the health of the population and study changes from year to year. The NHANES survey is an extensive survey to estimate what the population is eating, their exercise and disease prevalence and other areas as well. A short description of two surveys (NHIS and NHANES) follows.

NHIS (National Health Interview Survey)

The National Health Interview Survey (NHIS) provides vital information on the health of the civilian noninstitutionalized population of the United States and is one of the major data collection programs of the National Center for Health Statistics (NCHS) which is part of the Centers for Disease Control and Prevention (CDC). NHIS data are used widely throughout the Department of Health and Human Services (DHHS) to monitor trends in illness and disability and to track progress toward achieving national health objectives. The data are also used by the

public health research community for epidemiologic and policy analysis of such timely issues as characterizing those with various health problems, determining barriers to accessing and using appropriate health care, and evaluating Federal health programs.

The National Health Interview Survey is a cross-sectional household interview survey. Sampling and interviewing are continuous throughout each year. The sampling plan follows a multistage area probability design that permits the representative sampling of households and noninstitutional group quarters (e.g., college dormitories). The sampling plan is redesigned after every decennial census. The first stage of the current sampling plan consists of a sample of 428 primary sampling units (PSU's) drawn from approximately 1,900 geographically defined PSU's that cover the 50 States and the District of Columbia. A PSU consists of a county, a small group of contiguous counties, or a metropolitan statistical area.

Within a PSU, two types of second-stage units are used: area segments and permit segments. Area segments are defined geographically and contain an expected eight, twelve, or sixteen addresses. Permit segments cover housing units built after the 2000 census. The permit segments are defined using updated lists of building permits issued in the PSU since 2000 and contain an expected four addresses. The NHIS sample is drawn from each State and the District of Columbia. The total NHIS sample is subdivided into four separate panels, or subdesigns, such that each panel is a representative sample of the U.S. population. The expected NHIS sample size (completed interviews) is approximately 35,000 households containing about 87,500 persons. The annual response rate of NHIS is close to 90 percent of the eligible households in the sample.

Data are collected through a personal household interview conducted by interviewers employed and trained by the U.S. Bureau of the Census according to procedures specified by the

NCHS. For the Family Core component, all adult members of the household 17 years of age and over who are at home at the time of the interview are invited to participate and to respond for themselves. For children and for adults not at home during the interview, information can be provided by a responsible adult family member, 18 years of age and over, residing in the household.

NHANES (National Health and Nutritional Exam Survey)

The two complex surveys utilized in this thesis are the National Health Interview Survey from 2012 (NHANES 2012) and the National Health Interview Survey II (NHANES II). This survey is one of the major data collection programs of the National Center for Health Statistics (NCHS) which is part of the Centers for Disease Control and Prevention (CDC). NHANES is arguably the largest and longest-running national source of objectively measured health and nutrition data. The survey is unique in that it combines interviews and physical examinations to collect data. Through the physical examinations, clinical and laboratory tests, and personal interviews, NHANES provides a "snapshot" of the health and nutritional status of the U.S. population. Findings from NHANES provide health professionals and policymakers with the statistical data needed to determine rates of major diseases and health conditions (e.g., cardiovascular disease, diabetes, obesity, infectious diseases) as well as identify and monitor trends in medical conditions, risk factors, and emerging public health issues, so that the appropriate public health policies and prevention interventions can be developed.

Five NHANES have been conducted since 1970. NHANES I, the first cycle of the NHANES studies, was conducted between 1971 and 1975 and included a national sample of

approximately 30,000 individuals between one and seventy-four years of age. NHANES II (1976–1980) included just slightly over 25,000 participants and expanded the age of the first NHANES sample somewhat by including individuals as young as 6 months of age. In addition, children and adults living at or below the poverty level were sampled at higher rates than their proportions in the general population ("oversampled") because these individuals were thought to be at particular nutritional risk.

The second National Health and Nutrition Examination Survey, NHANES II, is a nationwide probability sample of 27,801 persons from 6 months to 74 years of age. From this sample, 25,286 people were interviewed and 20,322 people were examined, resulting in an overall response rate of 73 percent. Because children and persons classified as living at or below the poverty level were assumed to be at special risk of having nutritional problems, they were sampled at rates substantially higher than their proportions in the general population. Adjusted sampling weights were computed within 76 age-sex income groups in order to inflate the sample to closely reflect the target population at the midpoint of the survey.

Beginning in 1999, NHANES became a "continuous survey." That is, unlike the previous NHANES surveys, which were conducted over a period of approximately four years with a "break" of at least one year between survey periods, the 1999–2000 survey was (and all subsequent surveys will be) conducted without breaks, on a yearly basis. As the survey period is shorter in length, the subject sample will be smaller. The 1999–2000 survey included nutritional and medical data on approximately 8,837 individuals up to 74 years of age.

