Surveys, Sampling and Observational Data

Derek Li

Contents

1 Review **4**

1.1 Basic Definition . . . 4

1.2 Basic Notations . . . 4

1.3 Population Parameters . . . 4

1.3.1 Population Mean (μ_y) . . . 4

1.3.2 Population Variance (σ_y^2) . . . 5

1.3.3 Population Total (τ_y) . . . 5

1.3.4 Population Proportion . . . 5

1.3.5 Population Ratio . . . 5

1.4 Basic Rules from Probability . . . 5

1.5 Sample . . . 6

1.6 Estimation . . . 6

1.6.1 Variance of Estimator . . . 6

1.6.2 Properties of Estimators . . . 7

1.6.3 Error of Estimation . . . 7

2 Elements of the Sampling Problem **8**

2.1 Definition . . . 8

2.2 Design of Sample Survey . . . 8

2.2.1 Preparation and Execution . . . 8

2.2.2 Methods to Select a Sample . . . 9

2.3 Random Sampling . . . 9

2.4 Errors in Surveys . . . 10

2.4.1 Inadequate Frame-Coverage Error . . . 10

2.4.2 Selection and Interviewer Bias . . . 10

2.4.3 Response Error . . . 10

2.4.4 Non-Response Bias . . . 11

2.4.5 Summary of Common Sampling Mistakes . . . 11

2.5 Data Collection . . . 11

2.6 Designing a Questionnaire . . . 12

2.6.1 Basic Problems of a Questionnaire . . . 12

2.7 Others Methods of Studies: Observational or Experimental . . . 12

2.7.1 Observational Study . . . 12

3 Simple Random Sampling 13
3.1 Probability Sampling Designs 13
3.1.1 Condition 13
3.1.2 Basic Notation 13
3.2 Simple Random Sampling 13
3.2.1 Combinatorial Notation 14
3.2.2 Simple Random Sampling without Replacement 14
3.3 How to Draw a SRS 14
3.3.1 Table of Random Numbers (TRN) 15
3.4 Inference 15
3.4.1 Estimation of the Population Mean 15
3.4.2 Estimation of the Population Variance 17
3.4.3 Estimation of the Population Standard Deviation 18
3.4.4 Finite Population Correction 18
3.4.5 Estimation of the Population Total 19
3.4.6 Confidence Intervals of μ and τ 19
3.4.7 Estimation of the Population Proportion p Using SRS 19
3.5 Selecting Sample Size for Estimating μ, τ and p Using SRS 20

4 Ratio, Regression, and Difference Estimation 22
4.1 Ratio Method Using SRS 22
4.1.1 Ratio Estimation of the Population Ratio $R = \frac{\mu_y}{\mu_x}$ 22
4.1.2 Ratio Estimation of the Population Total τ_y 22
4.1.3 Ratio Estimation of the Population Mean 22
4.1.4 Sample Size Required for Estimating R 23
4.1.5 Sample Size Required for Estimating μ_y 23
4.1.6 Sample Size Required for Estimating τ_y 23
4.2 Regression Estimation Using SRS 23
4.2.1 Regression Estimation of the Population Mean 24
4.3 Difference Estimation Using SRS 24
4.3.1 Difference Estimation of the Population Mean 24

5 Stratified Random Sampling 25
5.1 How to Draw a Stratified RS 25
5.2 Estimation 25
5.2.1 Estimation of the Population Mean 26
5.2.2 Estimation of the Population Total 26
5.2.3 Estimation of the Population Proportion 26
5.3 Sample Size 27
5.3.1 Sample Size Estimation 27
5.3.2 Allocation of Sample Size 27
5.3.2.1 Equal Allocation 27
5.3.2.2 Proportion Allocation 27
5.3.2.3 Neyman Allocation 27
5.3.2.4 Optimal Allocation 28
5.3.3 Sample Size Required for Estimating μ and τ 28
5.3.4 Sample Size Required for Estimating p 29
5.4 Ration Estimation 29
5.4.1 Separate Ratio (SR) Estimator 29
5.4.1.1 SR for Estimating R 29

5.4.1.2 SR for Estimating μ_y . . . 30
5.4.2 Combined Ratio (CR) Estimator . . . 31
5.4.2.1 CR for Estimating R . . . 31
5.4.2.2 CR for Estimating μ_y . . . 31
5.4.2.3 SR and CR for Estimating τ . . . 31

6 Systematic Sampling **32**
6.1 How to Draw a Systematic Sample . . . 32
6.2 Estimation . . . 33
6.2.1 Estimation of μ and τ . . . 33
6.2.1.1 Sampling Distribution . . . 33
6.2.1.2 Population ANOVA . . . 34
6.2.1.3 ICC . . . 35
6.2.2 Estimation of $\text{Var}[\overline{y}_{\text{sys}}]$ . . . 35
6.2.2.1 Difference Method . . . 36
6.2.2.2 Repeated Systematic Sampling (RSS) . . . 36
6.2.2.3 Rules . . . 37
6.3 Selecting Sample Size for Estimating μ, τ and p . . . 37

7 One-Stage Cluster Sampling **38**
7.1 Ratio Estimation . . . 38

1 Review

1.1 Basic Definition

Definition 1.1. ***Random experiment*** is the process of observing the outcome of a chance event.

Definition 1.2. ***Elementary outcomes*** are all possible results of the random experiment.

Definition 1.3. ***Sample space*** (Ω) is the set of all the elementary outcomes.

Definition 1.4. ***Random variable*** Y is a real-valued function defined over a sample space.

Definition 1.5. ***Variable*** is the function defined on population elements, characteristic of population elements. Variable can be quantitative (numerical) or qualitative (categorical).

Definition 1.6. ***Distribution*** or ***frequency distribution*** is the proportion of elements with value in an interval $[a, b], \forall a, b$.

1.2 Basic Notations

- Population: $E = \{e_1, e_2, \cdots, e_N\}$ with population size N, where e_i's are elements.
- Variable: $y, x, z, t, \cdots$.
- Range: $\{y(e), e \in E\}$.
- Probability: In discrete case,

$$P(y_i) = \frac{|\{e, y(e) = y_i\}|}{N} = \frac{N_i}{N}.$$

In continuous case,

$$P(a, b) = P(a < y < b) = \int_a^b f(y)\mathrm{d}y,$$

where $f(y)$ is the density function s.t.

$$f(y) \geqslant 0, \forall y \text{ and } \int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1.$$

1.3 Population Parameters

1.3.1 Population Mean (μ_y)

- Using distribution:

$$\mu_y = \sum_{i=1}^{k} y_i P(y_i) = \frac{1}{N}\sum_{i=1}^{k} N_i y_i.$$

- Using population values:

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

1.3.2 Population Variance (σ_y^2)

- Using distribution:

$$\sigma_y^2 = \sum_{i=1}^{k}(y_i - \mu_y)^2 P(y_i) = \frac{1}{N}\sum_{i=1}^{k} N_i(y_i - \mu_y)^2.$$

- Using population values:

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

- Population standard deviation is $\sigma_y = \sqrt{\sigma_y^2}$.

1.3.3 Population Total (τ_y)

$$\tau_y = \sum_{i=1}^{N} y(e_i) = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} N_i y_i = N\mu_y.$$

1.3.4 Population Proportion

Define

$$y(e) = \begin{cases} 0, & e \text{ does not have the property} \\ 1, & e \text{ has the property} \end{cases},$$

then

$$p = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{M}{N} = \mu,$$

where M is the number of elements with the property.

1.3.5 Population Ratio

Ratio of two variables' means or totals:

$$R_{y/x} = \frac{\mu_y}{\mu_x} = \frac{N\mu_y}{N\mu_x} = \frac{\tau_y}{\tau_x}.$$

1.4 Basic Rules from Probability

In probability, the covariance of X and Y is

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

In statistics, the covariance of x and y is

$$\begin{aligned}\mathrm{Cov}(x,y) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \mu_x\mu_y \\ &= \frac{1}{N}\sum_{i,j} N_{ij}(x_i - \mu_x)(y_j - \mu_y) = \frac{1}{N}\sum_{i,j} N_{ij} x_i y_j - \mu_x\mu_y.\end{aligned}$$

In probability, the correlation of X and Y is

$$\rho_{X,Y} = \rho_{Y,X} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

In statistics, the correlation of x and y is

$$\rho_{x,y} = \rho_{y,x} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}.$$

1.5 Sample

Definition 1.7. ***Sample*** is a subset of the population.

Definition 1.8. ***Random sample*** is a sequence of random variables (independent or dependent)

$$Y_1 = y_1, \cdots, Y_n = y_n,$$

where Y_i is the random variable and y_i is the obtained value.

Definition 1.9. ***Statistic*** is the function of sample values, such sample mean (average), sample variance, etc.

Definition 1.10. ***Sampling distribution*** is the distribution of the sample function. This distribution depends on the population distribution of y and function f.

Theorem 1.1 (Central Limit Theorem)**.** If one takes random sample from a population of mean μ and standard deviation σ, then as n gets large, $\overline{X}$ approaches the normal distribution with mean μ and variance $\frac{\sigma}{\sqrt{n}}$.

1.6 Estimation

Definition 1.11. An ***estimator*** $\widehat{\theta} = \phi(y_1, \cdots, y_n)$ is a sample function used to estimate population parameter θ.

Definition 1.12. An ***estimate*** of $\theta : \widehat{\theta}_n = \phi_n(y_1, \cdots, y_n)$ is the value of the sample function $\widehat{\theta}$ obtained from sample of size n.

Example 1.1. Suppose $\eta = \psi(\theta)$, then $\widehat{\eta} = \widehat{\psi}(\theta) = \psi(\widehat{\theta})$.

1.6.1 Variance of Estimator

Suppose $\mathrm{Var}[\widehat{\theta}] = \psi(\theta)$, then $\widehat{\mathrm{Var}}[\widehat{\theta}] = \psi(\widehat{\theta})$.

