P08802 Survey Fundamentals

Methods of Estimation, Surveying, and Sampling

Kayvan Nejabati Zenouz[1]

Oxford Brookes University

June 7, 2019

[1] Office: PG1.06 Email: knejabati-zenouz@brookes.ac.uk

Table of Contents

Module Introduction

Module Aims

The aim of this course is to provide an overview of sampling and estimation fundamentals. In particular, you will learn to

1. Select appropriate methods for sampling from finite populations, including the most common sampling designs.
2. Appropriately estimate finite population parameters and assess estimation errors.
3. In conjunction with others, create effective and non-biased surveys taking account of social and cultural expectations.

Module Introduction

Topics to be Covered

1. Basic Concepts
2. Surveys & Questionnaires
3. Estimation
4. Simple Random Sampling
5. Ratio Estimation
6. Stratified Sampling I
7. Stratified Sampling II
8. Cluster Sampling
9. Systematic Sampling I
10. Systematic Sampling II
11. Non-sampling Errors

Assessment

1. Group questionnaire design and pilot 15%
2. Group revision of questionnaire 25%
3. Individual report on findings of questionnaire research 60%

Reading List

1. Knottnerus, P. (2002) Sample Survey Theory. Springer.
2. John A. Rice (1987) Mathematical Statistics and Data Analysis
3. Andres, L. (2012) Designing and Doing Survey Research. Sage.
4. Fink, A. (2016) How to conduct surveys: a step-by-step guide. Sage.
5. Fowler, F. J., (2014) Survey Research Methods, 5th Edition. Los Angeles: Sage.
6. Kent, R. (2001) Data Construction and Data Analysis for Survey Research. Palgrave.
7. Sapsford, R. (2011) Survey Research, 2nd Edition. Sage.

Week 1
Review of Basic Concepts

Intended Learning Outcomes

In this session you will review

1. Basics about surveys sampling vs census.
2. Relevant theory on probability distributions and estimation.

Sample Surveys

Definition

Sample surveys are used to obtain information about a large population by examining only a small fraction of that population.

Numerical parameters for characteristics of the population, called **statistics**, e.g., mean, proportion, variance, are estimated using surveys.

- A sample statistic is calculated and this provides an estimate of the population value.
- New samples collected in the same way will produce new estimates.
- These estimates will not, in general, be equal, since each one is based on a different collection of values. This fact is called **sampling variability**.

Surveys vs Census

A Census:

- Aims for complete coverage of the population.
- May use too much time/money/effort.
- Should provide large numbers even in minority groups.
- May be of value to other studies/research projects.

A Sample Survey:

- Saves money/time/effort.
- Can provide a good enough level of accuracy, but involves a margin of error.
- Involves an element of risk.
- May produce only small numbers in minority groups.

Population Parameters

We shall review some basic statistical concepts...

Definition (Population Parameters)

Consider a population or set of N numbers $X_1, ..., X_N$. The population **mean** of variable x, also called the average value of x, is defined by

$$\overline{X} = \frac{X_1 + \cdots + X_N}{N}.$$

The population **variance** is defined by

$$\sigma_x^2 = \mathrm{Var}\left(X\right) = \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2.$$

Its square root σ_x is called the **standard deviation**.

Variance and Covariance

Exercise

Prove

$$\operatorname{Var}\left(X\right) = \overline{X^2} - \overline{X}^2.$$

Definition (Covariance)

The population covariance between two variables x and y is defined by

$$\operatorname{Cov}\left(X, Y\right) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right).$$

Random Variables

Definition (Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is a function from a set of possible outcomes Ω to real numbers space $\mathbb{R}$. The probability that X takes on a value in a set $S \subseteq \mathbb{R}$ is written as $\Pr\left(X \in S\right)$.

If the image of X is a discrete set, X can take on only a finite or at most a countably infinite number of values, then X is known as a **discrete** random variable, otherwise X is a **continuous** random variable.

Exercise

Suppose we flip a coin twice. What is the set of outcomes?
Let X be the number of heads. What is the image of X? What is $\Pr\left(X = 1\right)$?

Random Variables

Example

The Bernoulli distribution assumes that there are just two possible outcomes, which occur with probabilities θ and $1-\theta$ for some $\theta \in [0,1]$. We write $\Pr\left(X=1\right)=\theta$ and $\Pr\left(X=0\right)=1-\theta$.

Joint Probability Distribution

For two random variables X and Y we write

$$\Pr\left(X = X_i \text{ and } Y = Y_j\right).$$

for their joint probability distribution. Recall we have following.