In current practice, however, NHANES data are derived primarily from the first two sources; that is, via direct interview and direct clinical examination. The NHANES data collection procedures have changed slightly over the years. These changes reflect not only the

changing demographics of the United States over time, but also the changing nature of the survey (e.g., the inclusion of the nutrition component, the interest in the effects of environment upon health). Nonetheless, the basic trends of data collection, particularly with regards to sampling, are similar.

The primary sample design change for NHANES 2011-2012 is that there is an oversample of Non-Hispanic Asians in addition to the ongoing oversample of Hispanics, non-Hispanic Blacks, older adults, and low income whites/others. Since the total sample size in any year is fixed due to operational constraints, sample sizes for Hispanic persons and non-low income white and other persons were decreased in order to increase the sample sizes for Asians. Consequently, sample sizes for Mexican American Hispanic persons were also decreased compared to survey cycles prior to 2011.

NHANES uses a complex, multistage probability design to sample the civilian, noninstitutionalized population residing in the 50 states and D.C. Sample selection for NHANES followed these stages, in order:

1. Selection of primary sampling units (PSUs), which are counties or small groups of contiguous counties.

2. Selection of segments within PSUs that constitute a block or group of blocks containing a cluster of households.

3. Selection of specific households within segments.

4. Selection of individuals within a household.

In 2011-2012, 13,431 persons were selected for NHANES from 30 different study locations. Of those selected, 9,756 completed the interview and 9,338 were examined. Data are collected through a personal household interview conducted by interviewers employed and trained by the U.S. Bureau of the Census. Health interviews are conducted in respondents' homes. Health measurements are performed in specially-designed and equipped mobile centers, which travel to locations throughout the country. The study team consists of a physician, medical and health technicians, as well as dietary and health interviewers.

Chapter 6 Statistical Results from the Complex Surveys

The following tables in this chapter show various results obtained using R and STATA for the ratio estimations. Tables 6.1-6.4 show the frequency distribution of two variables before they are weighted (as a sample) then after applying the weights, which produce estimates of the population. Tables 6.5 and 6.6 compare various estimates of these variables weighted and unweighted. Following, the same estimates and variables are compared however utilizing the resampling techniques bootstrap and jackknife. The last two tables (6.7 and 6.8) use the software STATA to create and analyze a ratio variable. In survey data, it is very common to create ratio statistics. STATA is a computer program that allows proper analysis of a ratio variable.

Frequency Distributions and Extrapolations to U.S. Population

Table 6.1 is a frequency table of the variable DMDHHSIZ from the 2012 NAHANES data. DMDHHSIZ identifies the total number of people in a household. For example 2176 people in the sample have 4 people in their household. The number of people in a home ranges from 1 to 7. Homes that contained more than 7 people were coded as 7. Notice the frequencies of the number of household member's total 9756, which is the number of people sample.

# of People in House	1	2	3	4	5	6	7	Total
# of People	797	1891	1748	2176	1533	796	820	n= 9756

Table 6.1

Table 6.2 is a frequency table of the variable RIAGENDR, which tells the amount of males versus females in the sample of 9756. Looking at the figure one can see there are 4856 males and 4900 females in the sample of the NHANES data.

Gender	Male	Female	Total
# of People	4856	4900	n= 9756

Table 6.2

In order for the weights to be correctly applied to the data, the following R code must identify the variable that contains the weights as well as the survey design. The package survey needs to be loaded before hand in order to read the new survey design.

library(survey)

nhanes<-svydesign(id=~SDMVPSU, strat=~SDMVSTRA, weights=~WTMEC2YR, nest=TRUE, data=demo)

Tables 6.3 and 6.4 are frequency tables of the variable DMDHHSIZ and RIAGENDR respectively. These frequency tables represent DMDHHSIZ and RIAGENDR after the weights

have been applied. Therefore we can now look at the entire population of the U.S. rather than just a sample. Notice both tables total 306,590,681, which is the population of the United States in 2012. For example 66,729,892 people in the United States live with four people in their home and there are 149,634,950 males and 156,955,731 females in the United States.