Example 1.2. In sample random sampling, $\widehat{\mu} = \overline{y}, \widehat{\sigma}^2 = S^2$. We have

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}[\overline{y}] = \frac{\sigma^2}{n},$$

then

$$\widehat{\mathrm{Var}}(\widehat{\mu}) = \frac{S^2}{n}.$$

1.6.2 Properties of Estimators

Definition 1.13. If $\mathbb{E}[\widehat{\theta}] = \theta$, then the estimator is called ***unbiased***.

Definition 1.14. ***Bias*** of $\widehat{\theta}$ is defined as

$$B(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta.$$

Definition 1.15. If $\mathbb{E}[\widehat{\theta}] \to \theta$ as $n \to \infty$, then the estimator is called ***asymptotically unbiased***.

Definition 1.16. If $\widehat{\theta}_1, \widehat{\theta}_2$ are both unbiased estimators of θ and $\mathrm{Var}[\widehat{\theta}_1] < \mathrm{Var}[\widehat{\theta}_2]$, then $\widehat{\theta}_1$ is more ***efficient*** than $\widehat{\theta}_2$.

Note that estimator should be at least asymptotically unbiased and has small variance at least if sample size n is large. Also, an unbiased estimator may not have small variance.

Definition 1.17. The ***mean squared error*** of θ is

$$\mathrm{MSE} = \mathrm{Var}[\widehat{\theta}] + (B(\widehat{\theta}))^2 = \mathrm{Var}[\widehat{\theta}] + (\mathbb{E}[\widehat{\theta}] - \theta)^2.$$

Definition 1.18. $\widehat{\theta}$ is consistent for θ if

$$\lim_{n\to\infty} \mathrm{Var}[\widehat{\theta}] = 0.$$

We have some rules for comparing estimators:

- If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators for θ and $\mathrm{Var}[\widehat{\theta}_1] < \mathrm{Var}[\widehat{\theta}_2]$, then we prefer $\widehat{\theta}_1$.
- If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are biased estimators and $\mathrm{MSE}(\widehat{\theta}_1) < \mathrm{MSE}(\widehat{\theta}_2)$, then we prefer $\widehat{\theta}_1$.

1.6.3 Error of Estimation

The error of estimation $|\widehat{\theta} - \theta|$ is not known in practice.

Definition 1.19. The ***error bound*** B_α at level $1-\alpha$ is a value s.t.

$$P(|\widehat{\theta} - \theta| < B_\alpha) = 1 - \alpha.$$

We say B_α is an error bound with confidence $1-\alpha$ and the ***confidence interval*** is

$$[\widehat{\theta} - B_\alpha, \widehat{\theta} + B_\alpha].$$

Example 1.3. Suppose $\widehat{\theta} \sin \mathcal{N}(\theta, \sigma^2_{\widehat{\theta}})$, then

$$\begin{aligned} P(|\widehat{\theta} - \theta| < B_\alpha) &= P(-B_\alpha < \widehat{\theta} - \theta < B_\alpha) \\ &= P\left(-\frac{B_\alpha}{\sigma_{\widehat{\theta}}} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < \frac{B_\alpha}{\sigma_{\widehat{\theta}}}\right) = \left(-\frac{B_\alpha}{\sigma_{\widehat{\theta}}} < Z < \frac{B_\alpha}{\sigma_{\widehat{\theta}}}\right) = 1 - \alpha. \end{aligned}$$

Suppose $Z_{\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$ are the critical values of Z, then $P(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}) = 1 - \alpha$.

Therefore,

$$B_\alpha = Z_{\frac{\alpha}{2}} \sigma_{\widehat{\theta}} = Z_{\frac{\alpha}{2}} \sqrt{\mathrm{Var}[\widehat{\theta}]}.$$

2 Elements of the Sampling Problem

2.1 Definition

Definition 2.1. ***Element*** is the object on which measurement is taken.

Definition 2.2. ***Population*** is a set of elements defined according to the aims and objects of the survey.

Definition 2.3. ***Target population*** is the population intended to be investigated (sampled).

Definition 2.4. ***Sampling population*** is the population effectively sampled.

Definition 2.5. ***Sampling units*** is the non-overlapping collections of elements that cover the population effectively sampled.

Definition 2.6. ***Frame*** is the list or any technical device which provides sampling units, or access to sampling units.

Definition 2.7. ***Sample*** is the collection of sampling units selected from a frame.

2.2 Design of Sample Survey

2.2.1 Preparation and Execution

The general procedure is

- Identify target survey group.
- Develop questions.
- Pilot or test the questions/surveys.
- Determine the method of conducting the survey.
- Conduct the survey.
- Use an appropriate analysis technique to analyze the information collected.

Here are the steps in sampling process:

- Define the population.
- Identify the sampling frame.
- Select a sampling design or procedure (probability sampling or nonprobability sampling)
- Determine the sample size.
- Draw the sample.

2.2.2 Methods to Select a Sample

- ***Census***: Complete survey of a population.
- ***Probability/Random sampling***: Every element in the population has a known nonzero probability (not necessarily equal) of being sampled and involves random selection at some point.
- ***Nonprobability/Nonrandom sampling***: Accidental sampling, quota sampling and purposive/judgmental sampling.

  ∗ ***Quota sampling***: The criteria for selection of elements are based on quotas - assumptions regarding the population of interest. After the quotas are decided, the choice of actual sampling units to fit into quotas is mostly left to the interviewer.

The key problem with nonrandom sampling is the selection of elements does not allow proper estimation of sampling errors.

2.3 Random Sampling

- ***Simple random sampling***: Sample is selected completely at random. No special constraints on the sample are imposed. Note that it is an unbiased sample.
- ***Stratified sampling***: The population is divided into sub-populations (strata) and random sample is selected from every stratum. The constraint imposed is stratification.
- ***Cluster sampling*** (***one stage***): The population is divided into large number of (small) clusters, equal or nonequal and clusters are selected at random. The sample is all elements from selected clusters and the constraint imposed is clustering.

  ∗ For stratified sampling, population divided into **few** subgroups; for cluster sampling, population divided into **many** subgroups.

  ∗ For stratified sampling, **homogeneity within subgroups** and **heterogeneity between subgroups**; for cluster sampling, **heterogeneity within subgroups** and **homogeneity between subgroups**.

  ∗ For stratified sampling, choosing elements from within each subgroup; for cluster sampling, random choosing subgroups.

- ***Cluster sampling*** (***two stage***): The population is divided into large number of (bigger) clusters. Sample of clusters is selected and then sample comes from each selected cluster. The sample is all selected elements and the constraint imposed is clustering.
- ***Cluster sampling*** (***multi-stage***): The population is divided into clusters on several levels and sampling is performed at every stage. Sampling is performed at every stage. The sample is all sampling units selected at the last stage and the constraint imposed is multi-clustering.
- ***Double sampling*** (***two-phase sampling***): Sample from sample. Take a bigger sample with some basic measurements and then take a subsample from previously selected sample with more detailed measurements. The sample is the selected elements from the second phase.

  Example 2.1. Phase I: Large sample and stratification variable. Phase II: Subsample and variable of interest.

- ***Systematic sampling***: Elements are selected from an ordered sampling frame. First element is selected at random, and subsequent elements follow a predetermined pattern, usually an interval.

- ***Composite designs***: Most large scale surveys are done using cluster sampling combined with stratification.

  Example 2.2. Clusters are selected from each stratum and then elements are selected from each cluster.

2.4 Errors in Surveys

Definition 2.8. ***Sampling errors*** are due to random sampling, controlled by sample design, sample size and error bound.

Definition 2.9. ***Non-Sampling errors*** are caused by factors other than those related to sample selection. They are not easily identified or quantified.

Non-Sampling errors may come from:

- Imperfect sampling population (inadequate frame-coverage error)
- Poorly designed questionnaire
- Selection bias, sampling bias
- Non-Response problem
- Response error (inaccurate response)
- Systematic error (interviewer bias)
- Processing, editing, entering error

2.4.1 Inadequate Frame-Coverage Error

The sampling design excludes or under-represents a specific groups in the sample, deliberately or not. If the group is different, with respect to survey issues, bias will occur.

2.4.2 Selection and Interviewer Bias

- Voluntary response bias: Sample members are self-selected volunteers, as in voluntary samples. Individuals with strong opinions about the survey issues or those with substantial knowledge will tend to be over-represented, creating bias.

- Interviewer error: interviewers incorrectly record information; are not neutral or objective; influence the respondent to answer in a particular way; or assume responses based on appearance or other characteristics.

2.4.3 Response Error

Respondents intentionally of accidentally provide inaccurate responses. It occurs when concepts, questions or instructions are not clearly understood by the respondent; there are high levels of respondent burden and memory recall required; some questions can result in a tendency to answer in a socially desirable way; questions are sensitive.

2.4.4 Non-Response Bias

We fail to obtain a response from some units because of absence, non-contact, refusal, not-able or some other reason. There are two types:

1. Complete/Unit non-response: No data has been obtained at all from a selected unit .
2. Partial/Item non-response: The answers to some questions have not been provided by a selected unit.

Non-Response bias will occur if people who refuse to answer are different, with respect to survey issues, from those who respond. If the response rate is low, bias can occur because respondents may tend consistently to have views that are more extreme than those of the population in general.

2.4.5 Summary of Common Sampling Mistakes

- Use a bad sampling frame: An SRS from an incomplete or out-of-data sampling frame introduces bias because the individuals included may differ from the ones not in the frame.
- Undercoverage bias occurs when a portion of the population is systematically left out of the selection process.
- Nonresponse bias occurs when a portion of the selected sample declines to participate in the study.
- Response bias occurs when individuals give what the perceive to the preferred response rather than their true opinion. It refers to anything in the survey design that influences the responses.
- In convenience sampling, we simply include the individuals who are convenient for us to sample.
- In voluntary response sample, a large group of individuals is invited to respond, and those who choose to respond are counted. The respondents, rather than the researcher, decide who will be in the sample.

