An Introduction to Probability Sampling

Josué Guzmán, PhD
<JGuzmanPhD@Gmail.Com>

May 27, 2009

Contents

Preface **v**

1 Introduction to Sampling **1**

1.1 Basic Concepts 1

1.1.1 Some Advantages Probability Sampling 2

1.1.2 Some Limitations of Probability Sampling 3

1.1.3 Questions to Consider 3

1.2 Common Descriptive Measures 4

1.3 Random Sampling 5

1.4 Sampling Errors 5

1.4.1 Accuracy, Precision, and Bias 6

1.5 Non-Sampling Errors 7

1.6 Some Probability Sampling Schemes 8

2 Simple Random Sampling **11**

2.1 Introduction 11

2.1.1 Using Program *R* 13

2.2 Basic Formulae in *SRS* 13

2.3 Sampling Distribution 14

2.3.1 Covariance 15

2.4 Variance of Linear Functions 15

2.5 Variances in *SRS* 16

2.5.1 *SRSWR* 16

2.5.2 *SRSWOR* 16

2.6 Variances Estimation and Standard Errors 17

2.7 Estimation of a Proportion 18

2.8 Sample Size Under *SRS* 19

3 Systematic and Replicated Sampling **23**

3.1 Introduction to Systematic Sampling 23
3.2 Some Advantages of Systematic Sampling 25
3.3 Sampling Probability 25
3.4 Systematic Sampling Using Program *R* 27
3.4.1 Linear Systematic Sampling 27
3.4.2 Circular Systematic Sampling 28
3.5 Introduction to Replicated Sampling 28
3.6 Estimation of μ_x in *R-SRS* 29
3.6.1 Estimation of p_y in *R-SRS* 29
3.7 Replicated-SRS Using Program *R* 30
3.8 Estimation of μ_x in *R-SysRS* 30
3.8.1 Estimation of p_y in *R-SysRS* 30
3.9 Replicated Linear-Systematic Sample Using *R* 31
3.9.1 Replicated Circular-Systematic Sample Using *R* 32

4 Basics of Stratified Sampling **35**

4.1 Introduction 35
4.2 Some Advantages of Stratified Sampling 36
4.2.1 Some Disadvantages of Stratified Sampling 37
4.3 Proportional Stratified Sampling 37
4.4 Other Stratified Sampling Allocations 38
4.4.1 Proportional Allocation Using *R* (Two Strata) 38
4.4.2 Equal Allocation Using *R* (Two Strata) 39
4.4.3 Optimal Allocation Using *R* (Two Strata) 39
4.5 Estimation of the mean, μ_x, in the frame 40
4.5.1 Estimation of the proportion, p_y, in the frame 41
4.6 Variance Comparison 42
4.6.1 Proportional Allocation vs *SRS* 42
4.6.2 Proportional Allocation vs Optimal Allocation 43
4.7 Replicated Stratified Sampling 44
4.8 A Limitation of Stratified Sampling 45

5 Introduction to Clustered Sampling 49
5.1 Introduction . . . 49
5.2 Some Advantages of Clustered Sampling . . . 50
5.2.1 Some Disadvantages of Clustered Sampling . . . 50
5.3 Two-Stage Clustered Sampling . . . 51
5.3.1 Two-Stage Cluster Sample Using *R* . . . 51
5.3.2 Examples of Two-Stage Clustered Sampling . . . 52
5.4 Two-Stage Cluster Sampling Estimation . . . 53
5.4.1 Frame Total Estimation . . . 53
5.5 Clustered Sampling Allocation . . . 54
5.5.1 Clustered Sampling Allocation With *R* . . . 55
5.6 PPS Clustered Sampling . . . 55
5.6.1 Sample Selection Procedure . . . 56
5.6.2 A *PPS* Sample Selection Procedure Example . . . 57
5.6.3 A *PPS* Sample Selection Using *R* . . . 57

6 Introduction to Ratio and Regression Estimation 61
6.1 Ratio Estimation in *SRS* . . . 62
6.1.1 Ratio Estimation Example . . . 62
6.2 Ratio Estimation in Stratified Sampling . . . 65
6.2.1 Combined Ratio Estimation . . . 65
6.2.2 Separate Ratio Estimation . . . 66
6.3 Introduction to Regression Estimation . . . 66
6.3.1 Regression Trough The Origin, *RTO* . . . 67
6.3.2 Simple Regression . . . 68
6.3.3 Regression Estimation Using *R* . . . 69

7 The *R* Survey Package 73
7.1 Introduction . . . 73
7.2 Basic Estimation . . . 74
7.3 Describing Surveys To *R* . . . 75
7.4 Clustered Sampling . . . 76
7.5 Two-Stage Sampling . . . 77
7.6 *PPS* Sampling . . . 78
7.6.1 *PPS* Sampling Example . . . 78

7.7 Summary Statistics 79
7.8 Tables 80
7.8.1 Computing Over Subgroups 80
7.9 Cross-Tabulations 82

A Quick Introduction to *R* 83
A.1 Getting and Working with *R* 83
A.1.1 Some Restrictions 84
A.2 Graphics 84
A.3 Getting Help 84
A.4 *R* Packages 85
A.5 Data Types 85

B Accessing Data using *R* 87
B.1 Reading Text File 87
B.1.1 Using `read.table( )` 88
B.2 Text Files Export 88
B.3 Data From Other Statistics Programs 88
B.3.1 Other Statistics Programs 88
B.4 Reading Excel Spreadsheets 89

C Review of Basic Probability 91
C.1 Types of Probability 91
C.2 Probability Rules 92
C.3 Conditional Probability 93
C.4 Probability Table 93
C.4.1 Probability Table Example 93
C.5 Bayes' Rule 93
C.5.1 Bayes' Rule Example 94

Preface

The good thing about Statistics is that you meet the best people the world-over.[a]

[a] W Edwards Deming; May 1975. Personal communication.

Figure 1: W Edwards Deming

Remark

In this as well as the following chapters, we will begin and end with a quote from Dr. W Edwards Deming's writings.[1] It is hoped that these quotes will serve as reflexions or messages that will help the reader, both at work and live in general.

What is this all about?

The present manuscript was used as a basic instructional device for an introduction on *Probability Sampling*. Other complementary material was offered to

[1] It was our good fortune and privilege to have being a disciple of Dr. Deming during our graduate studies at Stern School of Business. Certainly, it is one of our greatest honors of having the opportunity to study under a great human being. And that is why these notes are dedicated to his memory.

the participants with the intended purpose that it would elaborate or clarify some of the ideas presented in the manuscript. The R[2] program is used as a supporting computing device for illustration of different sampling techniques and procedures.

As for this manuscript, we present the following (non-exhaustive) topics:

- Chapter One introduces the basic concepts on Probability Sampling;
- Chapter Two presents *Simple Random Sampling*, which forms the basic root for other sampling designs;
- Chapter Three covers both *Systematic* and *Replicated Sampling* methodologies;
- Chapter Four introduces *Stratified Sampling*design;
- Chapter Five covers some *Clustered Sampling* techniques;
- Chapter Six includes both *Ratio* and *Regression Estimation* with auxiliary information;
- Chapter Seven introduces the `survey` package, to be used with R program.

Several exercises are included in order to reinforce the concepts introduced to the reader.

We should notice that *Probability Sampling* has been one of the topics which we previously taught in courses, both at the Institute of Statistics, Faculty of Business Administration, Río Piedras Campus, and the Department of Biostatistics and Epidemiology, School of Public Health, Medical Sciences Campus, of the University of Puerto Rico.

Thanks

The author wants to recognize the support for this project, which include: Dr. Mario Marazzi, Executive Director of the Puerto Rico Institute of Statistics, and its staff, especially, Mr. Orville Disdier, Idania Rodríguez, Luz Mairym López, Héctor López, Francisco Acevedo, and Yashira Guzmán. Thanks to all of them for their dedication and collaboration with the first *Academy on Sampling* for statistics-engaged government employees.

We also want to recognize the dedication and interest on Probability Sampling that showed the following participants of the Academy: Gloria Rosado (Carolina Municipality), Carol Málaga (Turism Department), Aixa Díaz (Puerto

[2] An open-source statistics and graphics program, available free of charge at www.r-project.org

Rico Special Communities Office), Alex López (Agriculture Department), Idania Rodríguez (Puerto Rico Institute of Statistics), Myriam Ramos (Department of Health), and Rafael Silvestrini (Turism Department). Thanks to all for your patient and helpful suggestions on this manuscript. As always, we continue to learn from our students.

Our colleagues, Professors Wilfredo Camacho and Jairo Fúquene, of the Institute of Statistics, Faculty of Business Administration, of the University of Puerto Rico, read part of the manuscript and also made valuable suggestions. We appreciate their interest and help. Needless to say, but any errors, blemishes, and blunders are the complete and exclusive responsibility of the author.

Finally, but importantly, many thanks to my wife Maritza, for her patient, support, and encouragement throughout this and other professional engagements.

Chapter 1

Introduction to Sampling

Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability.[a]

[a] W Edwards Deming; 1950, 1986. *Some Theory of Sampling.* New York: Wiley, Dover. Page 2.

1.1 Basic Concepts

This is an introduction to *probability sampling*. With probability samples, we can calculate sampling errors; we can also eliminate selection biases (via random sampling) and non-response and estimation errors can be contained within known limits. We will present some of the various survey sampling designs, both in probability and non-probability sampling, including some of their advantages and disadvantages. Within probability sampling, we will cover the basic concepts of: simple random sampling (the basis of other designs), systematic, replicated, stratified, and clustered sampling. And within non-probability sampling, we will include subjective selection, quota, convenience, and snowball sampling. We emphasize that our main focus will be on probability sampling techniques.

We start describing some basic concepts in sampling:

Population — Formed by all elements or units of interest in a study, generally, at a time point. A population is defined by the context of the study. It could be made of people, but it could also be businesses, farms, schools, factory products, animals, and so on.

Sampling frame — List of sampling units that gives access to the population of interest. A frame could also be formed by maps, with explicit unit boundaries, in an area study, as we will see later. The *frame size* is made of N sampling

units; $\{U_i = U_1, U_2, \ldots, U_N\}$. In these notes, we will assume that N is finite or countable.

Sample — Part or fraction of a population selected from the frame. The *sample size* is made of $n \leq N$ sampling units; $\{u_i = u_1, u_2, \ldots, u_n\}$.

Census — It refers to a survey where all elements or units of interest are studied. Notice that a census is a special sample, where $n = N$.

Parameter — Suppose that we have a measure of interest in our study, Y, and this measure have some characteristics which we would like to estimate. If we examine the frame: $\{(U_i, Y_i) = (U_1, Y_1), (U_2, Y_2), \ldots, (U_N, Y_N)\}$, a function of Y_i in the frame is known as a *parameter*. For example: mean, median, percentiles, variance, standard deviation, and others.

1.1.1 Some Advantages Probability Sampling

Probability sampling provides measures of sampling errors with respect to the estimates that come from the data and we can generalize the findings of a study, based on an inference from the sample to the frame. Some of the advantages of probability sampling are:

- It improves an statistical program by clarifying the objectives of the study;
- It provides a quantitative measure of the extent of variation due to random effects (*sampling error*);
- It provides data of known quality, through statistical control procedures;
- It reduces the risk of non-response and the burden of response;
- It allows better interviewing or testing, better supervision, better data processing than a complete enumeration survey (census);
- It provides data in timely fashion, as compared to a complete enumeration survey;
- It provides acceptable data reliability at a reduced cost;
- Due to the smallness of the study, we obtain more control over non-sampling errors, as compared to a complete enumeration survey;
- Statistical inference and probability theory can be applied to analyze and interpret the data.

1.1.2 Some Limitations of Probability Sampling

Probability sampling involves several risks and limitations. In fact, because we are not studying the complete frame, we will always have a risk of sampling errors. Some of the limitations of probability sampling include:

- It depends on the "goodness" of the sampling frame (although this is also a limitation in complete enumeration studies);
- Because we are examining a portion of the frame, sampling does no give specific information on every sampling unit (people, account, inventory, patients, ...);
- Sampling errors are inherent in a probability sample, and can be large for some aims of a study;
- In small areas or rare sub-populattions, sampling errors may be high;
- Representativeness of the frame (or population) may be questionable or controversial.

1.1.3 Questions to Consider

In any study, either from a sampling design or from a complete enumeration survey, there will be several questions that need to be answered. Some of these questions are:

- What is your main research question? (study purpose);
- What is your population of interest? (target population);
- What do you know about this population? (previous study);
- Do you have a sampling frame? (access to the population);
- How good is the sampling frame? (appropriateness);
- Do you have an existing questionnaire? (data gathering instrument);
- When do you need your data and analysis? (time frame);
- How much money do you have? (cost of the study).

1.2 Common Descriptive Measures

Definition 1: Mean.
The parameter known as *mean* is the sum of all the Y values in a frame, divided by N. The mean of a whole frame is usually denoted by μ_y, while the mean of a sample is usually denoted by $\bar{y}$. (Note that this is the arithmetic mean; there are other means, which will be discussed later.)

Thus, the mean of measur Y in the frame is given by

$$\mu_y \equiv \frac{Y_1 + Y_2 + \cdots + Y_N}{N} \tag{1.1}$$

Definition 2: Median.
The median is the middle number of a set of Y values, arranged in numerical order. If the number of values in a set is even, then the median is the midpoint of the two central values.

The median is not affected by the magnitude of the extreme (smallest or largest) Y values. Thus, it is useful because it is not affected by one or two abnormally small or large values, and because it is very simple to calculate.

Unlike the median, the mean is sensitive to *any* change Y values, while a change to an extreme value (in the case of a median) usually has no effect.

Definition 3: Variance.
This is a measure of how items are dispersed about their mean. The variance, σ_y^2, of a frame is given by the equation

$$\sigma_y^2 \equiv \frac{\sum_{i=1}^{N}(y_i - \mu_y)^2}{N} \tag{1.2}$$

The variance, s_y^2, of a sample is usually calculated as:

$$s_y^2 \equiv \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{1.3}$$

Definition 4: Standard Deviation.
The standard deviation, σ_y (or s_y for a sample), is the positive square root of the variance.

Definition 5: Relative Variability.
The relative variability, also known as the *coefficient of variation*, of a frame is its standard deviation divided by its mean.

$$\mathrm{CV}_y \equiv \frac{\sigma_y}{\mu_y} \tag{1.4}$$

The relative variability is useful for comparing several groups.

1.3 Random Sampling

Assume that the sampling frame has $N = 1,000$ units, $\{U_i = 001, 002, \ldots, 1000\}$, and we want to select a sample of $n = 10$ units, $\{u_i = u_1, u_2, \ldots, u_{10}\}$, by means of a random procedure. The sample will be selected[1] using the following scheme:

1. The sampling selection is *without replacement*.
2. Each unit in the sampled frame has the same probability, $\frac{10}{1000}$, of entering the sample.
3. The sample is chosen by a process of automatic randomization.

Just imagine an urn with $1,000$ numbered chips, $\{001, 002, \ldots, 1000\}$, and suppose that 10 randomly selected chips, *without replacement*, are numbered:

$$\{204, 436, 714, 917, 873, 770, 398, 187, 778, 732\}.$$

Then, these constitute our selected sampling units for subsequent investigation and statistical analysis.

If the sampling procedure were *with replacement*, we would obtain a sample where any of the $N = 1,000$ units could appear more than once.

Random sampling selection will be contrasted with *non-random sampling*, in which generally, selection is made by judgement, where the intent is to get a "representative" sample based on subject-matter expertise. Latter on, we will mention several non-random sampling procedures.

1.4 Sampling Errors

Because of the fact that we are examining $n < N$ sampling units, there will be un-certentainty with respect to the measures based on said sample values. Furthermore, if random selection is repeated using the same sampling procedure, we will obtain samples that differ with respect to the selected units.

Suppose that θ is a general parameter of interest, and it is estimated using an *estimator* $\hat{\theta}$. Generally, in repeated sampling, from the same frame and using the same sampling procedure, $\hat{\theta}$ differs from sample to sample. Assume that from the i^{th} selected sample, you get an estimate $\hat{\theta}_i$ of θ. The difference, $\delta_i \equiv \hat{\theta}_i - \theta$ is known as an *estimation error*; notice that δ_i could be either positive, negative or zero.

[1] Cf. WG Cochran, F Mosteller & JW Tukey; 1954. *Statistical Problems of the Kinsey Report*. Washington, DC: American Statistical Association.

Definition 6: Expected Value.
Suppose that $\hat{\theta}$ could have D different values. The mean or expected value of $\hat{\theta}$ is,

$$\mathrm{E}(\hat{\theta}_i) \equiv \sum_{1}^{D} \hat{\theta}_i \Pr(\hat{\theta}_i), \tag{1.5}$$

where $\Pr(\hat{\theta}_i)$ is the probability of selection for the i^{th} estimator $\hat{\theta}_i$.

If this value equals the parameter θ in the frame, we say that $\hat{\theta}$ is an *unbiased estimator* of θ. The difference $\mathrm{E}(\hat{\theta}) - \theta \equiv \mathrm{B}(\hat{\theta})$ is known as the *bias* of the estimator.

Definition 7: Mean Square Error.
In repeated sampling, the average value of the squared estimation error is known as the *mean square error*, mse. Thus,

$$\mathrm{MSE}(\hat{\theta}) \equiv \sum \delta_i^2 \Pr(\theta_i) = \sum (\hat{\theta}_i - \theta)^2 \Pr(\hat{\theta}_i). \tag{1.6}$$

The mean square error measures the *accuracy* of the estimator $\hat{\theta}$ with respect to the parameter θ. And the positive square root of $\mathrm{MSE}(\hat{\theta})$ is known as the *root mean square error*.

Definition 8: Sampling Variance.
In repeated sampling, the average value of squared deviation of the estimator, $\hat{\theta}$, with respect to its expected value is known as the *sampling variance*.

$$\mathrm{Var}(\hat{\theta}) = \sigma_{\hat{\theta}}^2 \equiv \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 = \sum (\hat{\theta}_i - \mathrm{E}(\hat{\theta}))^2 \Pr(\hat{\theta}_i) \tag{1.7}$$

This variability measure is also known as the *precision* of the estimator $\hat{\theta}$. The positive square root of $\mathrm{Var}(\hat{\theta})$ is known as the *standard error* of the estimator, $se(\hat{\theta})$. And the ratio of $se(\hat{\theta})$ to its expected value, is known as the coefficient of variation of the estimator $\hat{\theta}$:

$$\mathrm{CV}(\hat{\theta}) \equiv \frac{se(\hat{\theta})}{\mathrm{E}(\hat{\theta})} \tag{1.8}$$

1.4.1 Accuracy, Precision, and Bias

It can be shown that

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 + (\mathrm{E}(\hat{\theta}) - \theta))^2,$$

where $(\mathrm{E}(\hat{\theta}) - \theta)^2 \equiv \mathrm{B}^2(\hat{\theta})$ is the square *bias* of the estimator.
Therefore,

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{B}^2(\hat{\theta}), \tag{1.9}$$

i.e., *accuracy* = *precision* + *square bias* of the estimator, $\hat{\theta}$.

1.5 Non-Sampling Errors

It should be emphasized that sampling errors are intrinsic to the sampling procedure, due to the fact that we consider a portion $n < N$ of the sampling frame. Thus, the sampling error diminishes as the sample size increases. However, *non-sampling errors* occur both in sampling and in a complete enumeration survey (census); and these errors could increase as the size of the study increases.

There are many causes for the occurrence of non-sampling errors,[2] including:

- Lack of a careful statement of the problem to be investigated and the type of statistical information needed;
- No operational or clear definition of the population to be investigated;
- A defective or inappropriate sampling frame to access the population of interest;
- Deficiencies in operational definitions for measures of interest in the study;
- Deficiencies in the questionnaire or the instrument for data gathering;
- Non-response errors (e.g., respondent not found or refuses to answer);
- Response errors (e.g., respondent knowingly gives the wrong answer because does not understand the question);
- Errors in coding and faulty editing;
- Tabulation errors due to faulty selection of characteristics, class intervals, too many or too few cross-tabulations;
- Un-trained personnel for both supervision and field work;
- Bad timing for the study;
- Errors induced by the purpose of the study;
- Errors induce by the sponsors of the study;
- Errors, both voluntary or involuntary, induced by the interviewer.

[2] For a more elaborate list of non-sampling errors, with an extensive discussion, see Chapter Two of *Some Theory of Sampling*, by W Edwards Deming.

1.6 Some Probability Sampling Schemes

In this section, we briefly describe some of the basic probability sampling designs and selection methods. We will elaborate these techniques in subsequent chapters.

***Simple Random Sampling,* SRS** — Each sampling unit has same chance, n/N, of being selected from the frame. *SRS* is an elementary probability sampling design, but it is fundamental for other sampling techniques. It will be the topic of next Chapter.

Systematic Sampling — Begin with a random start, r, chosen between 1 and $k = N/n$. Then, select every k^{th} element until the sample is completed:

$$r,\, r+k,\, r+2k, r\, +3k,\, \ldots,\, r+(n-1)k.$$

There are different variants of systematic selection. For example, suppose that $N = 1,000$ and $n = 100$; then, $k = 1000/100 = 10$. If you select random start, $r = 7$, then, select the following units from the frame:

$$7, 17, 27, 37, 47, 57, \ldots, 997.$$

We will discuss systematic sampling in Chapter Three.