# of People in House	1	2	3	4	5	6	7	Total
# of People	29647018	82094975	55164152	66729892	38925817	17956560	16072267	N= 306590681

Table 6.3

Gender	Male	Female	Total
# of People	149634950	156955731	N= 306590681

Table 6.4

Comparison of Statistical Estimates between Unweighted (Raw Data) and Weighted (Complex Design)

Table 6.5 is a comparison of the unweighted DMDHHSIZ, RIDAGEYR, and INDFMPIR variables and the weighted version of them. As mentioned previously DMDHHSIZ is the number of people in a household, RIDAGYR is the age of the participants at the time of the survey, and INDFMPIR is a ratio of household income to poverty guidelines. The mean, median, variance, and 95% confidence interval of the mean for each variable is reported for the unweighted and weighted version. Due to the fact that 5.0 was reported above a certain income, the variable INDFMPIR is highly skewed at 5.0. In order to get a more accurate reading, the 5.0s were removed and the mean, median, and confidence interval were recalculated. Once the 5.0s were dropped the mean dropped from 2.2 to 1.74. These calculations are denoted with *.

VARIABLE		**RAW (Unadjusted)** n= 9,756	**COMPLEX (Adjusted)** N= 306,590,681	
DMDHHSIZ Number of people in household	**Mean**	**3.76**	**3.39**	
	95% CI for mean	(3.56, 3.63)	(3.08, 3.3)	
	Median	**4**	**3**	
	Variance	**3.13**	**2.59**	
RIDAGEYR Age of participant	**Mean**	**31.4**	**37.1**	
	95% CI for mean	(30.91, 31.89)	(35.73, 38.47)	
	Median	**26**	**37**	
	Variance	**604.13**	**499.84**	
INDFMPIR Ratio of family income to poverty guidelines	**Mean**	**2.2**	**2.74**	**1.74 ***
	95% CI for mean	(2.17, 2.24)	(2.5, 2.98)	(1.72,1.77) *
	Median	**1.63**	**2.6**	**1.33 ***
	Variance	**2.68**	**2.87**	

Table 6.5

Looking at the tables above one can compare the estimates of the sample data and the population. The mean of the sample DMDHHSIZ and population DMDHHSIZ (number of people in a household) does not have a significant difference, with a sample mean of 3.76 and 95% confidence interval of (3.56, 3.63) and a population mean of 3.39 and CI of (3.08, 3.3). The median also decreased from 4 to 3 when the weights were applied. For the RIDAGEYR variable (participant age) the weights had the opposite effect of increasing the mean and median. The sample mean is 31.4 with (30.91, 31.89) confidence interval and the population mean is 37.1

with (35.73, 38.47) confidence interval. The sample median increases from 26 years in the sample to 37 years in the population and demonstrate the need for weighting. The third variable INDRMPIR (ratio of household income to poverty guidelines) had a similar result as RIDAGEYR after being weighted. The mean increases from 2.2 to 2.74 and confidence intervals of (2.17, 2.24) to (2.5, 2.98). The median also increases from the sample to the population of 1.63 to 2.6.

Comparison of Statistical Estimates from the Resampling Methods

Table 6.6 reports the mean, median, and 95% confidence interval of the three variables using the bootstrap and jackknife methods.

VARIABLE		BOOTSTRAP RAW n= 9,756	JACKKNIFE RAW n = 9,756	BOOTSTRAP COMPLEX N= 306,590,681
DMDHHSIZ # of people in household	Mean	3.76	3.76	3.396
	95% CI	(3.73, 3.80)	(3.72, 3.79)	(3.299, 3.490)
	Median	4	4	3
RIDAGEYR Age of participant	Mean	31.4	31.4	37.177
	95% CI	(30.92, 31.88)	(31.403, 31.404)	(35.884, 38.47)
	Median	26.07	26	37
INDFMPIR Ratio of family income to poverty guidelines	Mean	2.2	2.21	2.74
	95% CI	(2.17, 2.23)	(2.204, 2.205)	(2.53, 2.95)
	Median	1.63	1.63	2.6

Table 6.6

The results in the comparison table of bootstrap and jackknife are very similar to each other. The mean and median for DMDHHSIZ are identical to the sample mean and median of DMDHHSIZ in FIGURE 6.5, 3.76 and 4 respectively. The confidence intervals differ by .01 of a person, bootstrap method results in a 95% Confidence Interval of (3.73, 3.8) and jackknife (3.72, 3.79). RIDAGEYR has very close results to the sample data as well. Bootstrap produced 31.4 mean, (30.92, 31.88) CI, and 26.07 median. Jackknife gives 31.4 mean as well with a median of 25 and CI of (31.403, 31.404). The final variable INDFMPIR bootstrap resulted with a mean of 2.2, (2.17, 2.23) CI, and 1.63 median. Lastly the Jackknife gives 2.21 as the mean,(2.205618,2.205625) CI and 1.63 median as well. The last column of this table shows the

results from bootstrapping the weighted data. As one can see they are very similar to the complex design results.