- $\sum_{i,j} \Pr\left(X = X_i \text{ and } Y = Y_j\right) = 1$
- $\Pr\left(X = X_i\right) = \sum_j \Pr\left(X = X_i \text{ and } Y = Y_j\right)$
- $\Pr\left(Y = Y_j\right) = \sum_i \Pr\left(X = X_i \text{ and } Y = Y_j\right)$
- $\Pr\left(X = X_i \text{ and } Y = Y_j\right) = \Pr\left(X = X_i\right)\Pr\left(Y = Y_j\right)$ if X and Y are independent.

Moments of Random Variables

Definition (Expectation and Variance)

Let X and Y be discrete random variables. The expectation of X is defined by

$$\mathrm{E}\left(X\right) = \sum_{i=1} X_i \mathrm{Pr}\left(X = X_i\right).$$

The variance of X is defined by

$$\sigma_x^2 = \mathrm{Var}\left(X\right) = \mathrm{E}[X - \mathrm{E}\left(X\right)]^2 = \sum_{i=1} [X_i - \mathrm{E}\left(X\right)]^2 \mathrm{Pr}\left(X = X_i\right).$$

σ_x is called the **standard deviation**. Furthermore,

$$E\left(XY\right) = \sum_{i=1} \sum_{j=1} X_i Y_j \mathrm{Pr}\left(X = X_i, Y = Y_j\right).$$

Covariance and Correlation Coefficient

The covariance and the correlation coefficient between the random variables X and Y are defined by

$$\sigma_{xy} = \mathrm{Cov}\left(X, Y\right) = \mathrm{E}[\left(X - \mathrm{E}\left(X\right)\right)\left(Y - \mathrm{E}\left(Y\right)\right)]$$

$$\rho_{x,y} = \mathrm{Cor}\left(X, Y\right) = \frac{\mathrm{Cov}\left(X, Y\right)}{\sigma_x \sigma_y}.$$

Exercise

Prove the following.

(i) $\mathrm{E}\left(\alpha + \beta X\right) = \alpha + \beta \mathrm{E}\left(X\right)$.

(ii) $\mathrm{Var}\left(\alpha + \beta X\right) = \beta^2 \mathrm{Var}\left(X\right)$.

(iii) $\mathrm{E}\left(X + Y\right) = \mathrm{E}\left(X\right) + \mathrm{E}\left(Y\right)$.

(iv) $\mathrm{Var}\left(X + Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\mathrm{Cov}\left(X, Y\right)$.

(v) $\mathrm{Cov}\left(X + Y, Z\right) = \mathrm{Cov}\left(X, Z\right) + \mathrm{Cov}\left(Y, Z\right)$.

Continuous Random Variables

The same properties can be derived for continuous random variables. A random variable X is called (absolutely) continuous if there exists a non-negative function $f(x)$, called **density function**, with

$$\Pr\left(X \leq x\right) = \int_{-\infty}^{x} f(x)dx, \text{ for } -\infty \leq x \leq \infty.$$

The expectation and variance of X are defined by

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} xf(x)dx,$$
$$\mathrm{Var}\left(X\right) = \int_{-\infty}^{\infty} \left[x - \mathrm{E}\left(X\right)\right]^2 f(x)dx.$$

Continuous Random Variables

Normal Distribution

A **normal distribution** with **mean** μ and **standard deviation** σ written as $N(\mu, \sigma^2)$ has a **probability density function** given by

$$\varphi\left(x \mid \mu, \sigma^2\right) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < +\infty.$$

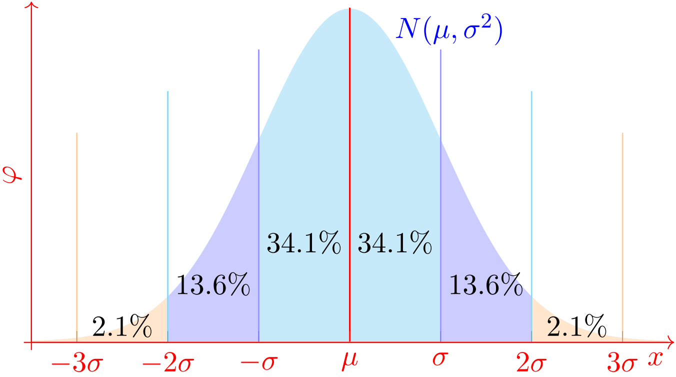
Therefore, $\mathrm{P}\left(x \leq z\right)$, probability of x taking a value less than z, is obtained by

$$\mathrm{P}\left(x \leq z\right) = \int_{-\infty}^{z} \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx,$$

this value is usually denoted by $\phi(z)$.