Replicated Sampling — It consists on splitting the sample into g, for $g \geq 2$, independent subsamples, each of size m, so that each of these subsamples can provide valid estimates of the parameter of interest. Notice that $mg = n$, the overall sample size.

This technique is also known as *interpenetrating sub-samples*, a phrase coined by PC Mahalanobis. Replicated sampling will also be covered in Chapter Three.

Stratified Sampling — Before selection, the frame is divided into few, H say, homogeneous groups, known as *strata*. The objective is to obtain as much as possible homogeneity within a stratum and to mantain heterogeneity between strata.

For example, suppose that we have three strata (e.g., *Small*, *Medium*, and *Large*) of sizes N_1, N_2, N_3; where $\sum N_i = N$. Make random selection of sampling units from each stratum: n_i of N_i, for $i = 1, 2, 3$; where $\sum_{i=1}^{3} n_i = n$. We will elaborate on stratified sampling in Chapter Four.

Clustered Sampling — Before selection, the frame is divided into a large number, M, of non-homogeneous groups, known as *clusters*. Sampling is performed in at least two stages:

- Stage I — A random selection of $m < M$ clusters;
- Stage II — A random selection of n_i sampling units from each of the m selected clusters, $\sum_{i=1}^{m} n_i = n$.

The objective is to get heterogeneity within clusters and to maintain homogeneity between clusters. Clustered sampling will be the topic of Chapter Five.

Exercises

The practical man is the man who practices the errors of his forefathers— TH Huxley.[a]

[a]Cited in, W Edwards Deming; 1960, 1990. *Sampling Design in Business Research.* New York: Wiley Classics Library. Page 276.

1. Your are assigned a task of developing a sampling frame for a study of prevalence of a male disease in San Juan Metropolitan Area (SJMA). The study will implement a probability sample survey of men, age 21 years or more. Explain how you would make such a frame. SJMA is made of the municipalities of: Bayamón, Carolina, Cataño, Guaynabo, Trujillo Alto, and San Juan.

2. For the study mentioned in the previous exercise, discuss some possible limitations of the sampling frame.

3. Prepare a sampling frame of your workplace with the following employees' information: id number $(1, 2 \ldots N)$; sex, years of experience, and whether (s)he travels to work by: car, bus, or train.

4. What would you consider as a sampling frame for a study of un-employement in Puerto Rico. Explain.

5. What would you consider as a sampling frame for a study of food sales in Puerto Rico. Explain.

Chapter 2

Simple Random Sampling

The theory of knowledge teaches us that a statement, if it to convey knowledge, predicts future outcome, with risk of being wrong, and that it fits without failure observations of the past ... It is extension of application that discloses inadequacy of a theory, and need for revision, or even a new theory. Again, without theory, there is nothing to revise. Without theory, experience has no meaning. Without theory, one has no question to ask. Hence without theory, there is no learning.[a]

[a] W Edwards Deming; 1993. *The New Economics: For Industry, Government, Education.* Cambridge, MA: MIT, CAES. Pages 105 *et seq.*

2.1 Introduction

The basic design scheme in probability sampling is known as *Simple Random Sampling, srs*. Other design techniques are fundamentally based on *srs*, and their purpose is to increase efficiency, relative to *srs*, and to manage auxiliary information, and thus to reduce sampling errors.

Definition 9: SRS.
Suppose that we randomly select from a sampling frame, of size N, so that each sample of n *different* units has an equal probability of being the selected sample. There exist

$$\binom{N}{n} = \frac{N!}{n! \cdot (N-n)!}$$

possible samples under *srs without replacement, srswor,* and any of these samples could be the selected one, with selection probability

$$\Pr(S) = \frac{1}{\binom{N}{n}}.$$

Operationally, the selection under *srswor* is done using a *random digits table.*[1] We select, with equal probability, n different numbers which correspond to n of the N sampling units labeled: $\{U_i = 1, 2, 3, \ldots, N\}$ in the frame. Every sampling unit has equal probability of being selected in each of the n selections, but the previously selected units are disqualified if its id number is repeated while reading the random digits table.

If sampling is *with replacement, srswr*, the selected sampling units can be chosen again in subsequent selections. Under *srswr*, there are N^n possible samples of size n; therefore, each sample of n sampling units (not necessarily different) will have a probability of N^{-n} of being the selected sample.

Under *srswr*, the probability of selection for any sampling unit in one of the n selections, is $1/N$. Thus, in the n independent selections, the expected value that each sampling unit u, will be selected is,

$$\mathrm{E}(u) = \sum_{1}^{n} u \Pr(u) = \overbrace{1/N + 1/N + \cdots + 1/N}^{n \text{ times}} = \frac{n}{N}. \tag{2.1}$$

Under *srswor*, the probability of selection for any sampling unit the first time is $1/N$. For the second selection, the probability of being selected, is $1/(N-1)$, conditioned on the probability $(N-1)/N$ that is as not selected on the first occasion. Therefore, the probability of it selection on the second occasion is

$$\frac{1}{N-1} \times \frac{N-1}{N} = \frac{1}{N}.$$

Similarly, the probability of selection on the third occasion is $1/(N-2)$, conditioned on the probability that it was not selected on the two previous occasions, which is

$$\frac{N-1}{N} \times \frac{N-2}{N-1} = \frac{N-2}{N}.$$

Then, the joint probability of being selected on the third selection is $1/N$. Therefore, the total probability will be the sum of the n probabilities, each of which is $1/N$, will be

$$\overbrace{\frac{1}{N} + \frac{1}{N-1} \cdot \frac{N-1}{N} + \frac{1}{N-2} \cdot \frac{N-1}{N} \cdot \frac{N-2}{N-1} + \ldots}^{n \text{ times}} = \frac{n}{N}.$$

Alternatively, we notice that under *srswor*, there exist $\binom{N}{n}$ equally possible samples and among these, there exist $\binom{N-1}{n-1}$ which contain any specific unit. Therefore, the probability of it selection is

$$\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}, \tag{2.2}$$

and this is the same, previously obtained, result.

[1] See e.g., MG Kendall & BB Smith; 1961. *Random Sampling Numbers.* Cambridge University Press.

2.1.1 Using Program *R*

With the advent of fast, low-cost computing, random selection is implemented easily. For example, using open-source program R, we could make a *pseudo-random* selection of n out of N sampling units in a frame. R provides an easy-to-use function explicitly called `sample( )`. Thus, if we want to select a *srswor* of 10 out of the $1,000$ units in a frame using R, we could proceed as follows:

```
U = 1:1000
sample(U, 10).
```

And, if we want to select a *srswr* of 10 out of the $1,000$ units in a frame, we could proceed as follows:

```
U = 1:1000
sample(U, 10, replace=TRUE).
```

2.2 Basic Formulae in *SRS*

Let suppose that the frame has N sampling units, labeled $\{U_i = 1, 2, 3, \ldots, N\}$. Furthermore, suppose that it has quantitative measure X, and a binary variable $Y = (0, 1)$.

Definition 10: Expected Values.
If x_i is the i^{th} value of the randomly selected variable X; equivalently, if $y_i = (0, 1)$ is the i^{th} value of the randomly selected dichotomous variable Y, then, in repeated sampling, their respective *expected values* are:

$$\mathrm{E}(x_i) = \sum_1^N x_i \frac{1}{N} = \mu_x, \tag{2.3}$$

and

$$\mathrm{E}(y_i) = \sum_1^N y_i \frac{1}{N} = p_y. \tag{2.4}$$

Definition 11: Variances.
Again, in repeated sampling, the corresponding *variances* of x_i and y_i are, respectively:

$$\sigma_x^2 = \mathrm{E}(x_i - \mu_x)^2 = \sum_1^N (x_i - \mu_x)^2 \frac{1}{N}, \tag{2.5}$$

and

$$\sigma_y^2 = \mathrm{E}(y_i - p_y)^2 = \sum_1^N (y_i - p_y)^2 \frac{1}{N} = p_y(1 - p_y). \tag{2.6}$$

The positive square root of the variance is known as the *standard deviation*.

Definition 12: Relative Variance.
The *relative variance*, RV, is the ratio of the variance and the square of it corresponding mean value:

$$\mathrm{RV}_x \equiv \frac{\sigma_x^2}{\mu_x^2} \quad \text{and} \quad \mathrm{RV}_y = \frac{p_y(1-p_y)}{p_y^2} = \frac{1-p_y}{p_y}, \tag{2.7}$$

And the *coefficient of variation* is the corresponding positive square root of the RV:

$$\mathrm{CV}_x \equiv \frac{\sigma_x}{\mu_x} \quad \text{and} \quad \mathrm{CV}_y = \sqrt{\frac{1-p_y}{p_y}}. \tag{2.8}$$

2.3 Sampling Distribution

The *sampling distribution* of the statistic $\hat{\theta}$, is formed by its different values, $\{\hat{\theta}_i; \quad \text{for} \quad i = 1, 2, \ldots, D\}$, and it corresponding probabilities, $\Pr(\hat{\theta}_i)$. The expected, mean, value of the sampling distribution is, by definition,

$$\mathrm{E}(\hat{\theta}_i) \equiv \sum_1^D \hat{\theta}_i \Pr(\hat{\theta}_i). \tag{2.9}$$

If this value equals the parameter θ in the frame, we say that $\hat{\theta}$ is an *unbiased estimator* of θ. Otherwise, the difference $\mathrm{E}(\hat{\theta}) - \theta \equiv \mathrm{B}(\hat{\theta})$ is its bias.

The variance of the sampling distribution of the statistic $\hat{\theta}$ is the expected value of the square mean deviation:

$$\mathrm{Var}(\hat{\theta}) = \sum [\hat{\theta}_i - \mathrm{E}(\hat{\theta}_i)]^2 \Pr(\hat{\theta}_i) \tag{2.10}$$

The standard deviation of the sampling distribution of the statistic $\hat{\theta}$ is known as the *standard error*.

The *mean square error*, MSE, of the statistic $\hat{\theta}$ is, by definition,

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &\equiv \mathrm{E}(\hat{\theta} - \theta)^2 \\ &= \mathrm{E}\left(\hat{\theta} - \mathrm{E}(\hat{\theta})\right)^2 + \left(\mathrm{E}(\hat{\theta}) - \theta\right)^2 \\ &= \mathrm{Var}(\hat{\theta}) + \mathrm{B}^2(\hat{\theta}). \end{aligned} \tag{2.11}$$

2.3.1 Covariance

Definition 13: Covariance.
The covariance between measures X and Z is given by,

$$\mathrm{Cov}(XZ) \equiv \mathrm{E}[X - \mathrm{E}(X)][Z - \mathrm{E}(Z)] = \sigma_{xz}. \tag{2.12}$$

Definition 14: Correlation Coefficient.
If two variables are independent, their covariance is zero (0), and we say that they are not correlated. This is due to the fact that Karl Pearson's *linear correlation coefficient* is defined as:

$$\mathrm{Cor}(XZ) = \rho_{xz} \equiv \frac{\mathrm{Cov}(XZ)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Z)}} = \frac{\sigma_{xz}}{\sigma_x \cdot \sigma_z} \tag{2.13}$$

Notice that ρ_{xz} is a pure number, whose range of values is $-1 \leq \rho_{xz} \leq +1$.

2.4 Variance of Linear Functions

Suppose that we want to obtain the variance of a measure Y, which is a linear function of J variables $X_1, X_2, \ldots, X_J$, weighted by constant factors, $W_1, W_2, \ldots, W_J$. That is, if

$$Y = \sum_{j=1}^{J} W_j \, X_j. \tag{2.14}$$

Then, the variance of Y is given by

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}\left(\sum_{1}^{J} W_j \, X_j\right) \\ &= \sum W_j^2 \, \mathrm{Var}(X) + 2 \sum_{j<k} W_j W_k \, \mathrm{Cov}(X_j, X_k). \end{aligned} \tag{2.15}$$

An example is the sum or difference of two random variables X_1 and X_2, where $W_1 = 1$ y $W_2 = \pm 1$:

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \mathrm{Var}X_2) \pm 2 \, \mathrm{Cov}(X_1, X_2). \tag{2.16}$$

And if the variables are not correlated, the covariance disappears from the two previous equations.

2.5 Variances in *SRS*

2.5.1 *SRSWR*

If we randomly select n sampling units *with replacement, srswr*, the sample total $t_x = \sum x_i$ has variance

$$\mathrm{Var}(t_x) = \mathrm{Var}\left(\sum_1^n x_i\right) = \sum_1^n \mathrm{Var}(x_i) = n\sigma_x^2. \tag{2.17}$$

Since n is a fixed value, the sample mean, $\bar{x} = \sum x_i/n$, is obviously, a linear function of the obtained values, with constant factor $1/n$. Then, the variance of the sample mean is

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} n\sigma_x^2 = \frac{\sigma^2}{n}. \tag{2.18}$$

To estimate the total $\tau_x = N\mu$, in the sampling frame, we use $N\bar{x}$, as estimator. Its variance is estimated as

$$\mathrm{Var}(N\bar{x}) = N^2 \mathrm{Var}(\bar{x}). \tag{2.19}$$

A pertinent exercise is to demonstrate that the coefficient of variation of the mean, $\bar{x}$ and the total, t_x, are equal.

2.5.2 *SRSWOR*

Because the sample mean is an unbiased estimator of the mean in the frame, we note that the total t_x, is an unbiased estimator of $n\mu_x$; i.e., $\mathrm{E}(t_x) = n\mu_x$. Then, the variance of t_x is obtained as folows:

$$\begin{aligned}
\mathrm{Var}(x) &= \mathrm{E}\left[\left(\sum_1^n x_i - n\mu_x\right)^2\right] \\
&= \mathrm{E}\left[\sum_1^n (x_i - \mu_x)^2\right] \\
&= \mathrm{E}\left[\sum_1^n (x_i - \mu_x)(x_i - \mu_x)\right] \\
&= \mathrm{E}\left[\sum_1^n (x_i - \mu_x)^2 + \sum_{i\neq j} (x_i - \mu_x)(x_j - \mu_x)\right].
\end{aligned} \tag{2.20}$$

By squaring the n terms, we get an $n \times n$ matrix, where we separate the n variances in the main diagonal and the remaining $n(n-1)$ covariances. Under

srswor, for each of the $n(n-1)$ pair of covariances, the expected value is the mean value between the $N(N-1)$ pairs of covariances in the frame. Similarly, for each of the n terms of variance, the expected value is the variance of the sampling units in the frame. Therefore,

$$\begin{aligned}\mathrm{Var}(t_x) &= \frac{n}{N}\sum_1^N (x_i-\mu_x)^2 + \frac{n(n-1)}{N(N-1)}\left[\sum_1^N (x_i-\mu_x)(x_j-\mu_x)\right] \\ &= n\sigma_x^2 + \frac{n(n-1)}{N(N-1)}\left[\left(\sum_1^N (x_i-\mu_x)\right)^2 - \sum_1^N (x_i-\mu_x)^2\right] \quad (2.21)\end{aligned}$$

The N square terms in the frame result in an $N \times N$ matrix, $\sum_1^N (x_i-\mu_x)^2$, and subtracting the N terms of the variance from the main diagonal, it leaves the remaining covariance terms. We note that the first term in square brackets, `[ ]`, in the above equation disappears due to the fact that $\sum_1^N (x_i-\mu_x) = 0$; also, we note that $\sum_1^N (x_i-\mu_x)^2 = N\sigma_x^2$.

Therefore,

$$\begin{aligned}\mathrm{Var}(t_x) &= n\sigma_x^2 - \frac{n(n-1)}{N(N-1)}\sigma_x^2 \\ &= \frac{N-n}{N-1} n\sigma_x^2 \\ &\approx (1-f)n\sigma_x^2. \quad (2.22)\end{aligned}$$

where $f = n/N$ is known as the *sampling fraction*.

Under *srswor* the sample size n is fixed; then, the variance of the sample mean is given by

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}(t_x/n) = \frac{1}{n^2}\mathrm{Var}(x) \approx \frac{1-f}{n}\sigma_x^2. \quad (2.23)$$

Also,

$$\mathrm{Var}(N\bar{x}) = N^2\mathrm{Var}(\bar{x}) \approx \frac{N(1-f)}{n}\sigma_x^2. \quad (2.24)$$

2.6 Variances Estimation and Standard Errors

To estimate the mean of variable X in the frame, (μ_x), we use the mean of the n sampling units: $\bar{x} = t_x/n = \sum x_i/n$.

Under *srswor*, we estimate the variance of the sample mean, using the sample variance, s_x^2:

$$\widehat{\mathrm{Var}}(\bar{x}) \approx \frac{1-f}{n}s_x^2, \quad (2.25)$$

where

$$s_x^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2. \tag{2.26}$$

Therefore, the "*standard error of the mean*" is estimated as the positive square root of the variance of the sample mean,

$$s_{\bar{x}} \approx \sqrt{1-f}\frac{s}{\sqrt{n}}. \tag{2.27}$$

If we want to estimate the total of variable X in the frame, $N\mu_x$, we use $N\bar{x}$ as estimator, whose standard error is estimated,

$$s_{N\bar{x}} \approx N\sqrt{1-f}\frac{s}{\sqrt{n}} = N\, s_{\bar{x}}. \tag{2.28}$$

Because, in practice, most of the times the sampling fraction $f = n/N$ is small, then, the multiplier $1 - f \to 1$.

Note the following illustration:

$f =$	0.10	0.05	0.01	0.001	0.0001
$\sqrt{1-f} \approx$	0.95	0.97	0.99	0.999	0.9999

2.7 Estimation of a Proportion

To estimate a proportion, p_y, of a binary variable $Y = (0, 1)$ in the frame, we use the sampling proportion, $\hat{p}_y = t_y/n$, where $t_y = \sum_1^n y_i$ is the total of sampling units which have the characteristic of interest.

Under *srswor*, the variance of the sampling proportion is estimated using

$$\widehat{\mathrm{Var}}(\hat{p}_y) \approx (1-f)\frac{\hat{p}_y(1-\hat{p}_y)}{n-1} \tag{2.29}$$

Then, the standard error of the sampling proportion is estimated as

$$s_{\hat{p}_y} \approx \sqrt{(1-f)\frac{\hat{p}_y(1-\hat{p}_y)}{n-1}}. \tag{2.30}$$

Commonly, in the basic statistic texts, the denominator of the previous equation is n instead of $n-1$, even though sampling *without replacement* is assumed. This is due to the fact that for large sample sizes, there will be no substantial difference.

The coefficient of variation estimator of the sampling proportion is given by

$$\mathrm{CV}(\hat{p}_y) \approx \sqrt{(1-f)\frac{(1-\hat{p}_y)}{\hat{p}_y(n-1)}}. \tag{2.31}$$

Under *srswr*, the multiplier $1 - f$ does not appear because the size of the frame does not vary from selection to selection and, essentially, is equivalent to sample from an infinite population (or a process).

2.8 Sample Size Under *SRS*

Remark *The size of sample is no criterion of its precision, nor of its accuracy, nor of its usefulness. The procedure of stratification, the choice of sampling unit, the formulas prescribed for the estimations, are more important than size in the determination of precision. Once these features are fixed, then as we can increase the size of the sample drawn with random numbers, we gain precision (though the point of diminishing returns comes rapidly).*[a]

[a] W Edwards Deming; 1960, 1990. *Sample Design in Business Research.* New York: Wiley Classics Library. Page 28.

As we previously observed, under *SRS*, the variance of the sample mean is

$$\begin{aligned} \mathrm{Var}(\bar{x}) &\approx \frac{1-f}{n}\sigma_x^2 \\ &= \frac{\sigma_x^2}{n'}, \end{aligned} \tag{2.32}$$

where

$$n' = \frac{n}{1-f}. \tag{2.33}$$

Then, for a required variance of the sample mean (or its coefficient of variation), the sample size is calculated using

$$n' = \frac{\sigma_x^2}{\mathrm{Var}(\bar{x})} = \frac{\sigma_x^2/\mu_x^2}{\mathrm{Var}(\bar{x})/\mu_x^2} = \frac{\mathrm{RV}_x}{\mathrm{RV}_{\bar{x}}}. \tag{2.34}$$

Then,

$$n = \frac{n'}{1+n'/N}. \tag{2.35}$$

It is assumed that the researcher knows the variance (or coefficient of variation) in the frame. Of course, this measure is typically unknown; thus, we need an estimate based on a previous study, or from the variance of a similar variable.

Alternatively, to calculate the required sample size, we can use the relative variance of the sample mean. This is given by

$$\mathrm{RV}(\bar{x}) = (1-f)\frac{\mathrm{RV}_x}{n} = \frac{\mathrm{RV}_x}{n'} \tag{2.36}$$

where as before, $n' = n/(1-f)$.