Ratio Estimation

Ratios commonly calculated from surveys are more than likely statistically analyzed incorrectly. This is commonly seen in surveys because of the many variables collected in surveys and what they represent at that point in time. So individuals feel (or they don't know) that it is Ok just to take the ratio of two variables and then treat that number as a single entity. Usually people will often construct confidence intervals or even do tests of hypothesis on these numbers. The equations that are used in these reports (unless they state specifically that particular Ratio estimation equations were used), simply treat the number as if it were a single value sampled from some population. Then the standard errors calculated are based on asymptotic results. These lead to incorrect estimates of the variance and further produce inaccurate inferences. The bottom line is that when you take the ratio of two variables, especially when they are both considered random samples, you are taking the ratio of two random variables. And the variance of a single random variable (such as the ratio) is not the same as the variance of the ratio of two random variables. That's because there is no theoretical distribution of two random variables and the variance of the ratio is the ratio of variances. This is confusing and often difficult to explain, especially when there is a theoretical distribution for the ratio of two sample variances, which is the F-Distribution, this is true when the two populations are normally distributed. Again, this often just encourages people to take the ratio of any two variables (especially when they can come up with some interpretation) and enter it into a software package and perform inferential procedures.

But there is no known simple equation that you can use to calculate the variance of a ratio of two random variables. There are quite a few approximations to calculating the variance of the ratio using various assumptions. One such approximation considers the numerator of the ratio as a linear function involving the denominator. Incorporating the complexity of the survey design, including the sampling weights results in a non-linear equation, which requires differential equations to obtain an estimate of the variance. The actual derivation of the variance is beyond the scope of this thesis, but to show what the variance equation looks like, Shah (2004) gives the following equation.

Let x and y be two random variables obtained from a survey. They can be any two variables and then define

$$\hat{R} = \frac{\hat{Y}}{\hat{X}}$$

as the ratio of these two variables. Here $\hat{R}$ is the estimated ratio of the two statistics for each variable, and both $\hat{Y}$, $\hat{X}$ represent the statistical estimates from the two variables, such as the mean of Y and the mean of X, and you want to estimate the ratio of the two means. Then the variance of the ratio as: (Svyvariance, n.d.)

$$\hat{V}\ (\hat{R}\ ) = \frac{1}{\hat{X}^2}\{\hat{V}\ (\widehat{Y}) - \ 2\ \hat{R}\widehat{COV}(\hat{Y}\ ,\ \hat{X}) + \ \hat{R}^2\hat{V}\ (\hat{X}\ )\}.$$

Note that the variance of the estimated ratio involves both the variance of the numerator and the denominator plus some covariance between the numerator and denominator. The main point is

that the variance of the ratio of two random variables is not equal to the variance of the single value.

An example of a properly analyzed ratio is taking the systolic over diastolic blood pressure statistics from the NHANES II data. Doctors often look at blood pressure readings as a vital piece of information. The systolic-diastolic ratio tells the doctor how much pressure is being exerted on the arteries when the heart contracts and relaxes. Abnormally low blood pressure may be a symptom of dehydration, internal bleeding, certain inflammatory diseases or heart disease. High blood pressure is a potentially dangerous condition in itself, but the systolic-diastolic ratio can also warn the doctor that the patient may have a problem with his heart, kidneys or circulatory system. If the systolic-diastolic ratio is greater than 140/90, the doctor knows that the patient may have high blood pressure. So in this particular case, the ratio of two random variables does have an interpretation that is useful for clinical decision-making.

Table 6.7 below displays the ratio analysis given under simple descriptive statistics (treating the ratio as a single number) and Table 6.8 displays the proper analysis for the ratio of two random variables. The estimates obtained were done using STATA and the equations from Svyvariance.

	MEAN RATIO	**LINEARIZED STD. ERR.**	**95% Confidence Interval**
bpsystol/bpdist	1.601	.0022	(1.59, 1.61)

Table 6.7

	Mean RATIO	**LINEARIZED STD. ERR.**	**95% Confidence Interval**
bpsystol/bpdist	1.567	.0050	(1.56, 1.578)

Table 6.8

As one can see from Tables 6.7 and 6.8 there is a substantial difference in the estimates and confidence intervals of the two analyses. Remember the sample size for the NHANES II was in the thousands and the results here are extrapolated for the entire US population (over 300 million). But you can see the differences especially in the standard errors. The mean ratios are slightly different (1.6 vs. 1.57), while the standard errors are very different; again remember these are based on thousands of individuals so the estimates would be very small. But still note the difference; the adjusted standard error for the complex design is more than two times larger than the unadjusted (.005 vs. .002). Of course it should be larger, it contains variance due to the denominator and additional co-variation. Ultimately these result in quite different confidence intervals; note that these CI do not overlap.