2.5 Data Collection

- Direct measurement: observation
- Personal interview (face-to-face)
- Telephone interview
- Mailed questionnaire
- Traditional: paper-and-pencil interviewing
- Modern: computer-assisted interviewing
- Online/Internet survey
- Mobil data collection survey
- Mixed mode survey

2.6 Designing a Questionnaire

2.6.1 Basic Problems of a Questionnaire

- Question ordering
- Open and closed questions
- Response options
- Wording of question

2.7 Others Methods of Studies: Observational or Experimental

Definition 2.10. An ***observational study*** observes individuals and measures variables without influencing the responses.

Definition 2.11. An ***experiments*** applies a treatment to the individuals and observes or measures variables to see the effect of the treatment.

In order to observe a cause-and-effect relationship, an experiment is much better than an observational study.

Definition 2.12. A ***confounding variable*** is a variable that both affects the response variable and the explanatory variable. The effect of a confounding variable on the response variable cannot be separated from the effect of the explanatory variable.

Confounding variables are especially a problem in observational studies. Randomized experiments help control the influence of confounding variables.

Definition 2.13. In an experiment, the individuals are called ***subjects***, the explanatory variables are called ***factors***, and a ***treatment*** is a specific combination of values of the factors.

Definition 2.14. A ***randomized experiment*** is one in which the subjects are assigned at random to the different groups.

Randomizing protects us from the influence of all the features of our population by making sure that, on average, the sample looks like the rest of the population.

2.7.1 Observational Study

Definition 2.15. ***Retrospective study*** is an observation study in which subject are selected and then their previous condition or behaviors are determined. It does not need to be based on random sample and focus on estimating differences between groups or associations between variables.

Definition 2.16. ***Prospective study*** is an observation study in which subject are followed to observe future outcomes. It focus on estimation differences among groups that might appear as the groups.

Because no treatment are applied, a prospective study is not an experiment.

3 Simple Random Sampling

3.1 Probability Sampling Designs

3.1.1 Condition

- The set of samples $\{s\}$, that are possible to obtain with the sampling procedure.
- A known probability of selection $p(s)$ is assigned with each possible sample s.
- The procedure gives every element a nonzero probability of inclusion (unbiasedness) π.
- Random mechanism for selection. One sample is selected by the mechanism giving each possible sample exactly the probability $p(s)$.

A sample obtained under conditions above is called a probability sample. Most commonly, we first describe the selection mechanism and then find $p(s)$.

3.1.2 Basic Notation

Suppose N is the population size, i.e., there are N units in the universe or finite population of interest. The N units in the population are denoted by an index set of labels:

$$\varepsilon = \{1, \cdots, N\}.$$

From the population, a sample of n units is to be taken. Let s represent a sample of n units. A probability $p(s)$ is assigned to every possible sample s.

Let y_i be the value of the population characteristic of interest associated with unit i, the population of y values is $\{y_1, \cdots, y_N\}$.

The probability that unit i will be included in a sample is denoted by π_i and is called inclusion probability for unit i.

3.2 Simple Random Sampling

Definition 3.1. ***Simple random sampling*** of size n is the probability sampling design for which a fixed number of n units are selected from a population of N units s.t. every possible sample of n units has equal probability of being selected. A resulting sample is called a simple random sample.

It is the simplest sample design. Each element has an equal probability of being selected from a list of all population units (sample of n from N population). SRS are EPSEM samples, i.e., equal probability of selection method. There are two types of SRS: with replacement (SRSWR) and without replacement (SRS).

For SRSWR:

- One unit of element is randomly selected from population is the first sampled unit.
- Then the sampled unit is replaced in the population.
- The second sample is drawn with equal probability.
- The procedure is repeated until the requisite sample units n are drawn.

- The probability of selection of an element remains unchanged after each draw.
- The same units could be selected more than once.

Unlike SRSWR, once an element is selected as a sample unit, it will not be replaced in the population pool. The selected sample units are distinct.

In practice, SRSWR is not attractive: we do not want to interview same individuals more than once. But in mathematical term it is simpler to relate the sample to population by SRSWR. SRS provides two additional advantages: elements are not repeated and variance estimation is smaller than SRSWR with same sample size.

3.2.1 Combinatorial Notation

- $n!$: The number of unique arrangements or permutations of n distinct items.

$$n! = n(n-1)\cdots 2\cdot 1.$$

- $\binom{N}{n}$: The number of combinations of n items selected from a population of size N.

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

3.2.2 Simple Random Sampling without Replacement

Definition 3.2. Sample of size $n, s = \{e_{i_1}, \cdots, e_{i_n}\}$ is called ***simple random sample*** (SRS) if every sample of size n from the population $\varepsilon = \{e_1, \cdots, e_N\}$ has the same probability of being selected.

There are $\binom{N}{n}$ possible SRSs of size n selected from a population of size N. For any SRS of size n from a population of size N, we have

$$p(s) = \frac{1}{\binom{N}{n}}.$$

Example 3.1. Consider $\varepsilon = \{A, B, C, D\}, N = 4$. List of all possible sample of size $n = 2$.

Solution. There are $\binom{4}{2} = 6$ samples which are $\{A, B\}, \cdots$. Sampling will be simple random, if probability of selection is

$$P(AB) = \cdots = \frac{1}{6}.$$

3.3 How to Draw a SRS

Selecting directly from the list of samples would be very inconvenient even for small populations. Instead of selecting samples directly, we select elements into samples from a list of elements:

- Find/Create a list (frame) of elements.
- Select one element at a time at random.

Example 3.2. $\varepsilon = \{A, B, C, D\}, n = 2$. Selecting $\{A, B\}$ is the same as selecting A first and B second, or B first and A second.

Property 3.1. $\{e_1, \cdots, e_n\}$ can be selected any order ($n!$ different orders).

Proof. We have

$$\begin{aligned} P(\{e_1, \cdots, e_n\}) &= n! P(e_1 \cdots e_n) = n! P(e_1) P(e_2|e_1) \cdots P(e_n|e_1 \cdots e_{n-1}) \\ &= n! \frac{1}{N} \cdot \frac{1}{N-1} \cdots \frac{1}{N-(n-1)} = \frac{1}{\binom{N}{n}}. \end{aligned}$$

□

3.3.1 Table of Random Numbers (TRN)

Definition 3.3. ***Table of random numbers*** is a list of digits produced by an random number generator (RNG) convenient for manual/field/small size problems sampling.

We use TRN to simulate events with given probability:

- Assign certain digits, or groups of digits to the events $A_1, A_2, \cdots$ we want to simulate, depending on $P(A_1), P(A_2), \cdots$.
- Decide how we will read the table, that is, select some digits from the table.
- Read the table, and see which one of the events has occurred.

Example 3.3. A population consists of $N = 8743$ elements/sampling units. Select an SRS of size $n = 8$:

- List of the elements: $a_1, a_2, \cdots, a_{8743}$.
- Number of digits(N) is 4 so we use groups of 4 consecutive digits.
- Assign to every element on group of 4 digits.
- Read the table from left to right, starting with the first row and use first 4 digits out of every group of 5 digits until 8 elements are selected.

In R,

```
N = 8743
n = 8
# SRS without replacement
S1 = sample(N, n, replace = F)
# SRS with replacement
S2 = sample(N, n, replace = T)
```

3.4 Inference

3.4.1 Estimation of the Population Mean

The estimator of μ is

$$\widehat{\mu} = \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{\widehat{\tau}}{N},$$

where $\widehat{\tau}$ is the estimator of τ (population total):

$$\widehat{\tau} = \frac{N}{n} \sum_{i=1}^{n} y_i.$$

Property 3.2. In SRSWR, $\widehat{\mu}$ is unbiased estimator of $\mu : \mathbb{E}[\widehat{\mu}] = \mu$, and $\mathrm{Var}[\widehat{\mu}] = \frac{\sigma^2}{n}$.

Proof. In sampling with replacement, $y_1, \cdots, y_n$ are i.i.d., i.e., $\mathbb{E}[y_i] = \mu, \mathrm{Var}[y_i] = \sigma^2$. Note that, identically distributed means

$$P(y_i = y | y_{i-1}, \cdots, y_1) = P(y_i = y) = \frac{1}{N}, i = 1, \cdots, n.$$

Therefore,

$$\mathbb{E}[\widehat{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\mathbb{E}[y_1 + \cdots + y_n] = \frac{1}{n}(\mathbb{E}[y_1] + \cdots + \mathbb{E}[y_n]) = \frac{1}{n}n\mu = \mu,$$

and

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n^2}\mathrm{Var}[y_1 + \cdots + y_n] = \frac{1}{n^2}(\mathrm{Var}[y_1] + \cdots + \mathrm{Var}[y_n]) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

□

Property 3.3. In SRS (without replacement), $\widehat{\mu}$ is unbiased estimator of $\mu : \mathbb{E}[\widehat{\mu}] = \mu$, and $\mathrm{Var}[\widehat{\mu}] = \frac{(N-n)\sigma^2}{(N-1)n}$.

Proof. In sampling without replacement, $y_1, \cdots, y_n$ are dependent, but identically distributed. For instance,

$$P(y_1 = y) = \frac{1}{N}, P(y_2 | y_1) = \frac{1}{N-1}, y \neq y_1, \cdots .$$

But they have same marginal distribution

$$P(y_i = y) = \frac{1}{N}$$

since for example

$$P(y_2 = y) = P(y_1 \neq y)P(y_2 = y | y_1 \neq y) = \frac{N-1}{N}\frac{1}{N-1} = \frac{1}{N}.$$

We can easily show that $\mathbb{E}[\widehat{\mu}] = \mu$ since y_i's are identically distributed.