Then,

$$n' = \frac{\mathrm{RV}_x}{\mathrm{RV}_{\bar{x}}} \tag{2.37}$$

and, again,

$$n = \frac{n'}{1 + n'/N}. \tag{2.38}$$

Equivalently, for the sample proportion, we notice that its variance is,

$$\begin{aligned} \mathrm{Var}(\hat{p}_y) &\approx \frac{1-f}{n}\sigma_y^2 \\ &= \frac{\sigma_y^2}{n'}, \end{aligned} \tag{2.39}$$

where

$$n' = \frac{n}{1-f} \tag{2.40}$$

Then, for a required variance of the sample proportion,

$$n' = \frac{\mathrm{RV}_y}{\mathrm{RV}_{\hat{p}_y}} \tag{2.41}$$

and, again,

$$n = \frac{n'}{1 + n'/N}. \tag{2.42}$$

Alternatively, using the relative variance of the sample proportion,

$$\mathrm{RV}(\hat{p}_y) = (1-f)\frac{1-p_y}{np_y} = \frac{\mathrm{RV}_y}{n'} \tag{2.43}$$

where, as before, $n' = n/(1-f)$.

Then, for a required relative variance of the sample proportion,

$$n' = \frac{\mathrm{RV}_y}{\mathrm{Var}(\hat{p}_y)}. \tag{2.44}$$

Finally, we obtain as sample size (as above)

$$n = \frac{n'}{1 + n'/N}. \tag{2.45}$$

Exercises

Knowledge is a scarce national resource. *Knowledge in any country is a national resource. Unlike rare metals, which can not be replaced, the supply of knowledge in any field can be increased by education. Education can be formal, as in school. It may be informal, by study at home or on the job. It may be supplemented and rounded out by work and review under a master. A company must, for its very existence, make use of the store of knowledge that exists within the company, and learn how to make use of help from the outside when it can be effective.*[a]

[a] W Edwards Deming; 1986. *Out of the Crisis*. Cambridge, MA: MIT CAES. Page 466.

1. An urn contains five chips, numbered $\{x = 1, 2, \ldots, 5\}$; three are yellow, and two are green. You are going to select a sample of two chips.

- List the number of possible samples if sampling were without replacement;
- Form the sampling distribution of $\bar{x}$, the average of x in the sample;
- Form the sampling distribution of $\hat{p}_g$, the proportion green in the sample;
- Verify that $\mathrm{E}(\bar{x}) = \mu_x$, the mean of x in the frame;
- Verify that $\mathrm{E}(\hat{p}_g) = p_g$, the proportion green in the frame;
- Use the results of sampling theory to obtain $\mathrm{Var}(\bar{x})$ and $\mathrm{Var}(\hat{p})$.

2. Consider a frame of five employees: A, B, C, D, & E have sex: F, M, M, F, & F; and years of experience: 3.5, 5, 4.5, 2.5, & 6.

- Summarize their mean experience μ_x, and proportion female p_F;
- How many samples of size two can be selected (without replacement); enumerate them;
- For each of the selected samples estimate the mean experience $\bar{x}_i$ and proportion female $\hat{p}_F$;
- Obtain the mean of the sample means, and the mean of the sample proportions. The results should coincide with μ_x, mean experience of the all employees, and proportion female p_F, in the frame.

3. A sample (*srswor*) of 50 households drawn from a community of 800 households, reveals that: 12 of the head were single mothers.

- Estimate the proportion of single mothers in the community, its standard error, and give a 95% confidence interval for said proportion;

- What is the estimate of the total single mothers in the community? Obtain its standard error, and give a 95% confidence interval for said total;
- What is the estimated coefficient of variation in both cases?

4. Using program *R*, examine package UsingR, select its `cfb` data frame, with the following commands:

```
library(UsingR)
data(cfb)
attach(cfb)
str(cfb)
```

Using function `sample( )`, take a sample of 100 cases, without replacement, of variable `INCOME`. Prepare a histogram and a numerical summary of said variable. Notice that because of independent random selection, each participant will get different results. Compare your results with variable `INCOME` in the frame.

5. Repeat the previous exercise for variables `AGE, EDUC`, and `SAVING`.

Expert knowledge, judgement, sincerity, and honesty, are all necessary ingredients of any science, but they are not sufficient to make a sample. There is no substitute for the use of statistical theory.[a]

[a] W Edwards Deming; 1960, 1990. *Sample Design in Business Research.* New York: Wiley Classics Library. Page 28.

Chapter 3

Systematic and Replicated Sampling

Replication originated with Mahalanobis in 1936, and it is a pleasure to express my appreciation for the privilege of studying his method in India, in 1946 and again in 1951 and 1952. He uses the term interpenetrating subsamples, which has much merit ... Replicated sampling went under the name of the Tukey plan in my earlier papers and in my book Some Theory of Sampling ... in respect to my friend Professor John W Tukey, who in 1948 took the trouble to persuade me to use 10 interpenetrating subsamples in a certain application, to eliminate the labor of computing the standard errors.[a]

[a] W Edwards Deming; 1960, 1990. *Sample Design in Business Research*. New York: Wiley Classics Library. Page 186 *et seq*.

3.1 Introduction to Systematic Sampling

As we pointed out, *simple random sampling* forms the basis of probability sampling. Other sampling schemes and techniques aim to improve the efficiency, to reduce sampling errors. A convenient method for sample selection and analysis is *systematic sampling*. This technique maintains the characteristics of probability sampling, commencing the sample selection with one or more *random starts*, following certain pattern for subsequent sample selections until the samples size is obtained.

Let N be the size of the sampling frame, made of units $U_1, U_2, \ldots, U_N$, and n is the sample size. Compute $k = N/n$, known as the *sampling interval* (the reciprocal of the sampling fraction), and take the integer nearest to k.

Definition 15: Random Start.
A number r, randomly selected between 1 and k is known as a

random start. The rest of the sample selection depends on this random start, in which we select every other k^{th} sampling unit until the n units are obtained.

Definition 16: Linear Systematic Sampling.
In case that $k = N/n$ is an integer, the sample selection is an easy task. Just select a random start, r, between 1 and k; the n selected sampling units will be numbered:

$$r,\, r+k,\, r+2k,\, \ldots,\, r+(n-1)k. \tag{3.1}$$

This is known as *linear systematic sampling* with a random start.

For example, suppose that the frame has $N = 2000$ sampling units, $U_1, U_2, \ldots, U_{2000}$, and $n = 250$. Then $k = 8$, which means that we will select a random start, r, between 1 and 8. If for example, $r = 3$, then the n selected sampling units will be numbered:

$$3,\, 3+8,\, 3+2\cdot 8,\, \ldots,\, 3+(250-1)8. \tag{3.2}$$

In other words, the selected sampling units will be numbered U_3, U_{11}, U_{19} $\ldots$, U_{1995}. Notice that the selected sample is evenly distributed in reference to the sampling frame. Furthermore, as in *srswor*, no sampling unit appears more than once in the selected sample.

An alternative to linear systematic selection is the following method.

Definition 17: Circular Systematic Sampling.
Let $k = N/n$, or the nearest integer to N/n; then, select a random start, r, between 1 and N, the size of the sampling frame. The sample of n units is systematically obtained as those corresponding to the numbers:

$$\begin{aligned} r+jk \quad &\text{if} \quad r+jk \le N \\ r+jk-N \quad &\text{if} \quad r+jk > N. \end{aligned} \tag{3.3}$$

for $j = 0,\, 1\, \ldots,\, (n-1)$.

This method of selection is known as *circular systematic sampling* with a random start.

For example, suppose that we have a very simple frame made of 10 sampling units, $U_1, U_2, \ldots, U_{10}$, and that we want to select $n = 3$ units. This implies that $k = 3$. We choose a random start between 1 and 10; suppose that $r = 5$. Then, our sample will be made of those units corresponding to the numbers: 5, 8, and 1. That is, the selected sampling units will be U_1, U_5, and U_8.

3.2 Some Advantages of Systematic Sampling

Systematic sampling offers several advantages, besides selection convenience. It is operationally simple, which can be of great relevance in large-sample studies. The training of field supervisors and interviewers and of other people working in the study is greatly simplified. In a field study, systematic sampling can be used to select the sample progressively, where the field workers are provided with sampling intervals based on prior information.

Systematic sampling can even be used for quick preliminary tabulation of census results using say, a 10% sample of people in the study.

3.3 Sampling Probability

We note that in linear systematic sampling with a random start, the probability of selection for any possible sample is $1/k$. For example, in a simplified frame of six sampling units $U_1, U_2, \ldots, U_6$, and $n = 2$, then $k = 3$. The distribution of possible samples, with the corresponding probabilities are:

r	units	Pr
1	$1, 4$	$1/3$
2	$2, 5$	$1/3$
3	$3, 6$	$1/3$

If x is a quantity of interest, with mean μ_x and variance $\sigma_x^2 = \sum_{i=1}^{N}(x_i - \mu_x)^2/N$, the sampling distribution of mean $\bar{x}_r = \sum_{i=1}^{n} x_{ri}/n$, for $r = 1, 2, \ldots, k$ has expected value:

$$\mathrm{E}(\bar{x}_r) = \sum_{r=1}^{k} \bar{x}_r \frac{1}{k} = \mu_x. \tag{3.4}$$

And variance given by

$$\mathrm{Var}(\bar{x}_r) = \sum_{r=1}^{k} (\bar{x}_r - \mu_x)^2 \frac{1}{k} \equiv \sigma_b^2. \tag{3.5}$$

It can be shown that if we split the variance σ_x^2 into the *between-sample* variance, σ_b^2, and *within-sample* variance, σ_w^2: $\sigma_x^2 = \sigma_b^2 + \sigma_w^2$, where

$$\sigma_w^2 = \sum_{r=1}^{k} \sigma_{wr}^2 \frac{1}{k}, \quad \text{for} \quad \sigma_{wr}^2 = \sum_{i=1}^{n} (x_{ri} - \bar{x}_r)^2 \frac{1}{n}, \tag{3.6}$$

then,

$$\mathrm{Var}(\bar{x}_r) = \sigma_x^2 - \sigma_w^2. \tag{3.7}$$

Moral of Systematic Sampling

The above result can be interpreted by saying that in order to reduce the sampling error (and thus, to increase precision), we should arrange the sampling frame such that the sampling units within each systematic sample are as heterogeneous as possible with respect to the measure of interest. This implies that similar units within the sampling frame should be put together with respect to the measure of interest. Thus, one way of improving the sampling precision is by arranging the sampling units in certain *order* (ascending or descending) with respect to the measure of interest.

> **Definition 18: Intra-class Correlation**.
> The correlation between pairs of sampling units within the systematic sample is known as the *intra-class correlation coefficient*. This is given by,
>
> $$\rho_c = \frac{\sum_{r=1}^{k}\sum_{i=1}^{n}\sum_{i\neq i'}^{n}(x_{ri}-\mu_x)(x_{ri'}-\mu_x)}{N(n-1)\sigma_x^2}. \tag{3.8}$$

The intra-class correlation coefficient has range: $-\frac{1}{n-1} \leq \rho_c \leq 1$. And it can be shown that

$$\text{Var}(\bar{x}_r) = \frac{\sigma_x^2}{n} \times (1 + (n-1)\rho_c). \tag{3.9}$$

Therefore, if ρ_c takes as large, negative value, then $\text{Var}(\bar{x}_r)$ decreases, and thus, the sampling precision increases. And we end up with the same message, namely: the sampling units within each systematic sample should be as heterogeneous as possible with respect to the measure of interest.

Exercises

1. Suppose that a book of $N = 500$ pages will be examined in order to estimate the total number of errors. For our purpose, we will randomly select a sample of $n = 30$ pages.

- Explain how you would select the sample if you use *srswor*;
- Explain how you would select the sample if you use *linear systematic sampling*;
- Explain how you would select the sample if you use *circular systematic sampling*.

2. Suppose that you are assigned a task of selecting a sample of employees from an agency that has $N = 3,000$ employees. The purpose is to estimate various parameters related to years of experience. The proposed sampling plan is to randomly select a sample of $n = 250$ employees.

- Explain how you would select the sample if you use *srswor*;
- Explain how you would select the sample if you use *linear systematic sampling*;
- Explain how you would select the sample if you use *circular systematic sampling*.

3. Suppose that you are assigned a task of selecting a sample of units of a product from a lot that has $N = 1,000$ units. The purpose is to estimate the proportion p_d of defective units in the lot. The proposed sampling plan is to randomly select a sample of $n = 50$ units.

- Explain how you would select the sample if you use *srswor*;
- Explain how you would select the sample if you use *linear systematic sampling*;
- Explain how you would select the sample if you use *circular systematic sampling*.

3.4 Systematic Sampling Using Program *R*

What follows is a set of commands to obtain a systematic sample, using program *R*. First, we present the commands to obtain the required sample indicating both the sampling frame size, N, and the sample size, n.

3.4.1 Linear Systematic Sampling

Suppose that we want to solve exercise 1 above, using *linear* systematic sampling with program *R*. Then, we could use the following function:

```
linear.sample=function(N=500, n=30){
 k=round(N/n, 0)
 r = sample(1:k, 1)
 syst.samp= seq(r, r+k*(n-1), k)
 print(paste("Your Linear Systematic Sample is:  ", return(syst.samp)))
}
```

Then, for example, to obtain the desired sample, type the command,

```
linear.sample(500, 30)
```

3.4.2 Circular Systematic Sampling

Suppose that we want to solve exercise 1 above, using *circular* systematic sampling with program *R*. Then, we could use the following function:

```
circ.sample=function(N=500, n=30){
 k=round(N/n, 0)
 r = sample(1:N, 1)
 f=n-1
 circ.syst=r
 for (j in 1:f) {
 if (r + j*k <= N)
 circ.syst = c(circ.syst, r+j*k)
 else
 (circ.syst = c(circ.syst, r+j*k-N))
 }
 print(paste("Your Circular Systematic Sample is:  ", return(circ.syst)))
}
```

Then, for example, to obtain the desired sample, type the command,

```
circ.sample(500, 30)
```

3.5 Introduction to Replicated Sampling

As we pointed out in Chapter Two, *Replicated Sampling* consists on selecting $g \geq 2$ independent subsamples using the same sampling design procedure, so that each of these subsamples can provide valid estimates of the parameter of interest. Replicated sampling is also known as *interpenetrating sub-samples*, a phrase coined by PC Mahalanobis, who was Director of the Indian Statistical Institute in Calcutta. In Part II of his book, *Sample Design in Business Research*, Deming describes the use and usefulness of replicated sampling.

This sampling technique facilitates the estimation of the variance of the estimator, even for more elaborate sample designs, where it estimation could be complicated. If $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_g$ are estimates for parameter θ, based on g independent subsamples, then

$$\hat{\theta} = \sum_{i=1}^{g} \hat{\theta}_i \frac{1}{g} \tag{3.10}$$

will be an unbiased estimator of θ, where the variance of $\hat{\theta}$ is estimated as

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{g(g-1)} \sum_{i=1}^{g} (\hat{\theta}_i - \hat{\theta})^2. \tag{3.11}$$

Notice that if $g = 2$, then

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{4}(\hat{\theta}_1 - \hat{\theta}_2)^2. \tag{3.12}$$

The sample, so divided into g independent subsamples not only facilitates the estimation of sampling errors, but also helps in detecting and correcting non-sampling errors as well, thus increasing the reliability of sample surveys.

3.6 Estimation of μ_x in *R-SRS*

Suppose that we want to estimate the frame mean, μ_x, selecting n sampling units from a frame using *replicated-simple random sampling*. We randomly draw m units from each of the $g \geq 2$ independent subsamples, where $mg = n$. Then from each subsample, we get an estimator $\bar{x}_i = \sum_{j=1}^{m} x_j/m$, of μ_x, for $i = 1 : g$, with sampling variance σ_x^2/m. Then, by equations 1 and 2, above,

$$\bar{x} = \sum_{i=1}^{g} \bar{x}_i \frac{1}{g}, \tag{3.13}$$

and

$$\widehat{\text{Var}}(\bar{x}) = \frac{1}{g(g-1)} \sum_{i=1}^{g} (\bar{x}_i - \bar{x})^2. \tag{3.14}$$

Notice that if $g = 2$, then

$$\widehat{\text{Var}}(\bar{x}) = \frac{1}{4}(\bar{x}_1 - \bar{x}_2)^2. \tag{3.15}$$

3.6.1 Estimation of p_y in *R-SRS*

Similarly, if $y = (0, 1)$ is a dichotomous variable, and we want to estimate p_y, the proportion of interest in the frame, we randomly draw m units from each of the g independent subsamples ($mg = n$). Then from each subsample, we get an estimator $\hat{p}_i = \sum_{j=1}^{m} y_j/m$, of p_y, for $i = 1 : g$, with sampling variance $p_y(1 - p_y)/m$. Then, by equations 1 and 2, above,

$$\bar{p}_y = \sum_{i=1}^{g} \hat{p}_i \frac{1}{g}, \tag{3.16}$$

and

$$\widehat{\text{Var}}(\bar{p}_y) = \frac{1}{g(g-1)} \sum_{i=1}^{g} (\hat{p}_i - \bar{p}_y)^2. \tag{3.17}$$

Notice that if $g = 2$, then

$$\widehat{\text{Var}}(\bar{p}_y) = \frac{1}{4}(\hat{p}_1 - \hat{p}_2)^2. \tag{3.18}$$

3.7 Replicated-SRS Using Program *R*

Suppose that we want to select a *replicated simple random sample* of size n from a frame containing N sampling units. The n randomly selected units will be divided into $g \geq 2$ independent sub-samples, each containing $m = n/g$ units. Notice that m has to be a multiple of n and g.

A very simple R function, which accomplish this task follows:

```
# Replicated-SRS Using R
replic.srs = function(N=500, n=30, g=6){
m=n/g
samp=sample(1:N, n)
samp=matrix(samp, nrow=m)
samp
}
```

Then, for example, to obtain the desired sample, type the command,

```
replic.srs(500, 30, 6)
```

3.8 Estimation of μ_x in *R-SysRS*

Suppose that we want to estimate the frame mean, μ_x, selecting n sampling units from a frame using *replicated-systematic random sampling*. Now, we need two or more independent random starts, between 1 and $k = N/m$, if we use *linear* systematic sampling, or between 1 and N, if we are using *circular* systematic sampling. We systematically draw m units, using g different random starts from each of the $g \geq 2$ independent subsamples, $(mg = n)$. Then from each subsample, we get an estimator $\bar{x}_i = \sum_{j=1}^{m} x_j/m$, of the frame mean μ_x, for $i = 1 : g$, with sampling variance σ_x^2/m. Then, by equations 1 and 2, above,

$$\bar{x} = \sum_{i=1}^{g} \bar{x}_i \frac{1}{g}, \tag{3.19}$$

and

$$\widehat{\text{Var}}(\bar{x}) = \frac{1}{g(g-1)} \sum_{i=1}^{g} (\bar{x}_i - \bar{x})^2. \tag{3.20}$$

Notice that if $g = 2$, then

$$\widehat{\text{Var}}(\bar{x}) = \frac{1}{4}(\bar{x}_1 - \bar{x}_2)^2. \tag{3.21}$$

3.8.1 Estimation of p_y in *R-SysRS*

Suppose that we want to estimate a proportion of interest in the frame, p_y, selecting n sampling units using *replicated-systematic random sampling*. Now,

we need two or more independent random starts, between 1 and $k = N/m$, if we use *linear* systematic sampling, or between 1 and N, if we are using *circular* systematic sampling. If $y = (0, 1)$ is a dichotomous variable, and we want to estimate p_y, we systematically draw m units, using g different random starts from each of the $g \geq 2$ independent subsamples ($mg = n$). Then from each subsample, we get an estimator $\hat{p}_i = \sum_{j=1}^{m} y_j/m$, of p_y, for $i = 1 : g$, with sampling variance $p_y(1 - p_y)/m$. Then, by equations 1 and 2, above,

$$\bar{p}_y = \sum_{i=1}^{g} \hat{p}_i \frac{1}{g}, \tag{3.22}$$

and

$$\widehat{\text{Var}}(\bar{p}_y) = \frac{1}{g(g-1)} \sum_{i=1}^{g} (\hat{p}_i - \bar{p}_y)^2. \tag{3.23}$$

Notice that if $g = 2$, then

$$\widehat{\text{Var}}(\bar{p}_y) = \frac{1}{4}(\hat{p}_1 - \hat{p}_2)^2. \tag{3.24}$$

We remark that for *replicated sampling*, although the estimation formulae in both *simple random sampling* and *systematic sampling* are equivalent, they are based on different concepts. In *replicated-simple random sampling*, the number of possible samples is $g \times \binom{N}{m}$, while in *replicated-systematic sampling*, the number of possible samples is $g \leq k$.

We also emphasize that *replicated-systematic sampling* can be used in combination with other probability sampling designs like *stratified sampling* and *clusterred sampling*.

3.9 Replicated Linear-Systematic Sample Using *R*

Suppose that we want to select a *replicated linear-systematic sample* of size n from a frame containing N sampling units. The n randomly selected units will be divided into $g \geq 2$ independent sub-samples, each containing $m = n/g$ units. Now, we need g random starts between 1 and $k = N/m$. Notice that m has to be a multiple of n and g.

A very simple R function, which accomplishes this task follows:

```
# Replicated Linear-Systematic Sample Using R
linear.replic=function(N=500, n=30, g=6){
 m=round(n/g,0)
 k=round(N/m, 0)
 syst.samp=NULL
 for (i in 1:g){
```

```
 r = sample(1:k, 1)
 syst.samp= c(syst.samp, seq(r, k*m, k))
 }
 syst.samp=matrix(syst.samp, nrow=m)
print(paste("Your Replicated Linear-Systematic Sample is:  ", return(syst.samp)))
}
```

Then, for example, to obtain the desired sample, type the command:

```
linear.replic(500, 30, 6)
```

3.9.1 Replicated Circular-Systematic Sample Using *R*

Suppose that we want to select a *replicated circular-systematic sample* of size n from a frame containing N sampling units. The n randomly selected units will be divided into $g \geq 2$ independent sub-samples, each containing $m = n/g$ units. Now, we need g random starts between 1 and $k = N/m$. Notice that m has to be a multiple of n and g.