Chapter 7 Conclusion

The purpose of this thesis was to show various procedures for estimating the variance of specific statistics obtained from different types of survey designs and then demonstrating the methods for more advanced designs such complex surveys. The thesis began with defining various sampling designs and how they are performed. Next, the actual equations necessary for estimating the variances of the statistics for each design are demonstrated, showing the difficulty especially in estimating the variance of the statistics, as the designs get more complex. After demonstrating the resampling techniques with non-complex surveys, it was found that the bootstrap and jackknife methods produced very similar results in calculating the mean and variance. However the bootstrap technique computed a tighter mean confidence interval. The proper analysis of the two complex surveys(NHANES II and NHANES 2012) required more steps to complete. The first step was to apply the appropriate weights to the raw data in order to extrapolate the U.S. population. Once the weights were applied the mean, median, and variance for each variable changed, by increasing or decreasing.Meaning simple analysis of the raw data does not give an adequate picture of the entire U.S. population. Furthermore, as seen with the non-complex data the bootstrap and jackknife methods proved to accurately estimate the means, medians, and variances of the NHANES data, but the bootstrap gave a tighter confidence interval once again. Additionally the bootstrap resampling technique was the only method used to analyze the complex data due to the limitations of R. R could not jackknife the stratified survey design of the NHANES survey. Lastly, the proper analysis of a ratio of two random variables was examined. The difference between the ratio being treated as a single entity and then as two random variables wassignificant. The mean ratios differ by more than 0.1 (1.6 vs 1.567) and the confidence intervals do not overlap.

It is evident through this thesis that ignoring weighting schemes and survey designs of complex surveys does not accurately produce estimates of the population parameters. The same disregard for weighting variables and the mistreatment of ratio variables will give misleading statistical estimates.

REFERENCES

Aune-Lundberg, L., and Geir-Harald, S. (2014). Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data.*Environmental Modelling and Software,61*: 87-97.

Centers for Disease Control and Prevention. (2015). National Health and Nutrition Examination Survey. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Frongillo, E. (1996, October). StatNews #11: What is a Complex Survey? *Cornell Statistical Consulting Unit*.Retrieved from https://www.cscu.cornell.edu/news/statnews/stnews11.pdf

Kernler, D. (2014a). A visual representation of selecting a random sample using the cluster sampling technique. Retrieved from https://en.wikipedia.org/wiki/Sampling_%28statistics%29#/media/File:Cluster_sampling.PNG Made available under a CC BY SA 4.0 License: https://creativecommons.org/licenses/by-sa/4.0/deed.en

Kernler, D. (2014b). A visual representation of selecting a simple random sample. Retrieved from https://en.wikipedia.org/wiki/Sampling_%28statistics%29#/media/File:Simple_random_sampling.PNG. Made available under a CC BY SA 4.0 License: https://creativecommons.org/licenses/by-sa/4.0/deed.en

Kernler, D. (2014c). A visual representation of selecting a random sample using the stratified sampling technique. Retrieved from https://en.wikipedia.org/wiki/Sampling_%28statistics%29#/media/File:Stratified_samplin

g.PNG Made available under a CC BY SA 4.0 License: https://creativecommons.org/licenses/by-sa/4.0/deed.en

Kernler, D. (2014d). A visual representation of selecting a random sample using the systematic sampling technique. Retrieved from https://en.wikipedia.org/wiki/Sampling_%28statistics%29#/media/File:Systematic_sampling.PNG Made available under a CC BY SA 4.0 License: https://creativecommons.org/licenses/by-sa/4.0/deed.en

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Lohr, S. (1999). *Sampling: Design and analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, NJ: Wiley.

R Development Core Team. (2015). The R Project for Statistical Computing. Retrieved from http://www.r-project.org/

StataCorp. (2015).Variance Estimation. In *Stata Survey Data Reference Manual* (Release 14) (pp. 186-199). College Station, TX: Stata. Retrieved from http://www.stata.com/manuals13/svyvarianceestimation.pdf

Shao, J., and Tu, D. (1995). *The jackknife and bootstrap*. New York, NY: Springer.

Variance Issue in Systematic Sampling. (n.d.). AWF-Wik. Retrieved from http://wiki.awf.forst.uni-goettingen.de/wiki/index.php/Variance_issue_in_systematic_sampling

Were, F. (2015). A Design Unbiased Variance Estimator of the Systematic Sample Means. *American Journal of Theoretical and Applied Statistics, 4*(3): 201-210.