We have

$$\mathrm{Var}[\widehat{\mu}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{Var}[y_i] + \sum_{i\neq j}\mathrm{Cov}(y_i, y_j)\right) = \frac{\sigma^2}{n} + \frac{1}{n^2}\sum_{i\neq j}\mathrm{Cov}(y_i, y_j).$$

To find $\mathrm{Cov}(y_i, y_j)$, we first find the joint distribution of $(y_i, y_j) : p(y_i, y_j) = \frac{1}{N(N-1)}$ and so

$$\begin{aligned}
\mathbb{E}[y_i y_j] &= \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j p(y_i, y_j) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \\
&= \frac{1}{N(N-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j - \sum_{i=1}^{N} y_i^2\right) \\
&= \frac{1}{N(N-1)}\left((N\mu)^2 - N\sigma^2 - N\mu^2\right).
\end{aligned}$$

Thus,

$$\mathrm{Cov}(y_i, y_j) = \mathbb{E}[(y_i - \mu)(y_j - \mu)] = \mathbb{E}[y_i y_j] - \mu^2 = -\frac{\sigma^2}{N-1}.$$

Therefore,

$$\mathrm{Var}[\widehat{\mu}] = \frac{\sigma^2}{n} - \frac{(n-1)\sigma^2}{n(N-1)} = \frac{(N-n)\sigma^2}{(N-1)n}.$$

□

3.4.2 Estimation of the Population Variance

We want to estimate

$$\mathrm{Var}[\overline{y}] = \begin{cases} \frac{\sigma^2}{n}, & \text{SRSWR} \\ \frac{(N-n)\sigma^2}{(N-1)n}, & \text{SRS} \end{cases},$$

which are linear functions of σ^2.

Property 3.4. We consider the sample variance $S^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(y_i - \overline{y})^2$. We have

$$\mathbb{E}[S^2] = \begin{cases} \sigma^2, & \text{SRSWR} \\ \frac{N}{N-1}\sigma^2, & \text{SRS} \end{cases}$$

Proof. We have

$$\begin{aligned} \mathbb{E}[S^2] &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(y_i - \overline{y})^2] \\ &= \frac{1}{n-1}\left(\mathbb{E}\left[\sum_{i=1}^{n}(y_i - \mu)^2\right] - \mathbb{E}[n(\overline{y} - \mu)^2]\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}[y_i - \mu]^2 - n\mathbb{E}[(\overline{y} - \mu)^2]\right) \\ &= \frac{1}{n-1}(n\sigma^2 - n\mathrm{Var}[\overline{y}]). \end{aligned}$$

For SRSWR we have

$$\mathbb{E}[S^2] = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2.$$

For SRS, we have

$$\mathbb{E}[S^2] = \frac{1}{n-1}\left(n\sigma^2 - \frac{(N-n)\sigma^2}{N-1}\right) = \frac{N}{N-1}\sigma^2.$$

□

Hence,

$$\widehat{\sigma}^2 = \begin{cases} S^2, & \text{SRSWR} \\ \frac{N-1}{N}S^2, & \text{SRS} \end{cases},$$

which is unbiased, and thus the unbiased estimate of $\mathrm{Var}[\overline{y}]$ is

$$\widehat{\mathrm{Var}}[\overline{y}] = \begin{cases} \frac{S^2}{n}, & \text{SRSWR} \\ \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, & \text{SRS} \end{cases}.$$

3.4.3 Estimation of the Population Standard Deviation

A general note on estimation: If $\eta = g(\theta)$ and $\widehat{\theta}$ is an unbiased estimator of θ, then $\widehat{\eta} = g(\widehat{\theta})$ is not, in general an unbiased estimator of η, except when $g(\theta) = a\theta + b$, a linear function of θ. In this case,

$$\mathbb{E}[\widehat{\eta}] = \mathbb{E}[g(\widehat{\theta})] = \mathbb{E}[a\widehat{\theta} + b] = a\mathbb{E}[\widehat{\theta}] + b = a\theta + b.$$

We have $\sigma = \sqrt{\sigma^2} = g(\sigma)$, and $\widehat{\sigma} = g(\widehat{\sigma}^2) = \sqrt{\widetilde{S}^2} = \widetilde{S}$. $\widehat{\sigma} = \widetilde{S}$ is a biased estimator of σ but the bias is usually small.

The estimator of the standard deviation $\widehat{\sigma}_{\overline{y}}$ is

$$\widehat{\sigma}_{\overline{y}} = \widehat{\mathrm{SD}}(\overline{y}) = \sqrt{\widehat{\mathrm{Var}}[\overline{y}]} = \begin{cases} \frac{S}{\sqrt{n}}, & \text{SRSWR} \\ \sqrt{\left(1 - \frac{n}{N}\right)} \frac{S}{\sqrt{n}}, & \text{SRS} \end{cases} \sim \frac{S}{\sqrt{n}}.$$

Note that $\widehat{\sigma}_{\overline{y}}$ is biased and for large $N, \widehat{\sigma}_{\overline{y}} \sim \frac{S}{\sqrt{n}}$.

In SRS without replacement, we have

$$\begin{aligned} \widehat{\mu} = \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ \mathrm{Var}[\overline{y}] &= \frac{N-n}{N-1}\frac{\sigma^2}{n} \\ \widehat{\mathrm{Var}}[\overline{y}] &= \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \\ B = 2\widehat{\sigma}_{\overline{y}} &= 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{S^2}{n}}, \end{aligned}$$

where B is the estimated bound on the error of estimation.

Note that taking a square root of $\mathrm{Var}[\overline{y}]$ yields the ***standard deviation*** of the estimator; taking a square root of an estimated variance $\widehat{\mathrm{Var}[\overline{y}]}$ yields the ***standard error*** of the estimator.

3.4.4 Finite Population Correction

The term $1 - \frac{n}{N}$ in the estimated variance of $\overline{y}$ is called the ***finite population correction*** (***f.p.c.***).

If N is large (e.g., $N > 20n$), then f.p.c. can be ignored. In this case, the estimated variance is the more familiar quantity $\frac{S^2}{n}$. But since the f.p.c. will be less than 1, omitting the f.p.c. from the estimated variance formulas (i.e., replacing $1 - \frac{n}{N}$ by 1) will slightly overestimate the true variance - there is a small positive bias.

If N is not large relative to n (e.g., $N < 20n$,) then omitting the f.p.c. from the estimated variance formulas can seriously overestimate the true variance - there can be a large positive bias.

If N is known, we would rather not omit the f.p.c.. In many cases, the population size N is not clearly defined or is unknown and thus N can be assumed to be quite large and hence the f.p.c. can be ignored.

3.4.5 Estimation of the Population Total

From $\tau = N\mu$, we have

$$\text{Estimator of the population total } \tau : \widehat{\tau} = N\widehat{\mu} = N\overline{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$$

$$\text{Variance of } \tau : \text{Var}[\widehat{\tau}] = \text{Var}[N\overline{y}] = N^2\text{Var}[\overline{y}] = \frac{N^2(N-n)}{N-1}\frac{\sigma^2}{n}$$

$$\text{Estimated Variance for } \tau \text{ (Unbiased)} : \widehat{\text{Var}}[\widehat{\tau}] = \widehat{\text{Var}}[N\overline{y}] = N^2\widehat{\text{Var}}[\overline{y}] = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

$$\text{Estimated SD (Biased)} : \widehat{\sigma}_{\widehat{\tau}} = \widehat{\text{SD}}(\widehat{\tau}) = \sqrt{\widehat{\text{Var}}[\widehat{\tau}]} = N\sqrt{\widehat{\text{Var}}[\overline{y}]} = N\widehat{\text{SD}}(\overline{y})$$

$$\text{Estimated bound on the error of estimation} : B = 2\widehat{\sigma}_{\widehat{\tau}} = 2\sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}}$$

3.4.6 Confidence Intervals of μ and τ

The $100(1-\alpha)\%$ error of bound is given by

$$B = Z_{\frac{\alpha}{2}} \times \sigma.$$

For $\alpha = 5\%$, the $Z_{\frac{\alpha}{2}} = 1.96 \approx 2$ and thus the 95% bound on the error when estimating the population mean is

$$B_\mu \approx 2\sqrt{\text{Var}[\widehat{\mu}]}$$

and the 95% bound on the error when estimating the population total is

$$B_\tau \approx 2\sqrt{\text{Var}[\widehat{\tau}]}.$$

Note that the bound on the error of estimation is also called marginal of error.

For large samples, $100(1-\alpha)\%$ confidence interval for μ is $[\overline{y} \mp B_\mu]$ and for τ is $[\widehat{\tau} \mp B_\tau]$.

3.4.7 Estimation of the Population Proportion p Using SRS

Proportion of elements which poses certain property or belong to certain specified group. Define

$$y(e) = \begin{cases} 1, & e \text{ has the property} \\ 0, & e \text{ does not have the property} \end{cases}.$$

The population proportion is defined by

$$p = \frac{1}{N}\sum_{i=1}^{N} y(e_i) = \frac{M}{N} = \frac{\text{Number of elements with property}}{N}.$$

The population mean for y is $\mu_y = \mu = 0(1-p) + 1p = p$ and the population total is $\tau_y = \tau = N\mu = Np = M$.