A very simple program which accomplishes this task follows:

```
# Replicated Circular-Systematic Sample Using R
circular.replic=function(N=500, n=30, g=6){
 m=round(n/g,0)
 k=round(N/m, 0)
 circ.syst=NULL
 for (i in 1:g){
 r = sample(1:N, 1)
 f=m-1
 for (j in 0:f){
 if (r + j*k <= N)
 circ.syst = c(circ.syst, r+j*k)
 else
 (circ.syst = c(circ.syst, r+j*k-N))
 }
 }
 circ.syst=matrix(circ.syst, nrow=m)
print("Your Replicated Circular-Systematic Sample is:  ", return(circ.syst))
}
```

Then, for example, to obtain the desired sample, type the command,

```
circular.replic(500, 30, 6)
```

Exercises

1. Suppose that a book of $N = 500$ pages will be examined in order to estimate the total number of errors. For our purpose, we will randomly select a sample

of $n = 30$ pages.

- Explain how you would select the sample if you use *replicated-srswor* and $g = 6$;
- Explain how you would select the sample if you use *replicated-linear systematic sampling* and $g = 6$;
- Explain how you would select the sample if you use *replicated-circular systematic sampling* and $g = 6$.

2. Suppose that you are assigned a task of selecting a sample of employees from an agency that has $N = 3,000$ employees. The purpose is to estimate various parameters related to years of experience. The proposed sampling plan is to randomly select a sample of $n = 250$ employees.

- Explain how you would select the sample if you use *replicated-srswor* and $g = 10$;
- Explain how you would select the sample if you use *replicated-linear systematic sampling* and $g = 10$;
- Explain how you would select the sample if you use *replicated-circular systematic sampling* and $g = 10$.

3. Suppose that you are assigned a task of selecting a sample of units of a product from a lot that has $N = 1,000$ units. The purpose is to estimate the proportion p_d of defective units in the lot. The proposed sampling plan is to randomly select a sample of $n = 50$ units.

- Explain how you would select the sample if you use *replicated-srswor* and $g = 5$;
- Explain how you would select the sample if you use *replicated-linear systematic sampling* and $g = 5$;
- Explain how you would select the sample if you use *replicated-circular systematic sampling* and $g = 5$.

Chapter 4

Basics of Stratified Sampling

Remark. *Such a calculation of "significance" takes account only of the numerical data of this one experiment. An estimate of σ unless the observations have demonstrated randomness, ... and not unless the number of degrees of freedom ... amount to be 15 or 20, and preferably more. A broad of background of experience is necessary before one can say whether his experiment is carried of by demonstrably random methods. Moreover, even in the state of randomness, it must be borne in mind that unless the number of degrees of freedom is very large, a new experiment will give new values of both $\sigma(ext)$ and $\sigma(int)$, also of $P(\chi)$ and $P(z)$. Ordinarily, there will be a series of experiments, and a corresponding series of P values. It is the **consistency** of the P values of the series, under a wide variety of conditions, and not the smallness of any one P value by itself that determines a basis for action, particularly when we are dealing with a cause system underlying scientific law ... In the absence of a large number of experiments, related knowledge of the subject and scientific judgement must be relied on to a great extent in framing a course of action. Statistical "significance" by itself is not a rational basis for action .*[a]

[a] W Edwards Deming; 1943, 1964. *Statistical Adjustment of Data*. New York: Dover Publications. Page 30.
N.B. $\sigma(ext)$ and $\sigma(int)$ represent external and internal variability of an experiment.

4.1 Introduction

As we pointed out in Chapter One, in ***stratified sampling***, previous to the sample selection, the sampling frame is divided into few, H say, homogeneous groups, known as *strata*. These groups are formed based on certain auxiliary

variable, which would help to increase sampling precision of the measure(s) of interest in the research under study. These non-overlapping groups or strata, could be formed by *e.g.*, geographical areas, age-groups, or sex. A sample is randomly taken from each stratum, and then this sample is referred to as a *stratified random sample*. The process is called *statification* and its objective is to obtain as much homogeneity as possible within a group or *stratum* and to mantain heterogeneity between groups or strata. Thus, stratification is the process of grouping members of the frame into relatively homogeneous subgroups before sampling. Then random or systematic sampling is performed within each stratum. There are several designs or schemes of stratified sampling.

The strata should be mutually exclusive, *i.e.*, every element in the sampling frame must be assigned to only one stratum. The strata should also be collectively exhaustive, *i.e.*, no sampling unit in the frame can be excluded.

Let N be the size of the sampling frame, and N_h be the size of stratum h, where $\sum N_h = N$, for $h = 1 : H$. Stratified sampling proceeds by randomly selecting n_h sampling units from N_h units in stratum h, where $\sum n_h = n$. For example, suppose that we have two strata of sizes $N_1 = 200$ and $N_2 = 300$; where $\sum N_h = 500$. Take a random selection of $n_1 = 20$ sampling units from stratum one and, independently, take a random selection of $n_2 = 30$ sampling units from stratum two, where $n = \sum n_h = 50$. Notice that now, we can obtain separate estimates from each stratum and also, combine these estimates in order to get estimates of the entire sampling frame.

Some examples of situations where you might want to use stratified sampling are:

1. Health care costs – stratify based on patient's age;
2. Socio-economic survey – stratify based on home value;
3. Yield of farm product – stratify based on farm size, *e.g.*: *Small* (< 50 acres), *Medium* (between 50 and 100 acres), *Large* (> 100 acres);
4. Study of prevalence of a disease in a country – stratify by health region;
5. Employee income – stratify on years of experience and sex.

When choosing a criterium to stratify on, we should use an instrumental variable that is associated to the variable of interest as this should make the precision per stratum small.

4.2 Some Advantages of Stratified Sampling

When sub-populations vary considerably, it is advantageous to sample each sub-population or stratum independently. And, provided that the strata are formed so that members of the same stratum are as similar as possible with

respect of the characteristic of interest, stratification will always achieve greater precision than simple random sampling. The more heterogeneity or the bigger the differences between the strata, the greater the gain in precision.

Stratification often improves the representativeness of the sample by reducing sampling error. It can produce a weighted estimate that has less variability than the same estimate under a simple random sample of the sampling frame.

It is often administratively convenient to stratify a sample. Interviewers for example, can be specifically trained to deal with a particular age- or sex-group, or employees in a particular industry.

The results from each stratum may be of intrinsic interest in a study and by stratifying, can be analyzed separately.

Stratification also ensures better coverage of the population than simple random sampling.

Stratification also allows the use of different sampling techniques for different sub-groups.

4.2.1 Some Disadvantages of Stratified Sampling

On occasions, it is difficult to identify appropriate strata. Or simply, the frame is naturally stratified and there is no need for further stratification.

Usually, it is more complex to organize and analyze the results, compared with simple random sampling.

Sometimes, it can be difficult to select relevant stratification variables. It is not useful when the groups formed by the stratification process are not homogeneous with respect to the measure(s) of interest in a study.

Stratified sampling can be expensive. It requires accurate information about the population represented in the sampling frame.

4.3 Proportional Stratified Sampling

The technique known as *proportional stratified sampling* is the most common stratified sampling design. By definition, in proportional stratified sampling design, for each stratum h,

$$\frac{n_h}{n} = \frac{N_h}{N}, \quad \text{for } h = 1 : H. \tag{4.1}$$

This is known as *proportional allocation*, and it means that each stratum is represented in the sample in proportion to the stratum size in the sampling frame; this proportion is

$$P_h = \frac{N_h}{N}, \quad \text{where } \sum_{h=1}^{H} P_h = 1. \tag{4.2}$$

For example, for two strata, if the sampling frame is made of 60% units in stratum one and 40% units in stratum two, then the relative size of the two independent samples should reflect these proportions, *i.e.*, the sample size allocation will be $n_1 = 0.6\,n$ and $n_2 = 0.4\,n$.

4.4 Other Stratified Sampling Allocations

There are a number of different stratified schemes, besides *proportional allocation*, that can be used, some better than others. We mention two of them:

1. Equal allocation – Where we set $n_1 = n_2 = \cdots = n_H$;
2. Optimal allocation – We choose $n_1, \ldots, n_H$ which minimizes $\mathrm{Var}(\hat{\mu}_{str})$ for a given sample size n, *i.e.*,

$$\begin{aligned} n_h &= n\frac{P_h\,\sigma_h}{\sum_{j=1}^{H} P_j\,\sigma_j} \\ &\propto P_h\,\sigma_h \end{aligned} \tag{4.3}$$

   i.e., the sample allocation to stratum h will depend on both the proportion of sampling units in the frame, P_h, and the stratum variability, σ_h.

Both $\mathrm{Var}(\hat{\mu}_{str})$ and σ_h are given below. We notice that optimal allocation implies that there will be more observations in the more variable strata. And, if the cost of gathering information differs from stratum to stratum, it can be shown that,

$$n_h \propto P_h\,\sigma_h/\sqrt{c_h}, \tag{4.4}$$

where c_h is the per unit cost in stratum h. Thus, the larger the cost of obtaining information from stratum h, the smaller the sampling units selected from said stratum.

4.4.1 Proportional Allocation Using *R* (Two Strata)

Suppose that the N sampling units in the frame are divided into two strata: the first N_1 units belong to stratum one, the other N_2 units belong to stratum two. Furthermore, suppose that we will use a proportional allocation design.

A very simple R program to obtain a proportional stratified sample follows:

```
# Proportional Stratified Sample Using R
prop.sample=function(N=500, N1=200, N2=300, n=30){
P1=round(N1/N, 2); P2=round(1-P1, 2)
n1 = round(n*P1, 0); n2 = n - n1
stratum.1=1:N1; stratum.2=(N1+1):N
pr.samp.1=sample(stratum.1, n1)
```

```
pr.samp.2=sample(stratum.2, n2)
samp=c(pr.samp.1, pr.samp.2)
print("Your Proportional Stratified Sample is:  ", return(samp))
}
```

4.4.2 Equal Allocation Using *R* (Two Strata)

Suppose that the N sampling units in the frame are divided into two strata: the first N_1 units belong to stratum one, the other N_2 units belong to stratum two. Furthermore, suppose that we will use an equal allocation design.

A very simple R program to obtain an equal allocation sample follows:

```
# Equal Stratified Sample Using R
equal.sample=function(N=500, N1=200, N2=300, n=30){
n1 = round(n/2, 0); n2 = n1
stratum.1=1:N1; stratum.2=(N1+1):N
eq.samp.1=sample(stratum.1, n1)
eq.samp.2=sample(stratum.2, n2)
samp = c(eq.samp.1, eq.samp.2)
print("Your Equal Stratified Sample is:  ", return(samp))
}
```

4.4.3 Optimal Allocation Using *R* (Two Strata)

Suppose that the N sampling units in the frame are divided into two strata: the first N_1 units belong to stratum one, the other N_2 units belong to stratum two. Now, we need to know the sample variability of each stratum, σ_1 and σ_2. Furthermore, suppose that we will use an optimal allocation design, with equal per unit cost in each stratum, $c_1 = c_2$.

A very simple R program to obtain an optimal stratified sample follows:

```
# Optimal Stratified Sample Using R
optim.sample=function(N=500, N1=200, N2=300, sigma.1=50, sigma.2=25,
n=30){
P1=round(N1/N, 2); P2=round(1-P1, 2)
n1 = round(n*P1*sigma.1/(P1*sigma.1 + P2*sigma.2), 0); n2 = n - n1
stratum.1=1:N1; stratum.2=(N1+1):N
op.samp.1=sample(stratum.1, n1)
op.samp.2=sample(stratum.2, n2)
samp = c(op.samp.1, op.samp.2)
print("Your Optimal Stratified Sample is:  ", return(samp))
}
```

4.5 Estimation of the mean, μ_x, in the frame

Suppose your that the sampling frame is divided into H strata. Also, suppose that for stratum h, there are N_h sampling units from the frame ($\sum_{h=1}^{H} N_h = N$) and the X-value for the units in stratum h is given by $x_{1h}, x_{2h}, \ldots, x_{N_h h}$.

Let

$$P_h = \frac{N_h}{N} \qquad \text{and} \qquad \mu_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{ih} \tag{4.5}$$

And,

$$\mu_x = \frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h} x_{ih} = \frac{1}{N}\sum_{h=1}^{H} N_h\,\mu_h = \sum_{h=1}^{H} P_h\,\mu_h \tag{4.6}$$

Then, instead of taking a sample of n sampling units using srs from the whole frame, we can take a sample of n_h sampling units using srs from each stratum ($\sum_{h=1}^{H} n_h = n$).

Let $x_{1h}, x_{2h}, \ldots, x_{n_h h}$ be the sample from stratum h and let,

$$\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{ih} \qquad s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(x_{ih} - \bar{x}_h)^2 \tag{4.7}$$

be respectively, the sample mean and sample variance.

Then, an unbiased estimate of the frame mean μ_{str} is given by

$$\hat{\mu}_{str} = \sum_{h=1}^{H}\frac{N_h}{N}\bar{x}_h = \sum_{h=1}^{H} P_h\bar{x}_h. \tag{4.8}$$

The variance of $\hat{\mu}_{str}$ is given by

$$\mathrm{Var}(\hat{\mu}_{str}) \approx \sum_{h=1}^{H} P_h^2\frac{1}{n_h}(1 - f_h)\,\sigma_h^2, \tag{4.9}$$

where $f_h = \frac{n_h}{N_h}$, the sampling fraction in stratum h, and

$$\sigma_h^2 = \frac{1}{N_h}\sum_{i=1}^{N_h}(x_{ih} - \mu_h)^2, \tag{4.10}$$

i.e., the variance of stratum h.

We notice that whether stratified sampling is preferred to SRS depends on the condition,

$$\mathrm{Var}(\hat{\mu}_{str}) < \mathrm{Var}(\hat{\mu}_{srs}) = \frac{1}{n}(1 - f)\,\sigma_x^2, \tag{4.11}$$

i.e., it depends on the choice of sample sizes of the strata, n_h, the variation of the strata means, μ_h, and the strata variances, σ_h^2.

4.5.1 Estimation of the proportion, p_y, in the frame

Suppose your that the sampling frame is divided into H strata. Also, suppose that for stratum h, there are N_h sampling units from the frame ($\sum_{h=1}^{H} N_h = N$) and for variable $Y = (0, 1)$ (a dichotomous variable) the value for the units in stratum h is given by $y_{1h}, y_{2h}, \ldots, y_{N_h h}$.

Let

$$P_h = \frac{N_h}{N} \qquad \text{and} \qquad p_{y_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{ih}. \tag{4.12}$$

And,

$$p_y = \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{ih} = \frac{1}{N} \sum_{h=1}^{H} N_h \, p_{y_h} = \sum_{h=1}^{H} P_h \, p_{y_h}. \tag{4.13}$$

Then, instead of taking a sample of n sampling units using srs from the whole frame, we can take a sample of n_h sampling units using srs from each stratum ($\sum_{h=1}^{H} n_h = n$).

Let $y_{1h}, y_{2h}, \ldots, y_{n_h h}$ be the sample from stratum h and let,

$$\hat{p}_{y_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{ih} \qquad s^2_{\hat{p}_{y_h}} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{ih} - \hat{p}_h)^2 \tag{4.14}$$

be respectively, the sample proportion and sample variance estimates from stratum h.

Then, an estimate of the frame proportion $p_{y_{str}}$ is given by

$$\hat{p}_{y_{str}} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_{y_h} = \sum_{h=1}^{H} P_h \hat{p}_{y_h}. \tag{4.15}$$

The sampling variance of $\hat{p}_{y_{str}}$ is given by

$$\mathrm{Var}(\hat{p}_{y_{str}}) \approx \sum_{h=1}^{H} P_h^2 \frac{1}{n_h} (1 - f_h) \, \sigma_h^2, \tag{4.16}$$

where $f_h = \frac{n_h}{N_h}$, the sampling fraction in stratum h, and

$$\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (y_{ih} - p_{y_h})^2 = p_{y_h}(1 - p_{y_h}), \tag{4.17}$$

i.e., the variance of stratum h.

And, an unbiased estimator of $\mathrm{Var}(\hat{p}_{y_{str}})$ is given by

$$\widehat{\mathrm{Var}}(\hat{p}_{y_{str}}) = \sum_{h=1}^{H} P_h^2 \, (1 - f_h) \frac{\hat{p}_{y_h}(1 - \hat{p}_{y_h})}{n_h - 1}. \tag{4.18}$$

4.6 Variance Comparison

Let assume that for all of these stratified sampling schemes $n_h << N_H$ for all h so the $1 - f_h \approx 1$; thus, we can ignore it. Then, it can be shown that the sampling variance for each of these allocation schemes is given by:

1. Equal allocation (eq)

$$\text{Var}(\bar{x}_{eq}) = \frac{H}{n} \sum_{h=1}^{H} P_h^2 \, \sigma_h^2 \tag{4.19}$$

2. Proportional Allocation (pr)

$$\text{Var}(\bar{x}_{pr}) = \frac{1}{n} \sum_{h=1}^{H} P_h \, \sigma_h^2 \tag{4.20}$$

3. Optimal Allocation (op)

$$\text{Var}(\bar{x}_{op}) = \frac{1}{n} \left(\sum_{h=1}^{H} P_h \, \sigma_h \right)^2 \tag{4.21}$$

4.6.1 Proportional Allocation vs *SRS*

To show when stratified sampling is a better scheme than SRS, we need an expression of the variance in the frame. It can be shown that:

$$\begin{aligned} \sigma_x^2 &= \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{N_h} (x_{ih} - \mu_x)^2 \\ &= \sum_{h=1}^{H} P_h \, \sigma_h^2 + \sum_{h=1}^{H} P_h (\mu_h - \mu_x)^2 \end{aligned} \tag{4.22}$$

Definition 19: Within strata variance.

$$\sigma_w^2 \equiv \sum_{h=1}^{H} P_h \, \sigma_h^2 \tag{4.23}$$

this is known as the *weighted average variance within strata*.

Definition 20: Between strata variance.

$$\sigma_b^2 \equiv \sum_{h=1}^{H} P_h (\mu_h - \mu_x)^2) \tag{4.24}$$

this is known as the *variance between strata*.

Thus, we see that the total variance is

$$\sigma_x^2 = \sigma_w^2 + \sigma_b^2. \tag{4.25}$$

An important result is that proportional stratified sampling is as precise or more than a single SRS of the same total sample size n; *i.e.*, that

$$\text{Var}(\bar{x}_{pr}) \leq \text{Var}(\bar{x}_{srs}), \tag{4.26}$$

because

$$\begin{aligned} \text{Var}(\bar{x}_{srs}) - \text{Var}(\bar{x}_{pr}) &= \frac{1}{n}\left(\sum_{h=1}^{H} P_h\,\sigma_h^2 + \sum_{h=1}^{H} P_h(\mu_h - \mu_x)^2\right) - \frac{1}{n}\sum_{h=1}^{H} P_h\,\sigma_h^2 \\ &= \frac{1}{n}\sum_{h=1}^{H} P_h(\mu_h - \mu_x)^2 \geq 0; \end{aligned} \tag{4.27}$$

i.e., $\text{Var}(\bar{x}_{srs}) - \text{Var}(\bar{x}_{pr}) = \sigma_b^2/n \geq 0$. This result implies that the more heterogeneous the strata are (in terms of strata means), the better proportional sampling will be. Which also implies that when we select a stratification variable, we should to select one that is strongly correlated with our variable of interest.

4.6.2 Proportional Allocation vs Optimal Allocation

Definition 21: Average within strata standard deviation.

$$\bar{\sigma}_w \equiv \sum_{h=1}^{H} P_h\,\sigma_h, \tag{4.28}$$

this is known as the *average within strata standard deviation.*

Now, the advantage of optimal allocation over proportional allocation can be seen with the following result:

$$\begin{aligned} \text{Var}(\bar{x}_{pr}) - \text{Var}(\bar{x}_{op}) &= \frac{1}{n}\sum_{h=1}^{H} P_h\,\sigma_h^2 - \frac{1}{n}\left(\sum_{h=1}^{H} P_h\,\sigma_h\right)^2 \\ &= \frac{1}{n}\sum_{h=1}^{H} P_h(\sigma_h - \bar{\sigma}_w)^2 \geq 0 \end{aligned} \tag{4.29}$$

Thus, we obtain more precision from optimal allocation than with proportional allocation when there is high variability between strata. But in practice, usually, the gain by switching from SRS to proportional allocation is much bigger than by switching from proportional allocation to optimal allocation.