Normal Distribution

For normal distribution 68.2% of data lie within one standard deviation away from the mean, 95.4% of data lie within two standard deviations away from the mean, and 99.6% of data lie within three standard deviations away from the mean...



Exercises

Exercises

(i) The Binomial distribution $\text{Bin}(m, \theta)$ assumes that X has mass function

$$\Pr(X = x) = \binom{m}{x} \theta^x (1 - \theta)^{n-x}.$$

Find $\text{E}(X)$ and $\text{Var}(X)$.

(ii) The Exponential distribution $\text{Exp}(\theta)$ assumes that X has density function

$$f(x) = \theta e^{-\theta x} \text{ for } x > 0 \text{ and } \theta > 0.$$

Find $\text{E}(X)$ and $\text{Var}(X)$.

Summary

- Surveys are used to gather information about characteristics of populations.
- Numerical parameters called statistics are calculated using surveys.
- Discrete and continuous random variable are used to model data-generating processes.
- Next time we will look into the concept of estimation and start working on the coursework.

Week 2-3
Estimation and Questionnaires

Intended Learning Outcomes

In this chapter you will review

1. Fundamentals of estimation techniques.
2. Designing questionnaires.

Estimation

- Probability distributions **model** data-generating processes in real life.
- Models typically contains a **parameter**, the value of which we are free to choose.
- We may also want to determine certain statistics about the population.
- Using data we aim to choose the parameter to make the model distribution similar to the distribution of the data.
- Understanding the parameters of data is known as **statistical inference**.
- The two main tools of inference are **estimation** and **hypothesis testing**.

Estimation

The goal of **estimation** is to use the data $x_1, x_2, ..., x_x$ from a population to construct a best guess or estimate for the parameter value of interest.

Definition (Estimator)

Suppose θ is a parameter, or statistic, about a population $X_1, ..., X_N$. Then an estimator $\hat{\theta}$ of θ is a random variable that given a sample $\mathbf{x} = (x_1, ..., x_n)$ of the population calculates an estimate $\hat{\theta}(\mathbf{x})$ of θ.

Example

Suppose that $X = (X_1, ..., X_n)$ is an independent random sample from a population having common mean μ. Then three estimates of μ can be given by $\hat{\theta}_1(X) = \overline{X}$, $\hat{\theta}_2(X) = X_1$, and $\hat{\theta}_3(\mathbf{X}) = \frac{X_1+X_n}{2}$.

Properties of Estimators

A good choice for an estimator $\hat{\theta}$ will have a sampling distribution that is concentrated near the true value of θ. That is to say $\mathrm{E}(\hat{\theta}) = \theta$ and $\mathrm{Var}(\hat{\theta})$ is small.

Definition (Bias of an Estimator)

If $\hat{\theta}$ is an estimator of θ, then the **bias** of $\hat{\theta}$ is defined as

$$\mathrm{B}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$$

We say that $\hat{\theta}$ is an unbiased estimator of θ if $\mathrm{B}(\hat{\theta}) = 0$.

Definition (Standard Error of an Estimator)

If $\hat{\theta}$ is an estimator of θ, then the standard deviation of $\hat{\theta}$ is known as the **standard error** $\mathrm{Se}(\hat{\theta})$ of $\hat{\theta}$. The mean squared error of $\hat{\theta}$ is the quantity $\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]$.

Mean Squared Error

Theorem

We have

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

Proof.

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta}-\theta)^2] &= \mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}(\hat{\theta})+\mathrm{E}(\hat{\theta})-\theta\right)^2\right] \\ &= \mathrm{E}[(\hat{\theta}-\mathrm{E}(\hat{\theta}))^2] + \mathrm{E}[(\mathrm{E}(\hat{\theta})-\theta)^2] \\ &+ 2\underbrace{\mathrm{E}[(\hat{\theta}-\mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta})-\theta)]}_{=0} \\ &= \mathrm{Var}(\hat{\theta}) + (\mathrm{E}(\hat{\theta})-\theta)^2 = \mathrm{Var}(\hat{\theta}) + \mathrm{B}(\hat{\theta})^2.\end{aligned}$$

□

Exercises

Exercises

Suppose that $X_1, ..., X_n$ is an independent random sample from a population having common mean μ and variance σ^2.

(i) Show $\overline{X}$ is an unbiased estimator of μ and find its standard error.

(ii) Find the bias of the estimator

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(iii) Show the estimator

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is an unbiased estimator of σ^2.