Note that $y_i = y_i^2 = 0$ or 1, then

$$p = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{i=1}^{N} y_i^2}{N} \Rightarrow \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i^2 = Np$$

and thus

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i-p)^2 = \frac{1}{N}\left(\sum_{i=1}^{N}y_i^2 - Np^2\right) = \frac{1}{N}(Np-Np^2) = p(1-p).$$

Hence,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\widehat{p})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2 - n\widehat{p}^2\right) = \frac{1}{n-1}(n\widehat{p}-n\widehat{p}^2) = \frac{n}{n-1}\widehat{p}(1-\widehat{p}).$$

In SRS without replacement, we have

$$\text{Estimator of } p \text{ (Unbiased)} : \widehat{p} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\text{Variance of } \widehat{p} : \text{Var}[\widehat{p}] = \frac{N-n}{N-1}\frac{pq}{n}, q = 1-p$$

$$\text{Estimated variance of } \widehat{p} : \widehat{\text{Var}}[\widehat{p}] = \left(1-\frac{n}{N}\right)\frac{\widehat{p}\widehat{q}}{n-1}, \widehat{q} = 1-\widehat{p}$$

$$\text{Estimated bound on the error of estimation} : B = 2\widehat{\sigma}_{\widehat{p}} = 2\widehat{\text{SD}}(\widehat{p}) = 2\sqrt{\left(1-\frac{n}{N}\right)\frac{\widehat{p}\widehat{q}}{n-1}}, \widehat{q} = 1-\widehat{p}$$

3.5 Selecting Sample Size for Estimating μ, τ and p Using SRS

The number of observations needed to estimate a population mean μ with a bound on the error of estimation B is found by setting 2 × SD of the estimator $\mu = \overline{y}$, equal to B :

$$2 \times \text{SD}(\widehat{\mu}) = 2\sqrt{\text{Var}[\widehat{\mu}]} = 2\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} = B.$$

The required sample size can be found by solving equation for n and thus

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2}, D = \frac{B^2}{4}.$$

Similarly, the sample size required to estimate τ with a bound on the error B is

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2}, D = \frac{B^2}{4N^2}.$$

Note that $\sigma_y^2 = \sigma^2$ is unknown in practice and we can use S^2 to estimate, usually obtain by using one of the following three methods: (1) use a previous study, experience, educated guess; (2) use a presample (usually more complex studies); (3) quick and dirty method - use the range of variable $y, \widehat{\sigma} = \frac{\text{Range}}{4}$. If N is large (N unknown), then $N-1$ can be replaced by N in the denominator.

Sample size required to estimate p with a bound on the error estimation B is

$$n = \frac{Npq}{(N-1)D+pq}, q = 1-p \text{ and } D = \frac{B^2}{4}.$$

Note that p is unknown in practice and we can estimate from similar past surveys. If no past information is available, we can substitute $p = 0.5$ to obtain a conservative sample size (one that is likely to be larger than required).

Example 3.4. We want to conduct a survey to determine the proportion of students who favor a proposed honor code. Because interviewing $N = 2000$ student in a reasonable length of time is impossible, determine the sample size needed to estimate p with a bound on the error of estimation $B = 0.05$. Assuming that no prior information is available to estimate p.

Solution. We have

$$D = \frac{B^2}{4} = \frac{0.05^2}{4} = 0.000625$$

and

$$n = \frac{Npq}{(N-1)D + pq} = \frac{2000 \times 0.5 \times 0.5}{(2000-1) \times 0.000625 + 0.5 \times 0.5} = 333.56.$$

That is 334 students must be interviewed to estimate p with $B = 0.05$.

4 Ratio, Regression, and Difference Estimation

The estimation of the population mean and total are based on a sample of response measurements $y_1, \cdots, y_n$ obtained by SRS. Sometimes, other variables are closely related to the response y. By measuring y and one or more subsidiary (auxiliary) variable, we can obtain additional information for estimating the population parameter. We can use ratio, regression, and difference estimation.

4.1 Ratio Method Using SRS

4.1.1 Ratio Estimation of the Population Ratio $R = \frac{\mu_y}{\mu_x}$

Suppose that a SRS of size n is to be drawn from a fine population containing N elements. The estimation of R is

$$r = \frac{\overline{y}}{\overline{x}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

and the estimated variance of r is

$$\widehat{\text{Var}}[r] = \left(1 - \frac{n}{N}\right) \frac{S_r^2}{\mu_x^2 n},$$

where μ_x is the population mean for the r.v. X and

$$S_r^2 = \frac{1}{n-1} \sum_{i=1}^n e_i^2,$$

$e_i = y_i - rx_i$ is called residuals. Note that if μ_x is unknown, we use $\overline{x}^2$ to approximate μ_x^2 in the estimated variance.

4.1.2 Ratio Estimation of the Population Total τ_y

The ratio estimator of τ_y is

$$\widehat{\tau}_y = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}(\tau_x) = r\tau_x$$

and the estimated variance of $\widehat{\tau}_y$ is

$$\widehat{\text{Var}}[\widehat{\tau}_y] = \tau_x^2 \widehat{\text{Var}}[r] = (N\mu_x)^2 \left(1 - \frac{n}{N}\right) \frac{S_r^2}{\mu_x^2 n} = N^2 \left(1 - \frac{n}{N}\right) \frac{S_r^2}{n},$$

where μ_x and τ_x are the population mean and total for the r.v. X.

4.1.3 Ratio Estimation of the Population Mean

The ratio estimator of μ_y is

$$\widehat{\mu}_y = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}(\mu_x) = r\mu_x$$

and the estimated variance of $\widehat{\mu}_y$ is

$$\widehat{\text{Var}}[\widehat{\mu}_y] = \mu_x^2 \widehat{\text{Var}}[r] = \mu_x^2 \left(1 - \frac{n}{N}\right) \frac{S_r^2}{\mu_x^2 n} = \left(1 - \frac{n}{N}\right) \frac{S_r^2}{n}.$$

4.1.4 Sample Size Required for Estimating R

We set

$$2\sqrt{\text{Var}[r]} = 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{\mu_x^2 n}} = B.$$

Solve for n we have

$$n = \frac{N\sigma^2}{ND + \sigma^2}, D = \frac{B^2\mu_x^2}{4}.$$

In practice, we do not know σ^2. If no past information is available to calculate S_r^2 as an estimate of σ^2, we take a preliminary sample of size n' and compute

$$\widehat{\sigma}^2 = \frac{1}{n' - 1}\sum_{i=1}^{n'} e_i^2,$$

where $e_i = y_i - rx_i$. If μ_x is unknown, it can be replaced by $\overline{x}$, calculated from the n' preliminary observations.

4.1.5 Sample Size Required for Estimating μ_y

We have

$$n = \frac{N\sigma^2}{ND + \sigma^2}, D = \frac{B^2}{4}.$$

If no past information is available to calculate S_r^2 as an estimate of σ^2, we take a preliminary sample of size n' as before. Note that we do not need to know μ_x.

4.1.6 Sample Size Required for Estimating τ_y

We have

$$n = \frac{N\sigma^2}{ND + \sigma^2}, D = \frac{B^2}{4N^2}.$$

If no past information is available to calculate S_r^2 as an estimate of σ^2, we take a preliminary sample of size n' as before. Note that we do not need to know μ_x.

4.2 Regression Estimation Using SRS

Suppose there is evidence of linear relationship between observed y and x and we assume a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \cdots, n.$$

The estimated regression line is

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

where

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \text{ and } \widehat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

Regression line can be written as

$$\widehat{y}_i = \overline{y} + \widehat{\beta}_1(x_i - \overline{x}).$$

The predicted value of e_i is

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - (\overline{y} + \widehat{\beta}_1(x_i - \overline{x})).$$

4.2.1 Regression Estimation of the Population Mean

Regression estimator of μ_y is the predicted value of y when $x = \mu_x$:

$$\widehat{\mu}_{yL} = \overline{y} + \widehat{\beta}_1(\mu_x - \overline{x})$$

and the estimated variance of $\widehat{\mu}_{yL}$ is

$$\widehat{\text{Var}}[\widehat{\mu}_{yL}] = \left(1 - \frac{n}{N}\right) \frac{\text{MSE}}{n},$$

where MSE is the estimated mean square error defined as

$$\text{MSE} = \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}.$$

4.3 Difference Estimation Using SRS

The difference method of estimating a population mean or total is similar to the regression method in that it adjusts the $\overline{y}$ value up or down by an amount depending on the difference $(\mu_x - \overline{x})$. But in difference method, the regression coefficient $\widehat{\beta}_1$ is not computed, which is set equal to unity.

The difference method is then easier to employ than the regression method. Also, difference method frequently works well when the x values are highly correlated with the y values and both are measured on the same scale.

4.3.1 Difference Estimation of the Population Mean

Difference estimator of μ_y is

$$\widehat{\mu}_{yD} = \overline{y} + (\mu_x - \overline{x}) = \mu_x + \overline{d}, \overline{d} = \overline{y} - \overline{x}$$

and the estimated variance of $\widehat{\mu}_{yD}$ is

$$\widehat{\text{Var}}[\widehat{\mu}_{yD}] = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1},$$

where $d_i = y_i - x_i, i = 1, \cdots, n$.

5 Stratified Random Sampling

Definition 5.1. A ***stratified random sample*** is one obtained by separating the population elements into non-overlapping groups, called ***strata***, and then selecting a simple random sample from each stratum.

Note that we use stratified RS as stratified random sample and SRS means simple random sample without replacement.

The principle reasons for using stratifies RS rather than SRS are:

- Stratification may produce a smaller bound on the error of estimation than would be produced by a SRS of the same size. This result is particularly true if measurements within strata are homogeneous.
- The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.
- Estimates of population parameters may be desired for subgroups of the population. These subgroups should then be identifiable strata.

5.1 How to Draw a Stratified RS

First, specify the strata, the each sampling unit of the population is paced into its appropriate stratum. Then, we select a SRS from each stratum by using techniques given in SRS. We must be certain that the samples selected from the strata are independent, i.e., observations chosen in an stratum do not depend on those chosen in another.

Two basic rules of stratified sampling:

- A minimum of two-elements must be chosen from each stratum so that sampling errors can be estimated for all strata independently.
- The population should be homogeneous within stratum, and the population should be heterogeneous between the strata.