4.7 Replicated Stratified Sampling

In section 3.5 we introduced replicated sampling, as a technique that could be used with different sampling designs. Replicated sampling is based on $g \geq 2$ independent sub-samples, in which each sub-sample provides valid information from a frame. Now, we suppose that each stratum is divided into g independent sub-samples of size m_h, where $\sum m_h = n_h$ and $gm_h = n_h$. Then, an unbiased estimator of the mean of stratum h, μ_h, based on the g subsamples is given by

$$\hat{\mu}_h = \frac{1}{g}\sum_{i=1}^{g} \bar{x}_{hi}, \quad \text{for } h = 1 : H, \tag{4.30}$$

where $\bar{x}_{hi} = \sum_{j=1}^{m_h} x_{hij}/m_h, \quad$ for $i = 1 : g$ is the mean of X_h in each sub-sample. An unbiased estimator of the sampling variance of $\hat{\mu}_h$ in stratum h, is given by

$$\widehat{\text{Var}}(\hat{\mu}_h) = \frac{1}{g(g-1)}\sum_{i=1}^{g}(\bar{x}_{hi} - \hat{\mu}_h)^2 \tag{4.31}$$

Then, an unbiased estimator of μ_x in the frame, based on the g sub-samples for the H strata is obtained by

$$\hat{\mu}_{str} = \frac{1}{H}\sum_{h=1}^{H}\sum_{i=1}^{g} \bar{x}_{hi}, \tag{4.32}$$

whose sampling variance is estimated by

$$\widehat{\text{Var}}(\hat{\mu}_{str}) = \frac{1}{g(g-1)}\sum_{h=1}^{H}\sum_{i=1}^{g}(\bar{x}_{hi} - \hat{\mu}_{str})^2. \tag{4.33}$$

Alternatively, we could obtain an estimate of μ_h for the H strata, based on the g independent samples, as

$$\hat{\mu}_i = \sum_{h=1}^{H} \bar{x}_{hi} \quad \text{for } i = 1 : g, \tag{4.34}$$

Hence, a combined estimator for μ_{str} is the mean of the g estimates $\hat{\mu}_i$, with an unbiased estimator of the sampling variance simply given by

$$\widehat{\text{Var}}(\hat{\mu}_{str}) = \frac{1}{g(g-1)}\sum_{i=1}^{g}(\hat{\mu}_i - \hat{\mu}_{str})^2. \tag{4.35}$$

And, as we pointed out in Section 3.5 , if $g = 2$, then

$$\widehat{\text{Var}}(\hat{\mu}_{str}) = \frac{1}{4}(\hat{\mu}_1 - \hat{\mu}_2)^2. \tag{4.36}$$

4.8 A Limitation of Stratified Sampling

Not everything goes in favor of stratified sampling. It has the limitation that, in optimal allocation, we need to know the strata variances in order to obtain the sample sizes for each stratum, which could be problematic. However, for proportional sampling, we only need to know the fraction of units falling into each stratum. This information is much more readily available or at least, easier to approximate.

Related to this, is the estimation of standard errors. As the strata variances usually are not available, we need to estimate them using the sample variance of each stratum; *i.e.*,

$$\widehat{\text{Var}}(\bar{x}_{str}) \equiv s^2_{\bar{x}_{str}} = \sum_{h=1}^{H} P_h \frac{1}{n_h}(1 - f_h)\, s_h^2 \tag{4.37}$$

Whose corresponding confidence limits for μ_x are obtained by

$$\bar{x}_{str} \pm t_\nu \cdot s_{\bar{x}_{str}}.$$

Exercises

1. The following table present some results of a survey of farms in a country, stratified by size. The variable of interest was the yield per acre of certain product.

Table 4.1: Yield in a Stratified Sample of Farms

	SMALL	MEDIUM	LARGE
N_h	140	112	28
P_h	0.5	0.4	0.1
n_h	20	16	4
$\bar{x}_h$	30	25	20
s_h^2	36	150	90

- Can you determine the stratified design used in the study?
- What is the total number of farms in the country (frame)?
- What is the total number of farms in the sample?
- Obtain an estimate of the:
    1. mean yield for the country ($\bar{x}_{str}$);

2. variance of mean yield for the country ($s^2_{\bar{x}_{str}}$);
3. 95% limits for the mean yield, μ_{str}, in the country.

- Obtain an estimate of the:

1. total yield for the country ($\hat{\tau}_{x_{str}}$);
2. variance of the total yield for the country $\left(s^2_{\hat{\tau}_{x_{str}}}\right)$;
3. 95% limits for the total yield, τ_{str}, in the country.

2. Suppose that we have a deck of $N = 52$ cards, with four denominations called: *Spade, Club, Diamond, Heart.* The first two denominations are black, the others are red; each denomination contains the following cards: $A, 2, 3, \ldots, 10, J, Q, K$. Let suppose that this deck of cards represents an institution which has $N = 52$ employees; furthermore, that the institution has four departments, each of 13 employees. Let X represent an employee's years of experience, where $A = 1$, $J = 11$, $Q = 12$, and $K = 13$ years of experience; the rest of the numbered cards represents an employee's years of experience, *i.e.*, $X = 1, 2, \ldots, 13$. Also, a red card represents a female employee and a black card represents a male employee.

- Obtain μ_x and σ^2_x. Hints:

$$1 + 2 + \cdots + n = n(n+1)/2$$

$$1^1 + 2^2 + \cdots + n^2 = n(n+1)(2n+1)/6.$$

- Divide the frame by department, *i.e.*, into four strata; obtain μ_h and σ^2_h, for $h = 1 : 4$.

1. Using proportional allocation, obtain σ^2_w, σ^2_b and thus, σ^2_x.
2. What is the gain from such stratification scheme as compared with no stratification; *i.e.*, obtain

$$\text{Var}(\bar{x}_{srs}) - \text{Var}(\bar{x}_{pr})?$$

- Divide the frame by sex, *i.e.*, into two strata; obtain μ_h and σ^2_h, for $h = 1, 2$.

- Obtain μ_x and σ^2_x.

1. Using proportional allocation, obtain σ^2_w, σ^2_b and thus, σ^2_x.
2. What is the gain from such stratification scheme as compared with no stratification; *i.e.*, obtain

$$\text{Var}(\bar{x}_{srs}) - \text{Var}(\bar{x}_{pr})?$$

3. Suppose that you are assigned a task of selecting a sample of employees from an agency that has $N = 3,000$ employees; $N_1 = 1,200$ belong to Division One, and $N_2 = 1,800$ belong to Division Two. The purpose is to estimate various parameters related to years of experience. The proposed sampling plan is to randomly select a sample of $n = 250$ employees.

- Explain how you would select the sample if you use *proportional allocation*;
- Explain how you would select the sample if you use *equal allocation*;
- Explain how you would select the sample if you use *optimal allocation*, where $\sigma_1 = 5$ and $\sigma_2 = 1$.

4. Suppose that you are assigned a task of selecting a sample of units of a product from a lot that has $N = 1,000$ units. The lot has $N1 = 300$ units identified from Shift One; the rest are from Shift Two. The purpose is to estimate the proportion p_d of defective units in the lot. The proposed sampling plan is to randomly select a sample of $n = 50$ units.

- Explain how you would select the sample if you use *proportional allocation*;
- Explain how you would select the sample if you use *equal allocation*;
- Explain how you would select the sample if you use *optimal allocation*, where $\sigma_{p_1} = 0.5$ and $\sigma_{p_2} = 0.09$.

Chapter 5

Introduction to Clustered Sampling

> It seems to me that the prime requirement for a teacher is to possess some knowledge to teach. He who does no research possesses no knowledge and has nothing to teach. Of course, some people that do good research are also good teachers. This is a fine combination, and one to be thankful for, but not to expect. Two of the poorest teachers that I ever had ... were Professor Ernest Brown in mathematics at Yale and Sir Ronald Fisher at University College in London. Sir Ernest will be known for centuries for his work on lunar theory, and Sir Ronald for revolutionizing man's method of inference. People came from over the world to listen to their impossible teaching, and to learn from them, and learn they did. I would not trade my good luck to have had these men as teachers for hundred of lectures by lesser men but "good teachers." ....[a]

[a] W Edwards Deming; 1972. Letter to the Editor, *The American Statistician*. Volume 26.

5.1 Introduction

In clustered sampling the units sampled are chosen from mutually exclusive groups, known as *clusters*, in which the units are generally close to each other. Examples are households in a block, or successive items off a production line. Sample selection is made hierarchically, using at least two stages: First, the sampling frame is divided into a large number (M) of clusters, and a sample of $m \leq M$ of these clusters are randomly selected. Then, within each of the m selected clusters, n sampling units are chosen by simple random sampling or some other sampling technique. Ideally, the chosen clusters should have dissimilar sampling units (*i.e.* within cluster heterogenuity) so that the sample

of clusters is as representative as possible of the whole sampling frame. In practice, however, this is not generally true, and that is why we tend to select a large number of clusters in the first stage, followed by a small number of sampling units in the second stage.

It is generally expensive to spread out our sample across the frame as a whole. For example, travel can become expensive if we are using interviewers or enumerators to travel all over a region or a country. To reduce costs we may choose a clustered sampling technique. It could also happen that we do not have an exhaustive list of the sampling units of interest, *e.g.*, people or families, but we do have a list of household blocks from a map. Then, necessarily, we have to select blocks in the first stage of sampling and then, select a random sample of households from the chosen blocks. Other examples of clusters may be farms, schools, hospitals, and other geographic areas.

We remark that only sampling units from randomly selected clusters are included in the sample. Thus, units from non-selected groups are represented by those from the selected clusters. Notice also, that clustered sampling differs from a stratified sampling design, where sampling units are selected from each group or stratum; thus, the selected sampling units represent all strata in the frame.

5.2 Some Advantages of Clustered Sampling

Clustered sampling offers several advantages, some of which are the following:

- Reduced costs by saving of travelling time by supervisors and enumerators;
- Administrative convenience and simplified field work;
- Useful in surveying employees in a particular industry or patients in hospitals;
- Instead of having a sample scattered over the entire coverage area, the sample is more localized in relatively few groups or clusters;
- Only a listing of sampling units in the selected clusters is needed and not of all units in the whole frame.

5.2.1 Some Disadvantages of Clustered Sampling

On the other hand, clustered sampling offers several disadvantages, some of which are the following:

- Generally, less accurate results are often obtained from clustered sampling, due to higher sampling error, than in simple random sampling;

- Sampling units close to each other may be very similar and thus, less likely to represent the whole frame; *e.g.*, units within the same cluster do not possess independent information (known as the "intra-class (cluster) correlation");
- Decrease in sampling efficiency with increase in cluster size, although the loss in sampling efficiency is often compensated by cost reduction.

5.3 Two-Stage Clustered Sampling

In order to illustrate two-stage clustered sampling estimation, we will make use of the next table[1]:

Table 5.1: Some Formulas in Two-Stage Clustered Sampling

	Frame	Sample
Primary Units (PUs)	M	m
Sampling Units (SUs) in PU_i	N_i	n_i
All Sampling Units	$N = \sum_{i=1}^{M} N_i$	$n = \sum_{i=1}^{m} n_i$
Mean SUs per PU	$\bar{N} = N/M$	$\bar{n} = n/m$
X-Value in PU_i of SU_j	$X_{ij};\ j = 1 : N_i$	$x_{ij};\ j = 1 : n_i$
Total of X-Value on PU_i	$\sum_{j=1}^{N_i} X_{ij}$	$\sum_{j=1}^{n_i} x_{ij}$
Total of X-Value on all PUs	$\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}$	$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}$
Mean per SU in PU_i	$\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$	$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$
Mean Per SU	$\mu_x = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i} X_{ij}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}$
Variance Between SUs	$\sigma_b^2 = \frac{1}{M\bar{N}}\sum_{i=1}^{M} N_i(\mu_i - \mu_x)^2$	$s_b^2 = \frac{1}{m\bar{n}}\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2$
Variance per SU within PU_i	$\sigma_i^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(X_{ij} - \mu_i)^2$	$s_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$
Mean Within SU Variance	$\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{M} N_i\,\sigma_i^2$	$s_w^2 = \frac{1}{n}\sum_{i=1}^{m} n_i\,s_i^2$
Total Variance	$\sigma_x^2 = \sigma_b^2 + \sigma_w^2$	$s_x^2 = s_b^2 + s_w^2$

5.3.1 Two-Stage Cluster Sample Using *R*

Suppose that we have a sampling frame of M PUs, and $\bar{N}$ SUs. Furthermore, suppose that we want to sample $m \leq M$ PUs in the first stage, and $\bar{n}$ SUs in the second stage.

A very simple R function that helps us to implement this clustered sampling procedure is the following:

[1] Cf. W Edwards Deming; 1950, 1986. *Some Theory of Sampling.* New York: Wiley, Dover, Pages 142-145.

```
# Cluster Sample Using R
cluster.sample=function(M=100, m=20, Nbar=10, nbar=5){
 cl.samp=sample(1:M, m)
 s.samp=NULL
 for (i in 1:m){
 su.sampl=sample(1:Nbar, nbar)
 s.samp=c(s.samp, su.sampl)
 }
samp=c(cl.samp,s.samp)
print("Your Clustered Sample is :", return(samp))
}
```

Then for example, to obtain our desired clustered sample using the R program, we type the command:
`cluster.sample(100, 20, 10, 5)`.

Notice that this function returns as output our sampled PUs, followed by the corresponding sample of SUs.

A perhaps better alternative R function is the following:

```
# Cluster Sample Using R
cluster.sample=function(M=100, m=20, Nbar=10, nbar=5){
 s.samp=NULL
 for (i in 1:m){
 cl.sampl=sample(1:M, 1)
 su.sampl=sample(1:Nbar, nbar)
 s.samp=c(s.samp,cl.sampl, sort(su.sampl))
 }
samp=s.samp
print("Your Clustered Sample is :", return(samp))
}
```

This function returns as output each sampled PU, followed by a corresponding sorted sample of SUs, which is easier to read.

5.3.2 Examples of Two-Stage Clustered Sampling

As illustration of two-stage clustered sampling, we present the following examples, as obtained from the internet.

- Clustered sampling in an education study[2]— A good example where the country was first stratified by region and area; then, a sample of schools was selected in stage one, and a sample of school sections was selected for testing in stage two.

[2]E Puhakka; 1999. *Application of Two Stage Cluster Sampling in Finnish Data of IEA Civic Education Study*

- An assessment of an immunization coverage[3]— This is known as the 30×7 WHO Sample. Here, a random sample of "sites" (clusters) are selected from geographical areas of a country in stage one. Then, a selection of seven individuals from the appropriate age class within each selected site are non-randomly chosen in stage two.
- Compact segment sampling, to estimate vaccination coverage[4] — In order to correct the selection bias induced by the second stage of the 30×7 WHO Sample, the clusters are divided into households segments and a sample of segments are randomly chosen and investigated completely.

5.4 Two-Stage Cluster Sampling Estimation

Assume that we have a frame which is made of M mutually exclusive clusters, and that we want to randomly select a sample of m of them using *srswor* in a first stage. Then, in a second stage we randomly select a sample of $\bar{n} =$ out of $\bar{N} =$ sampling units, also using *srswor* from the selected clusters in the first stage.

Let x_{ij} be the X-value of the measure of interest from the j^{th} sampling unit from the i^{th} selected cluster. And let $\bar{x}_i$ be the sample mean from the i^{th} selected cluster (see Table 1). Then, an unbiased estimator of μ_x obtained by cluster sampling, is given by

$$\hat{\mu}_{clu} = \frac{1}{m\bar{n}} \sum_{i=1}^{m} \sum_{j=1}^{\bar{n}} x_{ij}, \tag{5.1}$$

The sampling variance of $\hat{\mu}_{clu}$ is then given by

$$\text{Var}(\hat{\mu}_{clu}) \approx (1 - f_1)\frac{\sigma_b^2}{m} + (1 - f_2)\frac{\sigma_w^2}{m\bar{n}} \tag{5.2}$$

where $f_1 = m/M$, for stage one sampling, $f_2 = \bar{n}/\bar{N}$, for stage two sampling, and $\sigma_x^2 = \sigma_b^2 + \sigma_w^2$.

5.4.1 Frame Total Estimation

Assume that we have a sample of m randomly selected clusters and within each selected cluster, we select $\bar{n}$ sampling units. Furthermore, we want to estimate the total of X in the frame, τ_x. Then an estimator of this total will be given by

$$\hat{\tau}_x = \frac{M}{m} \sum_{i=1}^{m} \frac{\bar{N}}{\bar{n}} \sum_{j=1}^{\bar{n}} x_{ij}, \tag{5.3}$$

[3]RH Henderson & T Sundaresan; 1982. Cluster sampling to assess immunization coverage: review of experience with a simplified sampling method. *Bulletin of the World Health Organization*. Vol. 6, No.2: 353–260.

[4]P Milligan, A Njie and S Bennett; 2004. Comparison of two cluster sampling methods for health surveys in developing countries. *International Journal of Epidemiology*. Vol. 33:18.

whose sampling variance is given by

$$\text{Var}(\hat{\tau}_x) = \frac{M^2}{m}(1-f_1)\sigma_b^2 + \frac{M}{m}\sum_{i=1}^{M}\frac{\bar{N}^2}{\bar{n}}(1-f_2)\sigma_w^2. \tag{5.4}$$

An estimator of $\text{Var}(\hat{\tau}_x)$ is obtained substituting σ_b^2 by s_b^2, and σ_w^2 by s_w^2, as given in Table 5.1, above.

5.5 Clustered Sampling Allocation

We now introduce a simplified cost function, where we assume an overhead cost c_0, a cost of including an additional PSU in stage one, given as c_1, and the cost per sampling unit examine in stage two, given as c_2; *i.e.*, the total cost is given as:

$$\begin{aligned} c &= c_0 + c_1 m + c_2 n \\ &= c_0 + c_1 m + c_2 m\bar{n}, \end{aligned} \tag{5.5}$$

where we assume that $n = m\bar{n}$. The sampling variance of $\hat{\mu}_{clu}$ was then given in section 5.4 as

$$\text{Var}(\hat{\mu}_{clu}) \approx (1-f_1)\frac{\sigma_b^2}{m} + (1-f_2)\frac{\sigma_w^2}{m\bar{n}}. \tag{5.6}$$

Suppose that we want find the values of m an $\bar{n}$ that minimize $\text{Var}(\hat{\mu}_{clu})$. From sampling theory, assuming that $\bar{n} << \bar{N}$, using calculus it can be shown that

$$\bar{n} = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_1}{c_2}} \tag{5.7}$$

For example, suppose that from a previous study, we know that $\sigma_w = 1.5$ and $\sigma_b = 0.20$. Also, suppose that $c_1 = 80$ and $c_2 = 30$

$$\bar{n} = \frac{1.5}{0.2}\sqrt{\frac{80}{30}} \approx 12; \tag{5.8}$$

i.e., 12 sampling units will examined in stage two.

Furthermore, suppose that we have \$30,000.00 to carry out our study. Then,

$$\begin{aligned} c_1\, m + c_2\, m\, \bar{n} &= \$30,000 \\ 80\, m + 30\, m\, (12) &= \$30,000 \end{aligned} \tag{5.9}$$

Then, $440\, m = 30,000$ or $m \approx 68$ clusters will be taken in stage one, and thus, $n = 68 \times 12 = 816$.

5.5.1 Clustered Sampling Allocation With *R*

Suppose that we want to obtain the number of clusters, m, in the first stage and the average number of sampling units, $\bar{n}$, in the second stage, assuming certain values for: σ_w, σ_b, marginal cost in stage one, c_1, and c_1, the marginal cost of stage two; furthermore, we assume a field study budget, B.

A very simple R program is given below:

```
sample.alloc = function(sigma.w=1.0, sigma.b=0.1, c1=160, c2=40, B=48000){
 n.bar = round(sigma.w/sigma.b*sqrt(c1/c2),0)
 m = round(B/(c1+c2*n.bar),0)
 n = round(m*n.bar,0)
 clusters = data.frame(m, n.bar, n)
 print("Clustered allocation is:", return(clusters))
}
```

Then, for example type the command:

```
sample.alloc(1.0, 0.1, 160, 40, 48000)
```

to obtain the desired result.

5.6 PPS Clustered Sampling

In this type of clustered sampling, primary sampling units (clusters) are selected according to their differing number of sampling units; *i.e.*, with probability proportionate to size (PPS). Then, necessarily clusters are sampled with replacement, which implies that a cluster can appear more than once in the list of selected clusters.

Let m be the number of clusters randomly selected with PPS in stage one. Then, within each selected cluster, we randomly select n_i sampling units out of the N_i, with probability

$$\pi_i = \frac{N_i}{\sum_{i=1}^{M} N_i} = \frac{N_i}{N}, \quad \text{for } i = 1 : M. \tag{5.10}$$

Now, the *Hansen-Hurwitz estimator*, HH, for the total τ_x is given by[5]

$$\hat{\tau}_{HH} = \frac{M}{m} \sum_{i=1}^{m} \frac{N}{N_i} \sum_{j=1}^{n_i} x_{ij} \tag{5.11}$$

Then, an HH estimator of μ_x is given by

$$\hat{\mu}_{HH} = \frac{1}{N} \frac{M}{m} \sum_{i=1}^{m} \frac{N}{N_i} \sum_{j=1}^{n_i} x_{ij}/n_i \tag{5.12}$$

[5]Hansen MM and WN Hurwitz; 1943. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*. Vol. 14, pages 333–362.