Solution (i)

Note,

$$\mathrm{E}(\overline{X}) = \frac{1}{n}\left(\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)\right) = \frac{n\mu}{n} = \mu,$$

so $\overline{X}$ is unbiased. For standard error, note we have

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\left(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

so

$$\mathrm{Se}(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Solution (ii)

To find the bias of σ_x^2 we work as follows.

$$\begin{aligned}
\mathrm{E}(\sigma_x^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\overline{X} + \sum_{i=1}^{n}\overline{X}^2\right) \\
&= \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i^2\right) - \mathrm{E}\left(\overline{X}^2\right),
\end{aligned}$$

so

Solution (ii)

$$\mathrm{E}(\sigma_x^2) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(X_i^2\right) - \mathrm{E}\left(\overline{X}^2\right).$$

Note we have

$$\mathrm{E}\left(X_i^2\right) = \mathrm{Var}\left(X_i\right) + \mathrm{E}\left(X_i\right)^2 = \sigma^2 + \mu^2$$

using rules of expectation and

$$\mathrm{E}\left(\overline{X}^2\right) = \mathrm{Var}\left(\overline{X}\right) + \mathrm{E}\left(\overline{X}\right)^2 = \frac{\sigma^2}{n} + \mu^2,$$

again using rules of expectation and (i), so we get

$$\mathrm{E}(\sigma_x^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

Thus $\mathrm{B}(\sigma_x^2) = \dfrac{n-1}{n}\sigma^2 - \sigma^2 = -\dfrac{\sigma^2}{n}.$

Solution (iii)

Note, we have

$$\begin{aligned} \mathrm{E}\left(S^2\right) &= \mathrm{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\ &= \mathrm{E}\left(\frac{n}{n-1}\sigma_x^2\right) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2. \end{aligned}$$

Example

Example

The proportion of individuals in a population who support a certain opinion is P, while the proportion who do not is $Q = 1 - P$. In a random sample of n independent individuals, r support the opinion. The sample proportion who support the opinion can be estimated by

$$\frac{r}{n},$$

where r is a Binomial random variable with parameters n and p. The standard error is

$$\sqrt{\mathrm{Var}\left(\frac{r}{n}\right)} = \sqrt{\frac{1}{n^2}\mathrm{Var}(r)} = \sqrt{\frac{nPQ}{n^2}}.$$

This can be estimated by $\sqrt{\frac{pq}{n}}$.

Relative Efficiency

In the example on page 20 we saw that there can be several estimators for our quantity of interested.

Can we find an unbiased estimator for our quantity of interest?
Answer. No in general. But we can find estimators which are **asymptotically unbiased**, i.e.,

$$\lim_{n\to\infty} \mathrm{B}(\hat{\theta}) = 0.$$

How can we judge which estimator is better?
Answer. In general if $\mathrm{MSE}(\hat{\theta}_1) < \mathrm{MSE}(\hat{\theta}_2)$, then $\hat{\theta}_1$ is a better estimator compared to $\hat{\theta}_2$. If the estimators are unbiased The **relative efficiency** of $\hat{\theta}_1$ compared to $\hat{\theta}_2$ is defined as

$$\mathrm{RE}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}.$$

Exercise

Exercise

Suppose that $X = (X_1, ..., X_n)$ is an independent random sample from a population having common mean μ and variance σ^2. Then three estimates of μ can be given by

$$\hat{\theta}_1(X) = \overline{X},\ \hat{\theta}_2(X) = X_1,\ \text{and}\ \ \hat{\theta}_3(\mathbf{X}) = \frac{X_1 + X_n}{2}.$$

Find the pairwise relative efficiencies of these estimators.

Confidence Interval (Bound on Error)

Once we choose an unbiased estimator $\hat{\theta}$, a bound B can be found such that

$$\Pr\left(\left|\hat{\theta}-\theta\right| < B\right) = 1-\alpha \text{ for some } \alpha.$$

This is due to the following theorem.

Theorem (Chebyshev's Inequality)

Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $k > 0$, we have

$$\Pr\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

Central Limit Theorem

Consider a sequence of independent and identically distributed (i.i.d.) random variables, $X_1, X_2, ...$ having the common mean and variance μ and σ^2. The sample mean of $X_1, X_2, ..., X_n$, we have

$$\mathrm{E}(\overline{X}) = \mu \text{ and } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

The central limit theorem says that, for a fixed number z we have

$$\Pr\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z\right) \Phi(z) \text{ as } n \to \infty,$$

where Φ is the cumulative distribution function of the standard normal distribution. In such a case for any $\delta > 0$ we have

$$\Pr\left(|\overline{X} - \mu| < \delta\right) \approx 2\Phi\left(\frac{\delta}{\frac{\sigma}{\sqrt{n}}}\right) - 1.$$

Confidence Interval

For $0 < \alpha < 1$ write $z_\alpha = \Phi^{-1}(\alpha)$. Then we have

$$\Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha.$$

Cramér-Rao Inequality

The Cramér-Rao Inequality gives a **bound on the standard error** of $\hat{\theta}$.