5.2 Estimation

Let L be the number of strata, N_i be the number of sampling units in stratum i, $N = N_1 + \cdots + N_L$ be the number of sampling units in the population, μ_i be the population mean for stratum i, τ_i be the population total for stratum i, $\tau = \tau_1 + \cdots + \tau_L$ be the population total, n_i be the sample size for stratum i, and $\overline{y}_i$ be the sample mean for the SRS selected for stratum i.

We have a SRS within each stratum. Therefore, we know

- $\widehat{\mu}_i = \overline{y}_i$ is unbiased estimator of μ_i.
- $\widehat{\tau}_i = N_i\overline{y}_i$ is unbiased estimator for the stratum total τ_i.

We denote the population mean estimator by $\overline{y}_{\text{st}}$ or $\widehat{\mu}_{\text{st}}$, and the population total estimator by $N\overline{y}_{\text{st}}$ or $\widehat{\tau}_{\text{st}}$.

5.2.1 Estimation of the Population Mean

Note that strata means are not additive. The population mean for strata is $\mu_i = \frac{\tau_i}{N_i}$ but

$$\mu = \tau_N = \frac{\tau_1 + \cdots + \tau_L}{N_1 + \cdots + N_L} \neq \frac{\tau_1}{N_1} + \cdots + \frac{\tau_L}{N_L} = \mu_1 + \cdots + \mu_L.$$

Population mean is weighted average, i.e.,

$$\mu = W_1\mu_1 + \cdots + W_L\mu_L = \sum_{i=1}^{L} W_i\mu_i$$

where $W_i = \frac{N_i}{N}$ are weights.

The estimator of μ is

$$\widehat{\mu}_{\text{st}} = \overline{y}_{\text{st}} = W_1\overline{y}_1 + \cdots + W_L\overline{y}_L = \sum_{i=1}^{L} W_i\overline{y}_i.$$

The estimated variance of $\widehat{\mu}_{\text{st}}$ is

$$\widehat{\text{Var}}[\widehat{\mu}_{\text{st}}] = \sum_{i=1}^{L} W_i^2 \widehat{\text{Var}}[\widehat{\mu}_i] = \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right).$$

5.2.2 Estimation of the Population Total

Note that strata totals are additive. The population total for strata i is $\tau_i = N_i\mu_i$ and

$$\tau = \tau_1 + \cdots + \tau_L = N_1\mu_1 + \cdots + N_L\mu_L.$$

The estimator of τ is

$$\widehat{\tau}_{\text{st}} = N\overline{y}_{\text{st}} = N_1\overline{y}_1 + \cdots + N_L\overline{y}_L = \sum_{i=1}^{L} N_i\overline{y}_i.$$

The estimated variance of $\widehat{\tau}_{\text{st}}$ is

$$\begin{aligned}\widehat{\text{Var}}[\widehat{\tau}_{\text{st}}] &= N_1^2\left(1 - \frac{n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + \cdots + N_L^2\left(1 - \frac{n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right) \\ &= \sum_{i=1}^{L} \widehat{\text{Var}}[\widehat{\tau}_i] = \sum_{i=1}^{L} N_i^2 \widehat{\text{Var}}[\widehat{\mu}_i] = \sum_{i=1}^{L} N_i^2\left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right).\end{aligned}$$

5.2.3 Estimation of the Population Proportion

Note that strata proportion are not additive. The population proportion for strata i is $p_i = \frac{\sum_{j=1}^{N} y_{ij}}{N_i}$ but

$$p = \frac{N_1}{N}p_1 + \cdots + \frac{N_L}{N}p_L = W_1p_1 + \cdots + W_Lp_L \neq p_1 + \cdots + p_L$$

where $W_i = \frac{N_i}{N}$ are called weights and $\sum\limits_{i=1}^{L} W_i = 1$. $\widehat{p}_{\text{st}}$ is

$$\widehat{\text{Var}}[\widehat{p}_{\text{st}}] = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \widehat{\text{Var}}[\widehat{p}_i] = \sum_{i=1}^{L} W_i^2\left(1 - \frac{n_i}{N_i}\right)\left(\frac{\widehat{p}_i(1-\widehat{p}_i)}{n_i - 1}\right).$$

5.3 Sample Size

5.3.1 Sample Size Estimation

Sample size estimation depends on variance estimation. Consider the variance of a mean for a variable y :

$$\widehat{\text{Var}}[\widehat{\mu}_{\text{st}}] = \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right).$$

Given bound B, we want

$$2\sqrt{\widehat{\text{Var}}[\widehat{\mu}_{\text{st}}]} = \frac{B^2}{4} \Rightarrow \widehat{\text{Var}}[\widehat{\mu}_{\text{st}}] = \frac{B^2}{4}.$$

But we still cannot solve equation for n unless we know the relationship $n_1, \cdots, n_L$ and n.

5.3.2 Allocation of Sample Size

There are many ways to divide the sample size n into the individual stratum sample sizes $n_1, \cdots, n_L$. Each division may result in a different variance for the sample mean. Hence, the objective is to use an allocation that gives a specified amount of information at minimum cost.

In each method, the number of observations n_i allocated to the ith stratum is some fraction of the total sample size n. We denote the allocation fraction by a_i and thus we can write

$$n_i = na_i, i = 1, \cdots, L.$$

5.3.2.1 Equal Allocation

Divide the number of sample units n equally among the L strata:

$$a_i = \frac{1}{L}, n_i = na_i = \frac{n}{L}.$$

Example 5.1. Suppose $n = 100, L = 4$ strata. We sample $n_i = \frac{100}{4} = 25$ in each stratum.

5.3.2.2 Proportion Allocation

The proportion allocation scheme is affected by only one factor:

- The total number of elements in each stratum.

We assume that the variability of observations within each stratum is the same for all strata: $\sigma_1 = \cdots = \sigma_L$ and the cost of obtaining an observation is the same for all strata: $c_1 = \cdots = c_L$. Under the method, the allocation fraction is

$$a_i = \frac{N_i}{\sum\limits_{k=1}^{L} N_k}.$$

5.3.2.3 Neyman Allocation

The Neyman allocation scheme is affected by two factors:

- The total number of elements in each stratum.

- The variability of observations within each stratum.

The cost of obtaining an observation is the same for all strata, $c_1 = \cdots = c_L$. Under the method, the allocation fraction is

$$a_i = \frac{N_i\sigma_i}{\sum\limits_{k=1}^{L} N_k\sigma_k}$$

where σ_i^2 denotes the populations variance for the $i$the stratum.

5.3.2.4 Optimal Allocation

The best allocation scheme is affected by three factors:

- The total number of elements in each stratum.
- The variability of observations within each stratum.
- The cost of obtaining an observation from each stratum.

Approximate allocation that minimizes cost for fixed value of $\text{Var}[\widehat{\mu}_{\text{st}}]$ or minimizes $\text{Var}[\widehat{\mu}_{\text{st}}]$ for a fixed cost is

$$a_i = \frac{N_i\sigma_i/\sqrt{c_i}}{N_1\sigma_1/\sqrt{c_1} + \cdots + N_L\sigma_L/\sqrt{c_L}} = \frac{N_i\sigma_i/\sqrt{c_i}}{\sum\limits_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}$$

where σ_i^2 denotes the populations variance for the ith stratum, and c_i denotes the cost of obtaining a single observation from the ith stratum. Note that n_i is directly proportional to N_i and σ_i and inversely proportional to $\sqrt{c_i}$.

5.3.3 Sample Size Required for Estimating μ and τ

Sample size required to estimate μ or τ using stratified RS with a bound on the error of estimation B is

$$n = \frac{\sum\limits_{i=1}^{L} N_i^2\sigma_i^2/a_i}{N^2 D + \sum\limits_{i=1}^{L} N_i\sigma_i^2}$$

where a_i is the fraction of observations allocated to stratum i and

$$D = \begin{cases} \frac{B^2}{4}, & \text{for estimating } \mu \\ \frac{B^2}{4N^2}, & \text{for estimating } \tau \end{cases}.$$

For optimal allocation: Substituting the $\frac{n_i}{n}$ given by the optimal allocation formula for a_i gives:

$$n = \frac{\left(\sum\limits_{i=1}^{L} N_k\sigma_k/\sqrt{c_k}\right)\left(\sum\limits_{i=1}^{L} N_i\sigma_i\sqrt{c_i}\right)}{N^2 D + \sum\limits_{i=1}^{L} N_i\sigma_i^2}.$$

5.3.4 Sample Size Required for Estimating p

Sample size required to estimate p using stratified RS with a bound on the error of estimation B is

$$n = \frac{\sum\limits_{i=1}^{L} N_i^2 p_i q_i / a_i}{N^2 D + \sum\limits_{i=1}^{L} N_i p_i q_i}$$

where a_i is the fraction of observations allocated to stratum i, $D = \frac{B^2}{4}$, and $q_i = 1 - p_i$.

Example 5.2. The fraction p_1, p_2 and p_3 are unknown and can be approximated by the estimates from the earlier study that $\widehat{p}_1 = 0.8, \widehat{p}_2 = 0.25$ and $\widehat{p}_3 = 0.5$. The cost of obtaining an observation is \$9 for either town A and town B and \$16 for the rural area. The number of households within the strata are $N_1 = 155, N_2 = 62$ and $N_3 = 93$. The firm wants to estimate the proportion p with a bound on the error of estimation equal to 0.1. Find the sample size n and the strata sample sizes n_1, n_2 and n_3.