And an estimator of $\text{Var}(\hat{\tau}_{HH})$ is

$$\widehat{\text{Var}}(\hat{\tau}_{HH}) = \frac{N^2}{m(m-1)} \sum_{i=1}^{m} (\bar{x}_i - \hat{\mu}_{HH})^2 \tag{5.13}$$

An alternative PPS estimator for the total τ_x, known as the *Horvitz-Thompson estimator*, HT, is obtained by[6]

$$\hat{\tau}_{HT} = \sum_{i=1}^{\nu} \frac{x_i}{\pi_i} \tag{5.14}$$

where ν is the number of distinct primary units in the sample, known also as the *effective sample size*.

The variance of this estimator is much elaborated; for more detail, see for example, the book titled *Sampling, 2nd Edition* by SK Thompson, Section 6.2, pages 53–56.[7]

5.6.1 Sample Selection Procedure

Suppose that we have M clusters and we want to select m of them with PPS. Assume that the the clustered frame has been arraigned as it appears in Table 5.2, below.

Table 5.2: Frame of Primary Sampling Units: PPS Selection

PSU	N_i	$\sum N_i$
1	N_1	N_1
2	N_2	$\sum_{i=1}^{2} N_i$
3	N_3	$\sum_{i=1}^{3} N_i$
$\vdots$	$\vdots$	$\vdots$
$M-1$	N_{M-1}	$\sum_{i=1}^{M-1} N_i$
M	N_M	N

Then, we select m random sampling numbers (r_i, for $i = 1 : m$), with replacement between 1 and M. The cluster selection is then made using the procedure that appears in Table 5.3.

We notice that since since the selection is based on PPS, a cluster can be selected more than once; *i.e.*, the selection of clusters is with replacement. In practice, we usually make a systematic selection of clusters in the first stage; a

[6] Horvitz DG and DJ Thompson; 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, Vol. 47, pages 663–685.

[7] See also, Särndal CE, B Sweenson, and J Wretman; 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Table 5.3: Cluster Selection Guide

Random No. Is	Select Cluster No.
$r_i \leq N_1$	1
$N_1 < r_i \leq \sum_{i=1}^{2} N_i$	2
$\sum_{i=1}^{2} N_i < r_i \leq \sum_{i=1}^{3} N_i$	3
$\vdots$	$\vdots$
$\sum_{i=1}^{M-1} N_i < r_i \leq N$	M

better procedure is to make replicated cluster selections, with different random starts.

5.6.2 A PPS Sample Selection Procedure Example

Suppose that a region is made of 15 municipalities where there are N_i families; for $(i = 1 : 15)$. We want to select 10 municipalities, based on PPS, and 20 families from each selected municipality (*i.e.*, a sample of $n = 200$) families. Thus, assume that the region is distributed as in Table 5.4, bellow.

Table 5.4: Frame of Families Per Municipality: PPS Selection

$Munic.$	N_i	$\sum N_i$	$Munic.$	N_i	$\sum N_i$	$Munic.$	N_i	$\sum N_i$
1	800	800	6	600	6400	11	600	11600
2	1200	2000	7	500	6900	12	800	12400
3	400	2400	8	1200	8100	13	400	12800
4	2000	4400	9	500	8600	14	900	13700
5	1400	5800	10	2400	11000	15	1300	15000

5.6.3 A PPS Sample Selection Using *R*

Suppose that we want to randomly select a sample of m clusters by the PPS procedure. Using program R, we create the following simple function:

```
sample.pps=function(N=1:15000, m=10)){
 samp=sort(sample(N, m, replace=TRUE))
 print("The m selected clusters are:  ", return(samp))
}
```

Suppose that we type the command:

```
sample.pps(1:15000, 10)
```

For example, if we obtain the following result:
2071 2761 5243 6647 7603 7717 8495 12449 13555 13930

Then, we select municipalities numbered: $3, 4, 5, 7, 8, 8, 9, 13, 14, 15$. Thus, in our sample, municipality number eight (8) appears twice, so we would select 40 families from it, and 20 families from the each of the other eight selected municipalities.

Exercises

1. Suppose that a book of $N = 500$ pages will be examined in order to estimate the total number of errors. For our purpose, we will randomly select a sample of $n = 30$ pages. The book is divided into $M = 100$ consecutive clusters, each of them has $\bar{N} = 5$ pages.

- Explain how you would select the sample if you use *clustered sampling* and $m = 6$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 15$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 30$.

2. Suppose that you are assigned a task of selecting a sample of employees from an agency that has $N = 3,000$ employees. The purpose is to estimate various parameters related to years of experience. The proposed sampling plan is to randomly select a sample of $n = 250$ employees. The frame is divided into $M = 300$ consecutive clusters, each of them has $\bar{N} = 10$ employees.

- Explain how you would select the sample if you use *clustered sampling* and $m = 25$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 50$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 125$.

3. Suppose that you are assigned a task of selecting a sample of units of a product from a lot that has $N = 1,000$ units. The purpose is to estimate the proportion p_d of defective units in the lot. The proposed sampling plan is to randomly select a sample of $n = 50$ units. The lot is divided into $M = 200$ consecutive clusters, each of them has $\bar{N} = 50$ patients.

- Explain how you would select the sample if you use *clustered sampling* and $m = 10$;

- Explain how you would select the sample if you use *clustered sampling* and $m = 25$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 50$.

4. Suppose that you are assigned a task of selecting a sample of list of patients from a hospital that has $N = 5,000$ patients. The purpose is to estimate health cost per patient. The proposed sampling plan is to randomly select a sample of $n = 150$ units. The list is divided into $M = 500$ consecutive clusters, each of them has $\bar{N} = 10$ units.

- Explain how you would select the sample if you use *clustered sampling* and $m = 25$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 75$;
- Explain how you would select the sample if you use *clustered sampling* and $m = 150$.

5. Suppose that we will perform a study of patients from certain disease A in hospitals in a country. Assume that we will visit m hospitals to study $\bar{n}$ hospital records in order to estimate the proportion of patients with condition A. Furthermore, we assume that the average cost $c_1 = \$160$ to bring an additional hospital into the sample, which include visiting the hospital's director to describe the purpose of the study, coordinate with the supervisor who will assist the enumerators, and to study the hospital records. The marginal cost $c_2 = \$40$ will be mainly that of studying hospital records to decide whether the patient is a case of condition A and to carry the fieldwork. From a similar study, we take $\sigma_w = 1.0$ and $\sigma_b = 0.10$.

- Calculate the corresponding number of records per hospital, $\bar{n}$;
- If the budget for this study is $\$48,000.00$, what will be the needed number of hospitals, m?
- What will be the total number of records, n?

6. Using the data from the example of municipalities above, select ten of them by PPS:

- Using linear systematic sampling;
- Using circular systematic sampling;
- Using replicated sampling with two sub-samples;

Chapter 6

Introduction to Ratio and Regression Estimation

> *Limitations of Statistical Inference.* All results are conditional on (a) the frame whence came units for test; (b) the method of investigation (the questionnaire or test-method and how it was used); (c) the people that carry out the interviews or measurements. In addition (d), the results of an analytic study are conditional on certain environmental states ... The exact environmental conditions for any experiment will never be seen again ... The gap beyond statistical inference can be filled only by knowledge of the subject-matter (economics, medicine, chemistry, engineering, psychology, agricultural science, etc.), which may take the formality of a model.[a]

[a] W Edwards Deming; November, 1975. On Probability As a Basis For Action, *The American Statistician.* Volume 29, No. 4, pp. 146–152.

The method of *ratio estimation* is a technique that uses available auxiliary information which is correlated with the variable of interest. Suppose that variable X is correlated with variable of interest Y; furthermore, that we have a paired random sample of n observations (x_i, y_i) for $i = 1 : n$. Then, a *ratio estimator* of

$$R \equiv \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}, \tag{6.1}$$

is obtained as

$$r \equiv \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}. \tag{6.2}$$

We notice that an important characteristic of ratio estimation is that both, the numerator and the denominator are random quantities.

The sampling variance of r is given by

$$\mathrm{Var}(r) \approx \frac{1}{\mu_x^2}(1-f)\frac{\sum_{i=1}^{N}(y_i - R\,x_i)^2}{nN}, \tag{6.3}$$

which is estimated as

$$\widehat{\mathrm{Var}}(r) \approx \frac{1}{\bar{x}^2}(1-f)\frac{\sum_{i=1}^{n}(y_i - r\,x_i)^2}{n(n-1)}. \tag{6.4}$$

Some example of ratio estimates are given in Table 6.1, below.

Suppose that we are interested in estimating the total, τ_y, of characteristic Y in the frame and that we know the total, τ_x, of characteristic X. Then, a ratio estimator of τ_y is given by

$$\hat{\tau}_y = r\,\tau_x \tag{6.5}$$

From sampling theory, we know that the above ratio estimator is biased, and that its bias, $\mathrm{B}(r)$, is given by

$$\begin{aligned} \mathrm{B}(r) &= R\left(\frac{\mathrm{Var}(\hat{\tau}_x)}{\tau_x^2} - \frac{\mathrm{Cov}(\hat{\tau}_x, \hat{\tau}_y)}{\tau_x\,\tau_y}\right) \\ &= \frac{1}{\tau_x^2}(\mathrm{RV}(\hat{\tau}_x) - \mathrm{Cov}(\hat{\tau}_x, \hat{\tau}_y)) \end{aligned} \tag{6.6}$$

6.1 Ratio Estimation in SRS

From sampling theory, we also know that the sampling variance of $\hat{\tau}_y$ is given by

$$\mathrm{Var}(\hat{\tau}_y) = \mathrm{Var}(\hat{\tau}_{y_{srs}}) - 2\,R\,\mathrm{Cov}(\hat{\tau}_x, \hat{\tau}_y) + R^2\,\mathrm{Var}(\hat{\tau}_x). \tag{6.7}$$

Then, $\hat{\tau}_y$ is more efficient than $\hat{\tau}_{y_{srs}}$ when $2\,R\,\mathrm{Cov}(\hat{\tau}_x, \hat{\tau}_y) > R^2\,\mathrm{Var}(\hat{\tau}_x)$.

As we pointed out above, under SRS, an estimator of $\mathrm{Var}(r)$ is given by

$$\widehat{\mathrm{Var}}(r) = \frac{1}{\bar{x}^2}\frac{\sum_{i=1}^{n}(y_i - r\,x_i)^2}{n(n-1)} \tag{6.8}$$

Thus, the 95% limits for τ_y are obtained by

$$\hat{\tau}_y \pm t_\nu \tau_x se(r), \tag{6.9}$$

where $se(r)$ is the square root of $\widehat{\mathrm{Var}}(r)$.

6.1.1 Ratio Estimation Example

We want to estimate the proportion non-conforming p_y of certain item, which comes in a ship of $N = 1,000$ lots. We randomly select $n = 10$ lots, and classify them by the number of units examined (X) and the number of non-conforming units (Y). Table 6.2, below, presents the results of our sample.

Using program R, we typed the following commands, with the corresponding results:

Table 6.1: Ratio Estimate Examples

X & Y Value	Ratio ($R = \frac{\tau_y}{\tau_x} = \frac{\mu_y}{\mu_x}$)
$X =$ Family Size $Y =$ Food Consumption	$R =$ Food Consumption per Capita
$X =$ Number of dairy farms $Y =$ Milk Production	$R =$ Milk Production per Farm
$X =$ Number of Hospitals $Y =$ Disease A Cases	$R =$ Disease A Cases per Hospital
$X =$ Number of Children $Y =$ Number Immunized	$R =$ Proportion Immunized Children
$X =$ Number of Families $Y =$ Number Under Puberty Level	$R =$ Proportion Under Puberty Level
$X =$ Labor Force Size $Y =$ Number Unemployed	$R =$ Unemployment Rate
$X =$ Cell Phones : 2000 $Y =$ Cell Phones : 2005	$R =$ Increase Rate
$X =$ Acreage Planted $Y =$ Production (tons)	$R =$ Yield Per Acre
$X =$ Number of Drinks $Y =$ Blood Alcohol Content	$R =$ BloodAlcoholContent Per Drink
$X =$ Speed (MPH) $Y =$ Distance Traveled (Miles)	$R =$ Distance Per Speed
$X =$ Man $-$ hours $Y =$ Number of Items Processed	$R =$ Productivity Rate

Table 6.2: Item Classification By Lot

Lot	x_i	y_i	Lot	x_i	y_i
1	20	4	6	20	4
2	20	2	7	40	7
3	30	5	8	30	5
4	25	4	9	10	1
5	15	2	10	20	3

```
x = scan()
20 20 30 25 15 20 40 30 10 20
y = scan()
4 2 5 4 2 4 7 5 1 3
plot(x,y)
n = length(x)
r = sum(y)/sum(x)
r
0.1608696
var.r = 1/mean(x)^2*sum(y-r*x)^2/(n*(n-1))
var.r
4.566893e-34
se.r = sqrt(var.r)
se.r
2.137029e-17
```

This means that our estimate of the proportion non-conforming is 16.1%. Then, the 95% limits for the proportion non-conforming in the shipment are obtained as:

$$r \pm t_\nu se(r), \tag{6.10}$$

Thus, in our example, we obtain the following, using program R:

```
r - qt(0.975, n-1)*se.r
0.1608696
r + qt(0.975, n-1)*se.r
0.1608696,
```

which, apparently in this case, do not change from the estimate r, because the margin of sampling error is very small.

N.B. In the *statistical quality control* literature, the corresponding estimate for the proportion non-conforming is usually expressed as $\bar{p} = \sum y/\bar{n}$, where $\bar{n}$ is the average sample size ($\bar{x}$ in our case). And the corresponding (three sigma) limits are given by

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}. \tag{6.11}$$

We notice that these estimators have the limitation that the sample size (x in our case) is not fixed, but a random variable.

6.2 Ratio Estimation in Stratified Sampling

Under a stratified sampling scheme, where we have H strata, the method of ratio estimation can be used in two ways, known as *combined* and *separate* ratio estimation.

6.2.1 Combined Ratio Estimation

Suppose that $\hat{\mu}_{x_{str}}$ and $\hat{\mu}_{y_{str}}$ are the estimators of μ_x and μ_y, respectively, under stratified sampling. Here, as the term indicates, we obtain a ratio estimator given by

$$r_{c.str} = \frac{\hat{\tau}_{y_{str}}}{\hat{\tau}_{x_{str}}}, \tag{6.12}$$

where

$$\hat{\tau}_{y_{str}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}, \tag{6.13}$$

and

$$\hat{\tau}_{x_{str}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hi}. \tag{6.14}$$

The sampling variance of $r_{c.str}$ is given by

$$\mathrm{Var}(r_{c.str}) \approx \frac{1}{\tau_x^2} \sum_{h=1}^{H} (1 - f_h) \frac{N_h^2}{n_h} \sigma_h^2. \tag{6.15}$$

Now,

$$\sigma_h^2 = \sigma_{y_h}^2 + R\,\sigma_{x_h}^2 - 2\,R\,\mathrm{Cov}(x_h, y_h), \tag{6.16}$$

where σ_{x_h} and σ_{y_h} are the corresponding within stratum standard deviations of X and Y.

An estimator of $\mathrm{Var}(r_{c.str})$ is given by

$$\widehat{\mathrm{Var}}(r_{c.str}) \approx \frac{1}{\hat{\tau}_x^2} \sum_{h=1}^{H} (1 - f_h) \frac{N_h^2}{n_h} s_h^2. \tag{6.17}$$

where these estimators are obtained by the corresponding sample values; *e.g.*,

$$s_{x_h}^2 = \frac{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}{n_h - 1}, \quad \text{for } h = 1 : H. \tag{6.18}$$

6.2.2 Separate Ratio Estimation

Assume that we know the total of variable X for each stratum, τ_{x_h}, and thus the total of variable X in the frame, $\tau_x = \sum_{h=1}^{H} \tau_{x_h}$. Also, suppose that for each stratum h, we obtain a ratio estimator

$$r_h = \frac{\sum_{i=1}^{n_h} y_{ih}}{\sum_{i=1}^{n_h} x_{ih}}, \quad \text{for } h = 1 : H. \tag{6.19}$$

Then, the total of variable Y in stratum h is estimated as

$$\hat{\tau}_{y_h} = r_h \, \tau_{x_h}, \quad \text{for } h = 1 : H. \tag{6.20}$$

Then an estimate of R in the frame is obtained by

$$r_{s.str} = \frac{1}{\tau_x} \sum_{h=1}^{H} r_h \, \tau_{x_h}, \tag{6.21}$$

whose sampling variance is given by

$$\text{Var}(r_{s.str}) \approx \frac{1}{\tau_x^2} \sum_{h=1}^{H} (1 - f_h) \tau_{x_h}^2 \text{Var}(\, r_h) \tag{6.22}$$

where $\tau_{x_h}^2 \text{Var}(\, r_h)$ is the within stratum h variance of the ratio estimated total $\hat{\tau}_{y_h}$. An estimator of $\text{Var}(r_h)$ is given by

$$\widehat{\text{Var}}(r_h) = \frac{1}{\bar{x}_h^2} \frac{\sum_{i=1}^{n_h} (y_{hi} - r_h \, x_{hi})^2}{n_h(n_h - 1)} \tag{6.23}$$

Therefore, an estimate of the total of variable Y in the frame, τ_y, is

$$\hat{\tau}_{y_{s.str}} = \sum_{h=1}^{H} \hat{\tau}_{y_h} = \sum_{h=1}^{H} r_h \, \tau_{x_h}. \tag{6.24}$$

whose estimated variance is

$$\widehat{\text{Var}}(\hat{\tau}_{s.str}) \approx \sum_{h=1}^{H} (1 - f_h) \tau_{x_h}^2 \, \widehat{\text{Var}}(\, r_h) \tag{6.25}$$

6.3 Introduction to Regression Estimation

When the auxiliary variable X is a predetermined (non-random) variable, we can obtain an alternative estimator to the ratio estimator. It is based on the concept of *least squared* method and it is known as *regression estimation.*

6.3.1 Regression Trough The Origin, *RTO*

Assuming that the relation between X and Y is

$$y_i = \beta x_i + \epsilon_i, \text{ for } i = 1 : n, \tag{6.26}$$

for the paired observations (x_i, y_i), the regression estimator of β is given as[1]

$$b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \left(\sum_{i=1}^n x_i^2\right)^{-1} \sum_{i=1}^n x_i y_i. \tag{6.27}$$

The variance of Y, σ_y^2, is known as the *mean square error* (MSE), and is estimated as

$$\text{MSE} \equiv \hat{\sigma}_y^2 = \frac{\sum_{i=1}^n (y_i - b\, x_i)^2}{n-1} \tag{6.28}$$

An estimator of the variance of b is then given by[2]

$$\widehat{\text{Var}}(b) = \frac{\text{MSE}}{\sum_{i=1}^n x_i^2} = \hat{\sigma}_y^2 \left(\sum_{i=1}^n x_i^2\right)^{-1}. \tag{6.29}$$

Therefore, using *RTO* estimation, an estimator for the mean of Y in the frame is obtained as

$$\hat{\mu}_{y_{reg}} = b\, \mu_x. \tag{6.30}$$

The confidence limits for $\mu_{y_{reg}}$ are obtained by

$$\hat{\mu}_{y_{reg}} \pm t_\nu \sqrt{\widehat{\text{Var}}(b)}\ \mu_x, \tag{6.31}$$

here, the degrees of freedom are $\nu = n - 1$. Similarly, an estimator for the total of Y in the frame is obtained as

$$\hat{\tau}_{y_{reg}} = b\, \tau_x \tag{6.32}$$

An estimator of the variance of $\hat{\tau}_{y_{reg}}$ is

$$\widehat{\text{Var}}(\hat{\tau}_{y_{reg}}) = \tau_x^2 \frac{\sum_{i=1}^n (y_i - b\, x_i)^2}{n(n-1) \sum_{i=1}^n x_i^2}. \tag{6.33}$$

The confidence limits for $\tau_{y_{reg}}$ are obtained by

$$\hat{\tau}_{y_{reg}} \pm t_\nu \sqrt{\widehat{\text{Var}}(b)}\ \tau_x. \tag{6.34}$$

And, in case that τ_x is unknown, we use $\hat{\tau}_x$ as its estimator in the above formulae.

[1] Notice the similarity between this expression and the *multiple regression* estimator, in matrix notation, $\hat{\beta} = (X^t X)^{-1} X^t y$.

[2] Cf. In multiple regression $\widehat{\text{Var}}(\hat{\beta}) = \hat{\sigma}_y^2 (X^t X)^{-1}$.

6.3.2 Simple Regression

More often, we study relationships between variables X and Y using the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ for } i = 1 : n; \tag{6.35}$$

i.e., where their relationship not necessarily goes through the origin. In such cases, we need to estimate both β_0 and β_1, using the least square method or equivalently, maximum likelihood estimation.

Regression theory shows that the estimator for β_1 is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{6.36}$$

Now, the variance of b_1 is estimated as

$$\widehat{\text{Var}}(b_1) = \frac{\text{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.37}$$

where, now

$$\text{MSE} \equiv \hat{\sigma}_y^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1\, x_i))^2}{n-2} \tag{6.38}$$

The estimator for β_0 is

$$b_0 = \bar{y} + b_1\, \bar{x} \tag{6.39}$$

Then, it can be shown that the mean of variable Y in the frame estimated by linear regression is

$$\hat{\mu}_{y_{reg}} = \bar{y} + b_1(\mu_x - \bar{x}); \tag{6.40}$$

an estimator of the variance of $\hat{\mu}_{y_{reg}}$ is given by

$$\widehat{\text{Var}}(\hat{\mu}_{y_{reg}}) = (1-f)\frac{\sum_{i=1}^{n}(y_i - b_0 - b_1\, x_i)^2}{n(n-2)}. \tag{6.41}$$

Therefore, the confidence limits for $\mu_{y_{reg}}$ are obtained as

$$\hat{\mu}_{y_{reg}} \pm t_\nu \sqrt{\widehat{\text{Var}}(\hat{\mu}_{y_{reg}})}, \tag{6.42}$$

now, the degrees of freedom are $\nu = n - 2$.

An estimator of the total for variable Y in the frame is obtained as

$$\hat{\tau}_{y_{reg}} = N\left[\bar{y} + b_1(\mu_x - \bar{x})\right]; \tag{6.43}$$

an estimator of the variance of $\hat{\mu}_{y_{reg}}$ is given by

$$\widehat{\text{Var}}(\hat{\tau}_{y_{reg}}) = N^2(1-f)\frac{\sum_{i=1}^{n}(y_i - b_0 - b_1\, x_i)^2}{n(n-2)}. \tag{6.44}$$

Therefore, the confidence limits for $\mu_{y_{reg}}$ are obtained as

$$\hat{\tau}_{y_{reg}} \pm t_\nu \sqrt{\widehat{\text{Var}}(\hat{\tau}_{y_{reg}})}, \tag{6.45}$$

6.3.3 Regression Estimation Using *R*

As general guidelines we recommend the following procedure:

- Perform *RTO* if there are theoretical reasons for regression through the origin;
- Otherwise, perform linear regression and verify if b_0 is statistically different from zero;
- Then, if b_0 is not statistically different from zero, perform *RTO*;
- Otherwise, perform linear regression (model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$).

Estimation of the regression line using program *R* is performed with the linear model function, `lm( )`. If we want to perform linear regression (model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$) we can type the following commands:

```
reg = lm(y ~ x)
summary(reg)
```

If we want to perform *RTO*, we can obtain an estimate of β and related statistics using program *R*, by the following commands:

```
reg = lm(y ~ -1 + x)
summary(reg)
```

Notice that we have to put -1 after the tilde, $\sim$, to instruct *R* that we are interested in performing *RTO*.

Exercises

1. We want to estimate the total number of certain cattle in certain region in 2008. The total number of cattle in 2007 was $\tau_x = 5,000$. The sampling unit was a farm in the region; assume that the number of farms in the frame is $N = 400$. A sample of $n = 20$ farms was randomly selected from the frame, and the following results were obtained:

- Obtain an X, Y plot the results for the two years. Indicate your observations;
- Obtain the sample correlation coefficient between X and Y;
- Obtain a ratio estimate of $R = \sum Y / \sum X$, where X represents the number of cattle in 2007, and Y represents the number of cattle in 2008;
- Estimate the total number of this type of cattle in the region, $\hat{\tau}_y$ in 2008;
- Obtain an estimate of $\text{Var}(r)$
- Estimate the 95 % limits for τ_y in 2008.

Table 6.3: Number of Certain Cattle in 2007 & 2008

Farm No.	2007	2008	Farm No.	2007	2008
1	10	15	11	9	13
2	25	30	12	15	20
3	35	40	13	11	11
4	10	13	14	20	25
5	10	13	15	15	15
6	30	25	16	25	30
7	15	17	17	9	11
8	20	22	18	15	14
9	9	11	19	12	15
10	25	30	20	20	22

Table 6.4: Supermarket Sales, By Type

Supermarkets	N_h	March Sales (X_h)	April Sales (Y_h)
Small	10	$3, 5, 4, 4, 3, 2, 5, 1, 1, 4$	$5, 7, 4, 5, 5, 3, 5, 2, 2, 5$
Medium	7	$12, 9, 7, 10, 8, 7, 9$	$10, 9, 7, 7, 5, 6, 5$
Large	3	$20, 15, 17$	$18, 12, 15$

2. The following table shows the supermarket sales ($000) in a city for two months:

- Obtain the frame ratio, R, of March-to-April sales;
- Randomly select a sample, where $n_1 = 3$, $n_2 = 2$, and $n_3 = 1$;
- Using the selected sample obtain a combined ratio estimate $r_{c.str}$;
- Obtain the variance $\text{Var}(r_{c.str})$;
- Estimate the total τ_y for April;
- Estimate the 95% limits for τ_y for April.

3. Obtain for the problem:

- The ratio estimates for each stratum, and thus the estimator $r_{s.str}$;
- The variance $\widehat{\text{Var}}(r_h)$ for each stratum h.
- An estimate of the total $\hat{\tau}_{y_{s.str}}$ and of $\hat{\tau}_{y_{s.str}}$

4. Using the data in problem 1, above:

- Obtain a regression estimate of β, where X represents the number of cattle in 2007, and Y represents the number of cattle in 2008;
- Estimate the total number of this type of cattle in the region, $\hat{\tau}_y$ in 2008;
- Obtain an estimate of $\text{Var}(\hat{\tau}_y)$
- Estimate the 95 % limits for τ_y in 2008.

5. For the following data:

X	105	113	125	137	141	153	165	177	188	198
Y	45	63	86	118	112	169	201	237	263	268

- Plot the graph for (x_i, y_i);
- Perform ratio estimation;
- Perform linear regression and decide whether RTO is needed.

6. Suppose that we wan to estimate the yield of certain product in a field of $N = 100$ plots. We randomly selected $n = 4$ plots and the amount of yield, y_i of each sampled plot was measured. It is known that the yield of a plot is related with the amount of fertilizer, x_i, applied to the plot, which is known for each plot in the field. the data is:

X	50	100	150	200
Y	141	169	168	185

- Plot the graph for (x_i, y_i);
- Estimate the regression line and decide whether RTO is needed;
- Estimate the mean yield and its confidence limits;
- Estimate the total yield and its confidence limits.

Chapter 7

The *R* Survey Package

Some remarks on the accuracy of an adjustment. *A least squares adjustment of sampling results must be regarded as a systematic procedure for obtaining satisfaction on the conditions imposed, and at the same time effecting an improvement of the data in the sense of obtaining results of a smaller variance than the sample itself, under ideal conditions of sampling from a stable universe. It must not be supposed that any or all of the adjusted m_{ij} in any table are necessarily "closer to the truth" than the corresponding sampling frequencies n_{ij}, even under ideal conditions. As for standard errors of the adjusted results, they can easily be estimated for the ideal case by making use of the calculated chi-square. For predictive purposes, however (which can be regarded as the only possible use of a census by any method, sample or complete), it is far preferable, in fact necessary, to get some idea of errors of sampling by actual trial, such as by a comparison of the sampling results with the universe, as can often be arranged by means of controls. There is another aspect to the problem of error—even a 100 per cent count, even though strictly accurate, is not itself useful for prediction, except so far as we can assert on other grounds what secular changes are taking place.*[a]

[a] W Edwards Deming and Frederick F Stephan; 1940. On a least squares adjustment of a sampling frequency table when the expected marginal totals are unknown. *The Annals of Mathematical Statistics*, Vol. XI, No. 4, Page 444.

7.1 Introduction

The following notes introduce the *R* program `survey` package. This package performs several survey sampling analyses, including: summary statistics, maximum likelihood estimation for multistage, stratified, cluster-sampled, unequally weighted samples. It also performs analysis of more complex survey samples, and several graphics procedures. The `survey` package is design-based, *i.e.*, in

which survey statistics the population data are regarded as fixed and the randomness comes entirely from the sampling procedure.

Some other *R* packages on sampling include: `pps`, `sampling`, and `sampfling`. These packages also focus on design, in particular *PPS* sampling without replacement.

The `survey` package was developed by Professor Thomas Lumley of the University of Washington. Current *R* version is 3.14 (April 15, 2009); version 2.3 was published in the *Journal of Statistical Software*, Vol. 9, Issue 8, on April, 2004[1] [2]

Some of `survey` package features include:

- Describing survey designs: `svydesign()`
- Database-backed designs;
- Summary statistics: mean, total, quantiles, design effect;
- Tables of summary statistics, domain estimation;
- Contingency tables: `svychisq()`, `svyloglin()`;
- Graphics: histograms, hexbin scatterplots, smoothers;
- Regression modelling: `svyglm()`, `svyolr()`;
- Multiply-imputed data.

As indicated by Lumley, the `survey` package always uses formulas to specify variables in a survey data set. Function `svydesign()` constructs an object that specifies the strata, PSUs, sampling weights or probabilities, and finite population correction for a survey sample. The resulting objects can be used in the statistical functions whose names begin 'svy': `svymean()`, `svyvar()`, `svyquantile()`, `svytable()`, and others. Standard errors are computed using Taylor series linearization (where available).

7.2 Basic Estimation

As in previous chapters, individuals are randomly sampled with known probabilities, π_i, from a frame of size N to end up with a sample of size n. Let

$$I_i = \begin{cases} 1 & \text{if individual } i \text{ is sampled;} \\ 0 & \text{if individual } i \text{ is not sampled.} \end{cases}$$

The design-based inference problem is to estimate what any statistic of interest would be if data from the whole frame were available.

[1] Available in the URL address, http://www.jstatsoft.org/v09/i08.
[2] For further information, visit the URL address, http://faculty.washington.edu/tlumley/survey/.

For a population total this is easy: an unbiased estimator of

$$\tau_y = \sum_{i=1}^{N} y_i \tag{7.1}$$

is

$$\hat{\tau}_y = \sum_{i:I_i=1} \frac{1}{\pi_i} y_i \tag{7.2}$$

Standard errors follow from formulas for the variance of a sum; the main complication is that we need to know $\text{Cov}(I_i, I_j)$.

7.3 Describing Surveys To *R*

We will focus on an example from the `survey` package: "Stratified independent sample (without replacement) of California schools". A measure of interest is the 'Academic Performance Index', *API*, which is computedd for all California schools, based on standardized testing of students. Suppose that we have a sample stratified by level of school (elementary, middle, high), in the data frame `apistrat`. The variable `snum` identifies a school, `stype` is the level of school, `fpc` is the number of schools in the stratum, and `pw` is the sampling weights. The initial step is to define a survey design object containing the data and metadata. Using the `survey` package, within program *R*:

```
library(survey)
data(api)
dstrat = svydesign(id=~1,strata=~stype, weights=~pw, data=apistrat,
fpc=~fpc)
```

where:

- `stype` is a factor variable for elementary/middle/high school;
- `fpc` is a numeric variable giving the number of schools in each stratum (if omitted we assume sampling with replacement);
- `id=~1` specifies independent sampling;
- `apistrat` is the data frame with all the data;
- `pw` contains sampling weights ($1/\pi_i$).

These could be omitted since they can be computed from the population size. Notice that all the variables are in the `apistrat` data frame, and are specified as formulas.

We can then type the command:

```
dstrat
```

to obtain:

```
Stratified Independent Sampling design
svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat,
fpc = ~fpc)
```

And then, we type the command:
`summary(dstrat)`
to obtain:

```
Stratified Independent Sampling design
svydesign(id = ~1, strata = ~stype, weights = ~pw, data = apistrat,
fpc = ~fpc)
Probabilities:
Min.    1st Qu. Median  Mean    3rd Qu. Max.
0.02262 0.02262 0.03587 0.04014 0.05339 0.06623
Stratum Sizes:
             E  H  M
obs        100 50 50
design.PSU 100 50 50
actual.PSU 100 50 50
Population stratum sizes (PSUs):
   E    M    H
4421 1018  755
Data variables:
[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pcttest" "api00"
...
```

Now, we can estimate for example, the mean *API* performance score, the total enrollment across California, typing the command:
`svymean(~api00+I(api00-api99), dstrat)`
to obtain:

```
                   Mean     SE
api00           662.287 9.4089
I(api00 - api99) 32.893 2.0511
```

Then, typing the command:
`svytotal( enroll, dstrat)`
we obtain:

```
         Total     SE
enroll 3687178 114642
```

7.4 Clustered Sampling

Here, we consider sampling 15 school districts and taking all the schools in each of these districts. The data frame is `apiclus1`: variable `dnum` identifies a school

district, `fpc` is the number of schools in the stratum, and `pw` is the sampling weights. The initial step is to define a survey design object containing the data and metadata.

Cluster sample of school districts, using all schools within a district, using the `survey` package, within program *R*:

```
dclus1 = svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
```

Notice that,

- `dnum` is a numeric identifier for school district;
- No stratification, so no strata = argument.

Then we type the command:

```
summary(dclus1)
```

to obtain:

```
1 - level Cluster Sampling design
With (15) clusters.
svydesign(id = ~dnum, weights = ~pw, data = apiclus1, fpc = ~fpc)
Probabilities:
Min.    1st Qu. Median  Mean    3rd Qu. Max.
0.02954 0.02954 0.02954 0.02954 0.02954 0.02954
Population size (PSUs): 757
Data variables:
[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pcttest" "api00"
...
```

7.5 Two-Stage Sampling

Now, we consider a two-stage cluster-sampled design in which 40 school districts are sampled and then up to five schools from each district. Using the `survey` package, within program *R*:

```
dclus2 = svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
```

Notice that,

- `dnum` identifies school district;
- `snum` identifies school;
- `fpc1` is the number of school districts in frame;
- `fpc2` is the number of schools in the district;
- Weights are computed from `fpc1` and `fpc2`

Then, we type the command:

```
summary(dclus2)
```

to obtain:

```
2 - level Cluster Sampling design
With (40, 126) clusters.
svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
Probabilities:
Min.     1st Qu.  Median   Mean     3rd Qu.  Max.
0.003669 0.037740 0.052840 0.042390 0.052840 0.052840
Population size (PSUs): 757
Data variables:
[1] "cds" "stype" "name" "sname" "snum" "dname"
[7] "dnum" "cname" "cnum" "flag" "pcttest" "api00"
...
```

7.6 *PPS* Sampling

Probability-proportional-to-size (*PPS*) is a general term for unequal probabilty sampling, because the most important application is sampling clusters proportional to size. The *Horvitz-Thompson estimator*, *HT*, is computationally difficult ($n \times n$ matrices) and depends on the pairwise sampling probabilities, which depend on the sizes of all clusters in the frame, not just those in the sample. The `survey` package uses a simple approximation that reduces exactly to the *HT* estimator for multistage stratified sampling and is accurate, more generally in *PPS* sampling.

Notice that totals are positively correlated with cluster size, even when proportions are negatively correlated.

7.6.1 *PPS* Sampling Example

Using the `survey` package, within program *R*, we type the commands:

```
plot(Bush~Kerry,data=election,log="xy")
plot(I(Bush/votes)~I(Kerry/votes), data=election)
dpps = svydesign(id=~1, weights=~wt, fpc=~p, data=election_pps, pps="brewer")
dppswr = svydesign(id=~1, weights=~wt, data=election_pps)
svytotal(~Bush+Kerry+Nader, dpps)
```

to obtain:

```
      Total    SE
Bush  64518472 2447629
Kerry 51202102 2450787
Nader 478530   102420
```

Then, we type the command:
`svytotal(~Bush+Kerry+Nader, dppswr)`
to obtain:

```
      Total    SE
Bush  64518472 2671455
Kerry 51202102 2679433
Nader 478530   105303
```

And the command:
`colSums(election[,4:6])`
to obtain:

```
Bush     Kerry    Nader
59645156 56149771 404178
```

7.7 Summary Statistics

The `survey` package functions: `svymean()`, `svytotal()`, `svyratio()`, `svyvar()`, `svyquantile()` are used for pertinent summary statistics. All take a formula and design object as arguments, return an object with `coef`, `vcov`, `SE`, `cv` methods.

Mean and total on factor variables give tables of cell means/totals. Mean and total have `deff` argument for design effects and the returned object has a `deff` method. Now, we type the command:
`svymean(~api00, dclus1, deff=TRUE)`
to obtain:

```
      Mean    SE     DEff
api00 644.169 23.542 9.3459
```

Then, we type the command:
`svymean(~factor(stype),dclus1)`
to obtain:

```
                Mean     SE
factor(stype)E 0.786885 0.0463
factor(stype)H 0.076503 0.0268
factor(stype)M 0.136612 0.0296
```

And, we type the command:
`svymean(~interaction(stype, comp.imp), dclus1)`
to obtain:

```
                                    mean     SE
interaction(stype, comp.imp)E.No  0.174863 0.0260
interaction(stype, comp.imp)H.No  0.038251 0.0161
interaction(stype, comp.imp)M.No  0.060109 0.0246
interaction(stype, comp.imp)E.Yes 0.612022 0.0417
interaction(stype, comp.imp)H.Yes 0.038251 0.0161
interaction(stype, comp.imp)M.Yes 0.076503 0.0217
```

If we type the command:
`svyvar(~api00, dclus1)`
we obtain:

```
      Variance SE
api00 11183    1386.4
```

And, if we type the command:
`svytotal(~enroll, dclus1, deff=TRUE)`
we obtain:

```
       Total   SE     DEff
enroll 3404940 932235 31.311
```

By typing the commands:
`mns = svymean(~api00+api99,dclus1)`
`mns`
we obtain the following:

```
      Mean   SE
api00 644.17 23.542
api99 606.98 24.225
```

7.8 Tables

The `survey` package has two main types of tables:

- Totals or proportions cross-classified by multiple factors;
- Arbitrary statistics in subgroups.

7.8.1 Computing Over Subgroups

Function `svyby()` computes a statistic for subgroups specified by a set of factor variables. We type the command:
`svyby(~api99, ~stype, dclus1, svymean)`
to obtain:

```
stype statistics.api99 se.api99
E E   607.7917         22.81660
H H   595.7143         41.76400
M M   608.6000         32.56064
```

Here: `api99` is the variable to be analysed, `stype` is the subgroup variable, `dclus1` is the design object, and `svymean` is the statistic to compute.

If we type the command:

```
svyby(~api99, ~stype, dclus1, svyquantile, quantiles=0.5,ci=TRUE)
```

we obtain:

```
stype statistics.quantiles   statistics.CIs            se      var
E E            615         525.6174, 674.1479    37.89113 1435.738
H H            593         428.4810, 701.0065    69.52309 4833.460
M M            611         527.5797, 675.2395    37.66903 1418.955
```

To obtain the following, we type the command:

```
svyby(~api99, list(school.type=apiclus1$stype), dclus1, svymean)
```

```
school.type statistics.api99 se.api99
E E                607.7917 22.81660
H H                595.7143 41.76400
M M                608.6000 32.56064
```

Then, typing the command:

```
svyby(~api99+api00, ~stype, dclus1, svymean, deff=TRUE)
```

we obtain:

```
stype statistics.api99 statistics.api00 se.api99 se.api00
E E   607.7917         648.8681         22.81660 22.36241
H H   595.7143         618.5714         41.76400 38.02025
M M   608.6000         631.4400         32.56064 31.60947

DEff.api99 DEff.api00
E 5.895734 6.583674
H 2.211866 2.228259
M 2.226990 2.163900

stype sch.wide statistic.api99 statistic.api00
E.No  E No     601.6667        596.3333
H.No  H No     662.0000        659.3333
M.No  M No     611.3750        606.3750
E.Yes E Yes    608.3485        653.6439
H.Yes H Yes    577.6364        607.4545
M.Yes M Yes    607.2941        643.2353
```

7.9 Cross-Tabulations

Functions `svyby()` or `svymean()` and `svytotal()`, with interaction, will produce the numbers, but the formatting is not pretty. On the other hand, function `ftable()` provides formatting. By typing the following commands:

```
d = svyby(~api99 + api00, ~stype + sch.wide, rclus1, svymean, keep.var=TRUE,
vartype=c("se","cvpct"))
round(ftable(d),1)
```

we obtain:

```
            sch.wide        No                              Yes
             statistics.api99 statistics.api00 statistics.api99 statistics.api00
stype
E       svymean         601.7            596.3            608.3            653.6
        SE               70.0             64.5             23.7             22.4
        cv%              11.6             10.8              3.9              3.4
H       svymean         662.0            659.3            577.6            607.5
        SE               40.9             37.8             57.4             54.0
        cv%               6.2              5.7              9.9              8.9
M       svymean         611.4            606.4            607.3            643.2
        SE               48.2             48.3             49.5             49.3
        cv%               7.9              8.0              8.2              7.7
```

Appendix A

Quick Introduction to *R*

An inference, if it is to have scientific value, must constitute a prediction concerning future data. If the inference is to be made purely with the help of the distribution theories of statistics, the experiments that constitute the evidence for the inference must arise from a state of statistical control; until that state is reached there is no universe, normal or otherwise, and the statistician's calculations by themselves are an illusion if not a delusion. The fact is that when distribution theory is not applicable for lack of control, any inference, statistical or otherwise, is little better than a conjecture. The state of statistical control is therefore the goal of all experimentation.—W Edwards Deming.[a]

[a] WA Shewhart (with the editorial assistance of W Edwards Deming); 1939. *Statistical Method: From the Viewpoint of Quality Control*. Washington, DC: Graduate School, USDA. Page iii.

A.1 Getting and Working with *R*

R is a program for statistical analysis and graphic presentation. It is open-source and it is available to download, free of charge, from CRAN's Website – **http://cran.r-project.org**

R is command-driven: just type a command and press Enter, then it executes the command and prints the result. Then, *R* waits for more input.