Solution. First we find a_1, a_2 and a_3. We have

$$\sum_{i=1}^{3} N_i \sqrt{\frac{p_i q_i}{c_i}} \approx 41.241.$$

Thus,

$$a_1 = \frac{N_1 \sqrt{\frac{p_1 q_1}{c_1}}}{41.241} \approx 0.501$$
$$a_2 = \frac{N_2 \sqrt{\frac{p_2 q_2}{c_2}}}{41.241} \approx 0.216$$
$$a_3 = \frac{N_3 \sqrt{\frac{p_3 q_3}{c_3}}}{41.241} \approx 0.282$$

Then we find the sample size needed

$$n = \frac{\sum\limits_{i=1}^{3} N_i^2 p_i q_i / a_i}{N^2 D + \sum\limits_{i=1}^{3} N_i^2 p_i q_i} \approx 63.$$

Hence, $n_1 = na_1 \approx 32, n_2 = na_2 \approx 14, n_3 = na_3 \approx 18$.

5.4 Ration Estimation

Suppose we have two variables, y is the variable of interest and x is auxiliary variable. We assume that we can take a large enough sample of both y and x. There are two different ways to produce estimates: separate ratio estimator and combined ratio estimator.

5.4.1 Separate Ratio (SR) Estimator

5.4.1.1 SR for Estimating R

Estimator of ratio $R_i = \frac{\mu_{y,i}}{\mu_{x,i}}$ for ith stratum is given by

$$\widehat{R}_i = \frac{\overline{y}_i}{\overline{x}_i}.$$

The estimated variance of $\widehat{R}_i$ is

$$\widehat{\text{Var}}[\widehat{R}_i] = \left(1 - \frac{n_i}{N_i}\right)\left(\frac{1}{\mu_{x,i}^2}\right)\frac{s_{R_i}^2}{n_i}$$

where

$$s_{R_i}^2 = \frac{\sum\limits_{j=1}^{n_i}(y_{ij} - \widehat{R}_i x_{ij})^2}{n_i - 1}.$$

The SR estimator of the population ratio $R = \frac{\mu_y}{\mu_x}$ is a weighted average of theses separate estimates:

$$\widehat{R}_{\text{SR}} = \sum_{i=1}^{L} W_i \widehat{R}_i$$

where $W_i = \frac{N_i}{N}$.

The estimated variance is

$$\widehat{\text{Var}}[\widehat{R}_{\text{SR}}] = \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{1}{\mu_{x,i}^2}\right)\frac{s_{R_i}^2}{n_i}.$$

If $\mu_{x,i}^2$ is unknown then it can be replaced by $\overline{x}_i^2$.

5.4.1.2 SR for Estimating μ_y

Estimator for $\mu_{y,i} = R_i \mu_{x,i}$ for the ith stratum is given by

$$\widehat{\mu}_{y,i} = \frac{\overline{y}_i}{\overline{x}_i}\mu_{x,i}.$$

The estimated variance is

$$\widehat{\text{Var}}[\widehat{\mu}_{y,i}] = \left(1 - \frac{n_i}{N_i}\right)\frac{s_{R_i}^2}{n_i}$$

where

$$s_{R_i}^2 = \frac{\sum\limits_{j=1}^{n_i}(y_{ij} - \widehat{R}_i x_{ij})^2}{n_i - 1}.$$

The SR estimator of $\mu_y = R\mu_x$ is a weighted average of these separate estimates:

$$\widehat{\mu}_{y,\text{SR}} = \sum_{i=1}^{L} W_i \widehat{\mu}_{y,i} = \sum_{i=1}^{L} W_i \frac{\overline{y}_i}{\overline{x}_i}\mu_{x,i}$$

where $W_i = \frac{N_i}{N}$.

The estimated variance is

$$\widehat{\text{Var}}[\widehat{\mu}_{y,\text{SR}}] = \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right)\frac{s_{R_i}^2}{n_i}.$$

5.4.2 Combined Ratio (CR) Estimator

5.4.2.1 CR for Estimating R

The method involves first estimating μ_y by the usual $\overline{y}_{\text{st}}$ and similarly estimating μ_x by $\overline{x}_{\text{st}}$. The CR estimator of R is

$$\widehat{R}_{\text{CR}} = \frac{\overline{y}_{\text{st}}}{\overline{x}_{\text{st}}}.$$

Estimated variance is

$$\widehat{\text{Var}}[\widehat{R}_{\text{CR}}] = \left(\frac{1}{\mu_x^2}\right) \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s^2_{R_{\text{CR},i}}}{n_i}$$

where

$$s^2_{R_{\text{CR},i}} = \frac{\sum_{j=1}^{n_i} (y_{ij} - \widehat{R}_{\text{CR}} x_{ij})^2}{n_i - 1}.$$

If μ_x^2 is unknown then it can be replaced by $\overline{x}^2_{\text{st}}$.

5.4.2.2 CR for Estimating μ_y

CR estimator of population mean μ_y is

$$\widehat{\mu}_{y,\text{CR}} = \widehat{R}_{\text{CR}} \mu_x = \frac{\overline{y}_{\text{st}}}{\overline{x}_{\text{st}}} \mu_x.$$

Estimated variance is

$$\widehat{\text{Var}}[\widehat{\mu}_{y,\text{CR}}] = \sum_{i=1}^{L} W_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s^2_{R_{\text{CR},i}}}{n_i}$$

where

$$s^2_{R_{\text{CR},i}} = \frac{\sum_{j=1}^{n_i} (y_{ij} - \widehat{R}_{\text{CR}} x_{ij})^2}{n_i - 1}.$$

Generally, the concern with the SR estimator is that with small sample sizes per stratum the individual stratum variance estimates will be biased and that bias is added across strata and thus it is recommended to use the SR estimator unless the stratum sizes are small ($n_i < 20$) or if the within-stratum ratios are approximately equal.

5.4.2.3 SR and CR for Estimating τ

$$\widehat{\tau}_{\text{SR}} = N \widehat{\mu}_{\text{SR}}$$

and

$$\widehat{\tau}_{\text{CR}} = N \widehat{\mu}_{\text{CR}}.$$

The variances of the estimators can be adjusted accordingly.

6 Systematic Sampling

Definition 6.1. ***1-in-k systematic sampling*** is a sampling method in which one of the first k members of the population is selected at random, then beginning with that member, every kth member is selected.

Example 6.1. Suppose $k = 10$. One of the first 10 members is selected at random (using `sample(10, size=1)`). Then beginning with that member, every 10th member is selected.

In general, 1-in-k systematic sampling involves random selection of one element from the first k elements and the selection of every kth element thereafter.

Systematic sampling provides a useful alternative to SRS for the following reasons:

- Systematic sampling is easier to perform in the field and hence is less subject to selection errors by fieldworkers than are either simple random sample or stratified random samples, especially if a good frame is not available.
- Systematic sampling can provide greater information per unit cost than simple random sampling can for populations with certain patterns in the arrangement of elements.

6.1 How to Draw a Systematic Sample

In general, for a systematic sample of n elements from a population of size N, we must have $k \leqslant \frac{N}{n}$. We select from an ordered sampling frame: Select first element/value (y_1) at random out of the first k elements, and then every kth element (with step k) from the list, $y_2, \cdots, y_n$.

The obtained sample is random, but it is not a simple random sample: If $k > 1$, two consecutive elements cannot appear in the sample.

Example 6.2. Suppose we want to sample rainfall at a site every 50th day over a year. Thus $N = 365$ and we wish to select a 1 in k systematic sample, with $k = 50$. The code is:

```
N = 365
k = 50
start = sample(1:k, 1)
s = seq(start, N, k)
```

Example 6.3. Suppose we want to draw a systematic sample of size $n = 4$ from a population of $N = 13$. We need to use `floor()` to calculate the appropriate value for the selection step $k = \frac{13}{4}$. The code is:

```
N = 13
n = 4
k = floor(N/n)
start = sample(1:k, 1)
s = seq(start, N, k)
```

6.2 Estimation

6.2.1 Estimation of μ and τ

We have

$$\widehat{\mu}_{\text{sys}} = \overline{y}_{\text{sys}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$\widehat{\tau}_{\text{sys}} = N\overline{y}_{\text{sys}} = \frac{N}{n}\sum_{i=1}^{n} y_i.$$

Now we suppose, in a population of size $N = kn$, there are k possibles systematic sample of size n. Population of k systematic samples is:

Systematic Sample Number	Sample number 1	2	3	$\cdots$	n	Sample Mean
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1n}	$\overline{y}_1 = \frac{1}{n}\sum\limits_{j=1}^{n} y_{1j}$
2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2n}	$\overline{y}_2 = \frac{1}{n}\sum\limits_{j=1}^{n} y_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	y_{k3}	$\cdots$	y_{kn}	$\overline{y}_k = \frac{1}{n}\sum\limits_{j=1}^{n} y_{kj}$

We have two approaches to compute $\text{Var}[\widehat{\mu}_{\text{sys}}]$: sampling distribution and population ANOVA table.

6.2.1.1 Sampling Distribution

For each systematic sample $i : \overline{y}_i = \frac{1}{n}\sum\limits_{j=1}^{n} y_{ij}$. Possible values of $\widehat{\mu}_{\text{sys}}$ are $\overline{y}_1, \cdots, \overline{y}_k$. Probability that $\widehat{\mu}_{\text{sys}}$ takes each value is $P(\widehat{\mu}_{\text{sys}} = \overline{y}_i) = \frac{1}{k}$.

We have

$$\begin{aligned}\mathbb{E}[\widehat{\mu}_{\text{sys}}] &= \sum_{i=1}^{k} \overline{y}_i P(\widehat{\mu}_{\text{sys}} = \overline{y}_i) = \frac{1}{k}\sum_{i=1}^{k} \overline{y}_i = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n}\sum_{j=1}^{n} y_{ij} \\ &= \frac{1}{nk}\sum_{k=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{N}\sum_{i=1}^{k} y_i = \mu.\end{aligned}$$

Thus $\widehat{\mu}_{\text{sys}}$ is unbiased estimator for μ.

The variance of $\widehat{\mu}_{\text{sys}}$ is

$$\text{Var}[\widehat{\mu}_{\text{sys}}] = \frac{1}{k}\sum_{i=1}^{k} (\overline{y}_i - \mu)^2.$$

From the table above, we know that selecting 1-in-k systematic sample is equivalent to select a SRS of $n = 1$ element from the population of the k means $\overline{y}_1, \cdots, \overline{y}_k$. Hence we have the sample mean is

$$\widehat{\mu} = \sum_{i=1}^{n} \overline{y}_i = \sum_{i=1}^{1} \overline{y}_i = \overline{y} = \overline{y}_{\text{sys}}.$$

The variance is

$$\text{Var}[\overline{y}_{\text{sys}}] = \frac{k-n}{k-1}\frac{\sigma_{\overline{y}}^2}{n} = \sigma_{\overline{y}}^2 = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \mu)^2.$$

6.2.1.2 Population ANOVA

Suppose population of size $N = kn$ with k possible systematic samples, each of size n. The ith sample mean is $\overline{y}_i = \frac{1}{n}\sum\limits_{j=1}^{n} y_{ij}$. The overall mean per element is $\overline{\overline{y}} = \frac{1}{k}\sum\limits_{i=1}^{k}\overline{y}_i = \frac{1}{k}\sum\limits_{i=1}^{kn}\frac{1}{n}y_{ij}$.

Based on a population of measurement, we can make ANOVA table. We first define the following sum squares:

- ***Between-Cluster Sum Square***:

  $$\text{SSB} = n\sum_{i=1}^{k}(\overline{y}_i - \overline{\overline{y}})^2.$$

  The df of SSB is $k-1$.

- ***Within-Cluster Sum Square***:

  $$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2.$$

  The df of SSW is $k(n-1)$.

- ***Total Sum Square***:

  $$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{\overline{y}})^2.$$

- ***Mean Between-Cluster Sum Square***:

  $$\text{MSB} = \frac{\text{SSB}}{k-1}.$$

- ***Mean Within-Cluster Sum Square***:

  $$\text{MSW} = \frac{\text{SSW}}{k(n-1)}.$$

The variance of the estimated mean can also be written as

$$\text{Var}[\widehat{\mu}_{\text{sys}}] = \frac{\sigma^2}{n}[1 + (n-1)\rho]$$

where σ^2 is the population variance, given by

$$\sigma^2 = \frac{\text{SST}}{N} = \frac{\text{SST}}{kn}$$

and ρ is the intracluster correlation (ICC), given by

$$\rho = \frac{(k-1)n\text{MSB} - \text{SST}}{(n-1)\text{SST}} = 1 - \frac{n}{n-1}\frac{\text{SSW}}{\text{SST}}.$$

For N large (unknown),

$$\rho \approx \frac{\text{MSB} - \text{MST}}{(n-1)\text{MST}}$$

where $\text{MST} = \frac{\text{SST}}{nk-1}$.

6.2.1.3 ICC

ICC measures the correlation among elements in the same cluster or systematic sample. It provides a measure of homogeneity within a cluster, i.e., it tells us how similar elements in the same cluster are.

If ICC is positive, then element within the systematic sample tend to be similar, and then systematic sampling is less efficient (will yield higher variance of the sample mean than SRS).

If ICC is negative, then element within the systematic sample tend to be different, and then systematic sampling is more efficient (will yield lower variance of the sample mean than SRS).

If ICC close to 0, systematic sampling is roughly equivalent to SRS.

An approximation of ρ is

$$\widehat{\rho} = \frac{1}{n-1}\left(\frac{n\widehat{\mathrm{Var}}[\overline{y}_{\mathrm{sys}}]}{\widehat{\sigma}^2} - 1\right).$$

6.2.2 Estimation of Var$[\overline{y}_{\mathrm{sys}}]$

An unbiased estimate of Var$[\overline{y}_{\mathrm{sys}}]$ cannot be obtained by using data from only one single systematic sample without additional information about the population. To estimate the variance, we need to know the structure of the population.

Definition 6.2. A population is ***random*** if the elements of the population are in random order, i.e., the list or sampling frame is in random order.

In many situation, the ordering of the population is unrelated to the characteristics of interest. In random order population, systematic sampling is likely to produce a sample that behaves like an SRS and we expect $\rho \approx 0$. In this situation, simple random and systematic samplings will give similar results. We can use SRS formulas to estimate Var$[\overline{y}_{\mathrm{sys}}]$ as

$$\widehat{\mathrm{Var}}[\overline{y}_{\mathrm{sys}}] = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

where s^2 is the sample variance.

Definition 6.3. A population is ***ordered*** if the elements of a population have values that trend upward or downward when they are listed, i.e., the sampling frame is in increasing or decreasing order.

Example 6.4. Class list ordered by first test results and variable of interest is second test results.

In this situation, Var$[\overline{y}_{\mathrm{sys}}]$ is less than the variance of the sample mean in SRS of the same size since $\rho < 0$. To estimate the variance without bias, we can use repeated systematic sample.

Definition 6.4. A population is ***periodic*** if the elements of a population have values that tend to cycle upward and downward in a regular pattern when listed, i.e., the sampling frame has a periodic pattern.

If we sample at the same interval as the periodicity, systematic sampling will be less precise than SRS. The systematic sampling is most dangerous when the population is in a cyclical or periodic order, and the sampling interval coincides with a multiple of the period.

6.2.2.1 Difference Method

Difference method is useful for ordered population. Suppose sample of size $n : y_1, \cdots, y_n$ with $\mathbb{E}[y_i] = \mu, \mathrm{Var}[y_i] = \sigma^2$. Consider successive differences

$$d_i = y_{i+1} - y_i, i = 1, \cdots, n-1.$$

A sample of size n yield $n_d = n - 1$ successive differences $d_1, \cdots, d_{n_d}$. We have

$$\mathbb{E}[d_i] = 0, \mathrm{Var}[d_i] = 2\sigma^2$$

and thus $\frac{1}{n_d} \sum\limits_{i=1}^{n_d} d_i^2$ is unbiased estimator of $\mathrm{Var}[d_i]$ and thus

$$S_d^2 = \frac{1}{2n_d} \sum_{i=1}^{n_d} d_i^2$$

is an estimator for σ^2.

The difference estimator of $\mathrm{Var}[\overline{y}_{\text{sys}}]$ is

$$\widehat{\mathrm{Var}}_d[\overline{y}_{\text{sys}}] = \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}.$$

6.2.2.2 Repeated Systematic Sampling (RSS)

To avoid problem with additional information about the structure of the population entirely, we can use repeated systematic sampling (selecting several systematic samples without repetition, instead of just one) to obtain more information about population: $\overline{y}_1, \cdots, \overline{y}_k$ needed to estimate $\mathrm{Var}[\overline{y}_{\text{sys}}]$.

Let n_s be the number of repeated systematic samples, n' be the sample size of one systematic sample, k' be the selection step for one systematic sample, and $n = n_s n'$ be the total sample size. The process to draw RSS is: First we choose $k' = \frac{N}{n'}$. Then we select n_s random numbers between 1 and k'. Finally, the constant k' is added to each of these random starting points. The process of adding the constant is continued until n_s samples of size n' are obtained. The code is:

```
y = # population
N = length(y)

# Suppose we want to select ns=4 systematic sample of size n'=3
np = 3
ns = 4

# Choose kp: the selection step for each sample of size n'
kp = floor(N/np)

# Random selection of ns=4 integer between 1 and kp=7
set.seed(1)
first = sample(1:kp, ns) # Selecting sample start, ns times

# Select the first, second, third and fourth sample
sys1 = seq(first[1], N, kp)
sys2 = seq(first[2], N, kp)
sys3 = seq(first[3], N, kp)
sys4 = seq(first[4], N, kp)
```

The unbiased estimation of μ is

$$\widehat{\mu} = \overline{y}_{\text{sys, rep}} = \frac{1}{n_s}\sum_{i=1}^{n_s}\overline{y}_i = \overline{y}_{\text{sys}}.$$

The estimated variance is

$$\widehat{\text{Var}}[\overline{y}_{\text{sys, rep}}] = \frac{k' - n_s}{k'}\frac{S^2_{\text{rep}}}{n_s}, S^2_{\text{rep}} = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(\overline{y}_i - \widehat{\mu})^2.$$

The unbiased estimation of τ is

$$\widehat{\tau}_{\text{sys, rep}} = N\overline{y}_{\text{sys, rep}} = N\sum_{i=1}^{n_s}\frac{\overline{y}_i}{n_s}.$$

The estimated variance is

$$\widehat{\text{Var}}[\widehat{\tau}_{\text{sys, rep}}] = N^2\frac{k' - n_s}{k'}\frac{S^2_{\text{rep}}}{n_s}.$$

6.2.2.3 Rules

- We should use the variance estimate based simple random sampling only when we have good reason to believe that the population elements are in purely random order.
- We should use the variance estimate based successive differences whenever we suspect the population elements to be other than purely randomly ordered.
- Repeated systematic sampling allow the experimenter to estimate the population mean or total and the variance of the estimator without making assumptions about the nature of the population.

6.3 Selecting Sample Size for Estimating μ, τ and p

To estimate μ from a systematic sample within B bound on the error of estimation, the required sample size is given by

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}, D = \frac{B^2}{4}.$$

The sample size required to estimated τ with a bound on the error B is

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}, D = \frac{B^2}{4N^2}.$$

The sample size required to estimate p with a bound on the error of estimation B is

$$n = \frac{Npq}{(N-1)D + pq}, D = \frac{B^2}{4}.$$

In practice, we do not know p. An approximate sample size can be found by replacing p with an estimated value, usually obtained from similar past surveys. If no past information is available, we can substitute $p = 0.5$ into the above to obtain a conservative sample size (one that is likely to be larger than required).