Some examples:

```
2 * 5   # multiplication
log(10)   # natural logarithm
sqrt(357)   # square root
x = rnorm(1000)   # generate random numbers
```

`log( )`, `sqrt( )`, and `rnorm( )` are examples of *functions*. Function calls use parentheses; for example:

```
plot(x)
summary(x)
```

A.1.1 Some Restrictions

Variable names cannot start with a digit, names are **Case-Sensitive**. Some common letters already reserved by *R* for special purposes, for example: **c, q, t, C, D, F, I, T**

Elementary data types in *R* are all vectors. The `c(...)` construct is used to create vectors:

```
age = c(60, 72, 57, 40, 25, 72)
age
```

Common arithmetic operations, including: $+$, $-$, $*$, $/$, $\hat{}$, and mathematical functions, e.g.: `max( ), min( ), exp( ), log( )` work element-wise on vectors; and produce another vector. Example:

```
experience = c(23.5, 25, 16.5, 9, 4.5, 30)
summary(age); summary(experience)
```

A.2 Graphics

The simplest way to produce *R* graphics output is to use the `plot` function. For example, a scatter plot of age and experience:

```
plot(age, experience)
```

A histogram is obtained with:

```
hist(x)
```

And a box-plot is given by:

```
boxplot(x)
```

R has many graphic capabilities in one- two- and three-dimensions. You can see a demonstration by typing

```
demo(graphics)
```

A.3 Getting Help

The command, `help.start( )` starts a browser window with an HTML help interface. One of the best ways to get started is using a manual for beginners called **An Introduction to R**. You can find many references to R in CRAN's Contributed Documentation section. The documents are divided into those of 100 or more pages and others of less than 100 pages.

The command, `help(topic)` displays the help page for a particular topic or function. Every *R* function has a help page. The following are equivalent:

```
help(plot)
?  plot
```

If you want to know about a specific subject, but do not know which particular help page has the information, the command `help.search( )` is very useful. For example:

```
help.search("logarithm")
```

A.4 *R* Packages

R makes use of a system of packages. A **package** is a collection of routines with a common theme. And a **library** is a collection of packages. Some packages are already loaded when *R* starts up; other packages need be loaded using function `library( )`.

Many packages are available from CRAN's website, visiting
http://cran.us.r-project.org/src/contrib/PACKAGES.html

At any point, a list of currently loaded packages can be listed using function:

```
search( )
```

Other packages can be loaded by the user. We will be interested in the `sampling` and `survey` packages. Once installed, these can be loaded using:

```
library(sampling)
```

And

```
library(survey)
```

New packages can be downloaded and installed using function `install.packages()`. For example, to install the `UsingR` package, one can type:

```
install.packages("UsingR")
library(help = UsingR)
```

The last command gives a list of all help pages in said package.

A.5 Data Types

R works with four data types. These are:

vector A set of units of the same mode in a specified order. We have already seen several examples of vectors.

matrix A two-dimensional array of elements of the same mode. For example, a correlation matrix.

factor Vector of categorical data. For example, `sex = c("F", "M", "F", "F", "M")`

data frame Two-dimensional array whose columns may represent data of different modes. We will work with many examples of data frames in our course.

list A set of components that can be any other object type. Many of *R*'s output are in a form of a list.

Appendix B

Accessing Data using *R*

Use of data requires also understanding of the distinction between enumerative studies and analytic problems. An enumerative study produces information about a frame. The theory of sampling and design of experiments are enumerative studies. Our Census is an enumerative study ... The interpretation of a test or experiment is something else. It is prediction that a specific change in a process or procedure will be a wise choice, or that no change would be better. Either way the choice is prediction. This is known as an analytic problem, or a problem of inference, prediction. Test of significance, t-test, chi-square, are useless as inference—i.e., useless for aid in prediction. Test of hypothesis has been for half a century a bristling obstruction to understanding statistical inference.[a]

[a] W Edwards Deming; 1993. *The New Economics: For Industry, Government, Education*. Cambridge, MA: MIT CAES. Page 103 *et seq.*

B.1 Reading Text File

Suppose that we want to read data from external text files. Also, we may want to read data from statistical programs like: Stata, SPSS, SAS, Minitab, and others. We might want to read data from other formats, like for example, to read data from Excel.

In general, *R* is not well suited to manipulate large-scale data. Therefore, we can read text (ASCII) files, which is the easiest form to import into *R*. Function, `scan( )` is used to read real (numeric) data. And `read.table( )` is used to read data frames.

B.1.1 Using read.table( )

Suppose data set (matrix) called "numbers.txt" is in our working directory. To read said file type:

```
mat.1 = read.table("numbers.txt")
mat.1
fix(mat.1)
```

Now, suppose data frame called "savings&loan.txt" is in our working directory. Also, suppose it has column headers. To read said data frame type:

```
df.1 = read.table("savings&loan.txt", header = TRUE)
df.1
fix(df.1)
```

B.2 Text Files Export

Normally, a text file will be convenient to export. A common task is to write a matrix or data frame to a file, which is done by the functions write.table( ) and write( ).

Function write( ) writes out a matrix or vector in a specified number of columns (and transposes a matrix). And function write.table( ) is conveniently used to write out a data frame with row and column labels.

B.3 Data From Other Statistics Programs

Package foreign provides import facilities from programs: Stata, Minitab, SPSS, SAS, and others. Suppose for example, a Stata data set "Apparatus Quality.dta" is in our working directory. To read said data, type the following commands:

```
library(foreign)
apparatus = read.dta("Apparatus Quality.dta")
str(apparatus)
save(apparatus, file="apparatus.RData")
```

B.3.1 Other Statistics Programs

For other statistics programs, ask for help:

```
?  read.mtp # MTB Worksheet
?  read.spss # SPSS Data File
?  read.ssd # SAS Dataset
?  read.epiinfo # Epi-Info Data
?  read.systat # Systat Data
?  read.dbf # DBF File
```

B.4 Reading Excel Spreadsheets

To access Excel files, we have several options. With Excel data in tab-delimited or comma-separated format, use `read.delim( )` or `read.csv( )`, to import it into *R*. Another possibility is that you can export to a DIF file and read it using `read.DIF( )`.

We can copy-and-paste between the display of a spreadsheet in such a program and *R*, then use `read.table( )`. For Windows, the package `xlsReadWrite` has a function `read.xls( )` to read `.xls` files.

Example: Suppose data set `"Salaries.xls"` is in our working directory. To access it, type the following commands:

```
library(xlsReadWrite)
salaries = read.xls("Salaries.xls")
str(salaries)
save(salaries, file="salaries.RData")
```

Appendix C

Review of Basic Probability

What is consumer research? I have mentioned several times the need for statistical surveys for consumer research ... As I said earlier, the terms "good quality" and "quality control" have no meaning except with reference to the consumer's needs ... The main use of consumer research is to feed consumer reactions back into the design of the product ... Consumer research takes the pulse of the consumer's reactions and demands, and seek explanations for the findings ... Real consumer research, geared to design and production, is an indispensable modern tool for modern problems.[a]

[a] W Edwards Deming; 1950. *Elementary Principles of The Statistical Control of Quality.* Tokio: Japanese Union of Scientists and Engineers. Page 7.

C.1 Types of Probability

Classical Event's A probability is the ratio of the number of favorable outcomes M and all possible outcomes N in an experiment.

$$\Pr(A) = \frac{M}{N} \tag{C.1}$$

Empirical Event's A probability is the proportion of times that the event occurs, if the same experiment is repeated many times.

- Suppose that the experiment is repeated N times, and you observe that event A occurred $M < N$ times.
- Then,

$$\Pr(A) = \lim_{N \to \infty} \frac{M}{N} \tag{C.2}$$

Subjective It is an individual's degree of belief on the occurrence of an event.

Examples of classical probability — games of chance: urns, cards, dice, roulletes

Examples of empirical probability:

1. Buffon: $N = 4,040$ Coin tosses, $M =$ 2,048 Heads, then $\Pr(H) = 0.505$
2. K Pearson: $N = 12,000$ Coin tosses, $M =$ 6,019 Heads, then $\Pr(H) = 0.502$
3. K Pearson: $N = 24,000$ Coin tosses, $M =$ 12,012 Heads, then $\Pr(H) = 0.501$
4. Coin toss simulation using R is easy!

```
 # Simulation of 50,000 tosses
  x = c(0,1); y=sample(x, 50000, replace = T)
pr.H = sum(y)/50000; pr.H
```

Examples of subjective probability:

1. What is the probability for you of getting A in Economics?
2. What is the probability of life in Mars?
3. What is the probability that hypothesis H is true?

C.2 Probability Rules

Sample space (S) – includes all possible outcomes of interest

Null event (ϕ) – empty; contains no outcomes

$$\Pr(S) = 1 \quad \text{and} \quad \Pr(\phi) = 0.$$

If A & B are two events in S:

$$0 \leq \Pr(A) \leq 1 \quad \text{and} \quad 0 \leq \Pr(B) \leq 1.$$

Complementary event (A') – non-occurrence of event A

$$\Pr(A') = 1 - \Pr(A). \tag{C.3}$$

Mutually exclusive events – If events A & B cannot occur simultaneously

$$\Pr(A \cup B) = \Pr(A) + \Pr(B). \tag{C.4}$$

Otherwise:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{C.5}$$

C.3 Conditional Probability

We want the probability of event A, given that event B has occurred. By definition, in symbols:

$$\Pr(A \mid B) \equiv \Pr(A \cap B) \div \Pr(B). \tag{C.6}$$

Similarly:

$$\Pr(B \mid A) \equiv \Pr(A \cap B) \div \Pr(A). \tag{C.7}$$

Therefore:

$$\begin{aligned} \Pr(A \cap B) &= \Pr(A) \cdot \Pr(B \mid A) \\ &= \Pr(B) \cdot \Pr(A \mid B). \end{aligned} \tag{C.8}$$

C.4 Probability Table

A probability table contains both joint and marginal probabilities. It looks as follows:

Outcome	B	B'	**Marginal**
A	$\Pr(A \cap B)$	$\Pr(A \cap B')$	$\Pr(A)$
A'	$\Pr(A' \cap B)$	$\Pr(A' \cap B')$	$\Pr(A')$
Marginal	$\Pr(B)$	$\Pr(B')$	1.00

C.4.1 Probability Table Example

As an example, we classify people according to academic degree and position as follows:

Outcome	Mgt	Mgt'	**Marginal**
BBA	0.12	0.58	0.70
BBA'	0.08	0.22	0.30
Marginal	0.20	0.80	1.00

C.5 Bayes' Rule

Total Probability – Suppose that event B occurs only if event A or event A' has occurred, i.e.:

$$B \in A \cup A' = S.$$

Then,

$$B = (A \cap B) \cup (A' \cap B).$$

Therefore,

$$\begin{aligned} \Pr(B) &= \Pr(A \cap B) + \Pr(A' \cap B) \\ &= \Pr(A) \cdot \Pr(B \mid A) + \Pr(A') \cdot \Pr(B \mid A'). \end{aligned} \tag{C.9}$$

This is called the **marginal** or **total probability** of event B

Thomas Bayes (1763) wanted to know the inverse probability of event A, given that B occurred. Based on equation (C.9) above, by conditional probability:

$$\begin{aligned} \Pr(A \mid B) &= \Pr(A \cap B) \div \Pr(B) \\ &= \Pr(A) \cdot \Pr(B \mid A) \div \Pr(B) \\ &= \frac{\Pr(A) \cdot \Pr(B \mid A)}{\Pr(A) \cdot \Pr(B \mid A) + \Pr(A') \cdot \Pr(B \mid A').} \end{aligned} \tag{C.10}$$

And this is known as **Bayes' Rule of Inverse Probability**.

C.5.1 Bayes' Rule Example

Suppose that your firm adopts a drug testing program to all employees. We know that laboratory testing is not perfect. Imagine that:

- If an employee uses drug, the test indicates positive with 90% probability;
- If an employee does not use drug, the test indicates positive with 5% probability;
- Based on past history, 4% of selected employees use drugs;
- A randomly selected employee was tested and got a positive result;
- After test result, what is the probability that he uses drugs?

In tabular form, the probabilities that he uses drugs or not, before and after getting a positive test, are computed in the following table:

Event	Pr(:)	Pr(+| :)	Prod.	Pr(: |+)
Use	0.04	0.90	0.036	0.43
No Use	0.96	0.05	0.048	0.57
Total	1.00	xxx	0.084	1.00

Bibliography

[1] Abad A; 1982. *Introducción al Muestreo, 2da. Edición.* México: Limusa Noriega.

[2] Aday LA; 1996. *Designing and Conducting Health Surveys, 2nd. Edition.* San Francisco: Jossey-Bass.

[3] Aday LA & SB Cornelius; 2006. *Designing and Conducting Health Surveys: A Comprehensive Guide, 3rd Edition.* San Francisco: Jossey-Bass.

[4] Biemer PP & LE Lyberg; 2003. *Introduction to Survey Quality.* New York: John Wiley & Sons.

[5] Chambers RL & CJ Skinner (Editors); 2003. *Analysis of Survey Data.* New York: John Wiley & Sons.

[6] Chaudhuri A & H Stenger; 1992. *Survey Sampling: Theory & Methods.* M Dekker.

[7] Cochran WG; 1977. *Sampling Techniques, 3rd. Edition.* New York: John Wiley & Sons.

[8] Deming WE; 1950, 1986. *Some Theory of Sampling.* New York: John Wiley & Sons, Dover.

[9] Deming WE; 1960, 1990. *Sampling Design in Business Research.* Wiley Classics Library.

[10] Dillman DA; 2007. *Mail and Internet Surveys: The Tailored Design Method, 2nd Edition.* New Jersey: John Wiley & Sons.

[11] Foreman EK; 1991. *Survey Sampling Principles.* M Dekker.

[12] Frey JH; 1989. *Survey Research By Telephone.* Newbury Park, CA: Sage.

[13] Ghosh, M & G Meeden; 1997. *Bayesian Methods for Finite Population Sampling.* Chapman & Hall.

[14] Govindarajulu Z; 1999. *Elements of Sampling: Theory and Methods.* New Jersey: Prentice-Hall.

[15] Groves RM & RL Kahn; 1979. *Survey By Telephone*. New York: Academic Press.

[16] Groves RM *et al.*; 2001. *Telephone Survey Methodology*. New York: John Wiley & Sons.

[17] Hansen MM, WN Hurwitz and WG Madow; 1953, 1993. *Sample Survey Methods & Theory*. (2 vols.) New York: John Wiley & Sons.

[18] Henry GT; 1990. *Practical Sampling*. Newbury Park, CA: Sage.

[19] Hess I; 1975. *Practical Sampling for Hospitals & Patients, 2nd. Edition*. Health Administration Press.

[20] Kalton G; 1985. *Introduction to Survey Sampling, 3rd. Edition*. Newbury Park, CA: Sage.

[21] Kiaer AN; 1897. *The Representative Method of Statistical Surveys*. Oslo: Kristiania.

[22] Kish L; 1965. *Survey Sampling*. New York: John Wiley & Sons.

[23] Kish L; 1987. *Statistical Design for Research*. New York: John Wiley & Sons.

[24] Korn EL & BI Graubard; 1999. *Analysis of Health Surveys*. New York: John Wiley & Sons.

[25] Lesser VM; 1992. *A Comparison of Periodic Survey Designs Employing Multi-Stage Sampling*. University of North Carolina.

[26] Lessler JT & WD Kalsbeek; 1992. *Nonsampling Error in Surveys*. New York: John Wiley & Sons.

[27] Lehtonen R & E Pahkinen; 2004. *Practical Methods for Design and Analysis of Complex Surveys, 2nd. Edition*. Chichester: John Wiley & Sons.

[28] Levy PS & S Lemeshow; 1999. *Sampling of Populations: Methods and Applications, 3rd. Edition*. New York: John Wiley & Sons.

[29] Levy PS & S Lemeshow; 1980. *Sampling for Health Professionals*. Wadsworth.

[30] Lohr SL; 1999. *Sampling: Design & Analysis*. Pacific Grove, CA: Duxbury Press.

[31] Lwanza SK; 1991. *Sample Size Determination in Health Studies: A Practical Manual*. Geneva: World Health Organization.

[32] Lybeg L *et al.* (Editors); 1997. *Survey Measurement & Process Quality*. New York: John Wiley & Sons.

[33] McCarthy PC. *The Bootstrap & Finite Population Sampling*. National Center for Health Statistics

[34] Murthy NN; 1967, 1977. *Sampling Theory & Methods, 2nd. Impression*. Calcutta: Statistical Publishing Society.

[35] Raj D; 1968. *Sampling Theory*. New York: McGraw-Hill.

[36] Raj D; 1972. *The Design of Sample Surveys*. New York: McGraw-Hill.

[37] Särndal CE, B Sweenson, and J Wretman; 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

[38] Scheafer RL *et al.*; 1986. *Elementary Survey Sampling, 3rd. Edition*. PWS Publishers.

[39] Scheafer RL *et al.*; 1987. *Elementos de Muestreo*. México: Editorial Ibero-América.

[40] Singh R & NS Mangat; 1996. *Elements of Survey Sampling*. Springer.

[41] Singh S; 2004. *Advanced Sampling Theory with Applications*. Springer.

[42] Slonim MJ; 1960. *Sampling*. New York: Simon & Schuster.

[43] Stuart A; 1984. *The Ideas of Sampling*. London: Griffin.

[44] Sudman S; 1976. *Applied Sampling*. New York: Academic Press.

[45] Sukhatme PV, BV Sukhatme, S Sukhatme, & C Asok; 1984. *Sampling Theory of Surveys with Applications, 3rd Revised Edition*. Ames: Iowa State University Press.

[46] Thompson ME; 1997. *Theory of Sample Surveys*. London: Chapman & Hall.

[47] Thompson SK; 2002. *Sampling, 2nd. Edition*. New York: John Wiley & Sons.

[48] United Nations; 2005. *Designing Household Survey Samples: Practical Guidelines*. Department of Economic and Social Affairs, Statistics Division Studies in Methods Series F No. 98.

[49] Yates F; 1960. *Sampling Techniques for Censuses & Surveys, 3rd. Edition*. Griffin.

Index

Advantages, 2
Average within strata standard deviation, 43

Bayes' rule, 93
Between strata variance, 42
bias, 6, 14

Census, 2
Circular Systematic Sampling, 24
Clustered Sampling, 8
Clustered sampling, 49
Clustered sampling advantages, 50
Clustered sampling allocation, 54
Clustered sampling estimation, 53
Clustered sampling two-stage, 51
Coefficient of variation, 14
coefficient of variation, 19
Correlation Coefficient, 15
Correlation coefficient, 15
Covariance, 15

definitions
  Average within strata standard deviation, 43
  Between strata variance, 42
  Circular Systematic Sampling, 24
  Correlation Coefficient, 15
  Covariance, 15
  Expected Value, 6
  Expected Values, 13
  Intra-class Correlation, 26
  Linear Systematic Sampling, 24
  Mean, 4
  Mean Square Error, 6
  Median, 4
  Random Start, 23
  Relative Variability, 4
  Relative Variance, 14
  Sampling Variance, 6
  SRS, 11
  Standard Deviation, 4
  Variance, 4
  Variances, 13
  Within strata variance, 42

Expected Value, 6
expected value, 17
Expected Values, 13

frame, 1

Intra-class Correlation, 26

Limitations, 3
Linear Systematic Sampling, 24

Mean, 4
Mean Square Error, 6
Mean square error, 14
Median, 4

Non-response errors, 7
non-sampling errors, 7

Parameter, 2
Population, 1
PPS clustered sampling, 55
PPS selection, 57
precision, 6
Probability, 91
probability, 12
Probability rules, 92
probability sampling, vii, 1
Proportion estimation, 18

R data access, 87
R data types, 85
R packages, 85
R program, 13, 27, 51, 55, 57, 62, 69
R program graphics, 84
R program introduction, 83
Random Start, 23
Random start, 23
Ratio estimation, 61
Ratio estimation in stratified sampling, 65
ratio estimation separate, 66
Regression estimation, 66
Regression simple, 68
Regression trough the origin, 67
Relative Variability, 4
Relative Variance, 14
Relative variance, 14
repeated sampling, 13
Replicated Sampling, 8
Replicated sampling, 28, 32
Replicated sampling estimation, 29, 30
Replicated stratified sampling, 44
Response errors, 7

Sample, 2
sample, 5
Sample size, 19
Sampling distribution, 14
sampling frame, 5
Sampling probability, 25
Sampling Variance, 6
Simple Random Sampling, 8, 11
SRS, 11
Standard Deviation, 4
Stratified Sampling, 8
Stratified sampling, 35
Stratified sampling limitation, 45
survey cross-tabulations, 82
survey estimation, 74
survey package, 73
survey PPS, 78
survey stratified, 75
survey summary, 79
survey tables, 80
survey two-stage, 77
Systematic Sampling, 8
Systematic sampling, 23, 24, 26
systematic sampling, 24

Thomas Lumley, 74

Variance, 4
variance, 4
variance between strata, 42
Variance comparison, 42
Variance of linear functions, 15
Variances, 13
Variances estimation, 17
Variances in SRS, 16

W Edwards Deming, vii, 1, 7, 9, 11, 21, 23, 35, 49, 61, 73, 83
Within strata variance, 42