Theorem (Cramér-Rao Inequality)

Suppose that $X_1, ..., X_n$ *is a random sample from a population having a common density function* $f(x \mid \theta)$ *depending on a parameter* θ*, and let* $\hat{\theta}$ *be an unbiased estimator of* θ*. If* $f(x \mid \theta)$ *is a smooth function of* y *and* θ*, then*

$$\mathrm{Var}\left(\hat{\theta}\right) \geq \frac{1}{nI(\theta)},$$

where

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial \theta^2} \log\left(f(x \mid \theta)\right)\right).$$

Planning a Survey

- **Statement of objectives:** Clear, brief, simple, understood by all workers, referred to repeatedly. Target population defined.
- **Sample design:** Sampling method(s), sampling units (elements), sample size(s), sampling frame.
- **Questionnaire design and piloting:** Determined by objectives & design.
- **Data collection (fieldwork):** Detailed planning. Well established lines of duty.
- **Data processing (management):** Huge amount of data may be collected. Quality control measures needed to ensure consistency between data collected & processed.
- **Data Analysis:** Outline of methods of analysis & results which will be included in the final report.
- **Report**

Methods of Data Collection

- Personal interview
- Telephone interview
- Self-completed questionnaire
- Direct observation
- Surveys of documents or records

Questionnaire Design

- **Question wording:** Balance the phrasing. Avoid arguments/counter-arguments. Avoid harsh/negatively charged questions. Use clearly defined questions.
- **Response options:** Avoid questions that are too demanding/time consuming. Make the response categories (options), mutually exclusive, clear and logical.
- **Question ordering**
- **Open or closed questions**
- **Layout**
- **Prompts**
- **Time taken**

Questionnaires should be tested before use: pre-testing individual questions. checking the overall design.

More on Questionnaires

- Give your questionnaire a title.
- Keep the questionnaire short.
- You may offer incentives for responding if appropriate.
- Use you creativity use colours and images to make it attractive.
- Make it convenient.
- When choosing your sample make sure it is representative of the population you are studying.
- State who you are.
- Outline what the purpose of the survey is and why their response is important
- Explain how answers will be treated with confidentiality and anonymity (unless agreed with the respondent).
- Provide clear instructions as to how each question should be answered.

Summary

Today we reviewed the following concepts...

- Estimation, bias, and standard error of an estimator.
- Questionnaire design.
- Read more on questionnaire design on the handouts on moodle.

Week 4
Random Sampling Methods

Random Sampling Methods

In this chapter you learn about

1. Simple sampling method
2. Estimation of parameters when conducting simple sampling method

Motivation

Many familiar statistical formulae are based on the assumption that sample values are independent of each other, having been obtained by random selection from an infinite population.

This will not be the case in real life situations.

- Population may be finite.
- Samples may be dependent. This occurs when sampling without replacement.

Random Sampling Methods

Random sampling

Random sampling methods, or probability sampling, give every member of the population a non-zero chance of being selected.

Benefits

- Minimises the risk of bias which results when non-random method such as quota sampling are used.
- Allows statistical inference to be used to draw conclusions about the wider population from which the sample is drawn.
- Require a sampling frame or other method of random selection e.g., random digit dialling.

Sampling Frame

A sampling frame is a list identifying every individual in the population.

Simple Random Sampling

When a sample of n individuals (sampling units) is chosen randomly from a population of size N. There are

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

possible samples which can be chosen.

- Every sample has the same probability of being drawn, that is
$$\frac{n!(N-n)!}{N!}.$$
- Every individual in the sample has the same probability of being drawn.

The ratio $f = \dfrac{n}{N}$ is called the **sampling fraction**.

Sampling without Replacement

Denote by $\zeta_1, ..., \zeta_m$ the distinct values assumed by the N population members. Denote the number of population members that have the value ζ_j by n_j for $j = 1, ..., m$.

Lemma

Suppose X_i is the value of the i^{th} member of the sample. Then X_i is a discrete random variable with probability mass function

$$\Pr(X_i = \zeta_j) = \frac{n_j}{N}.$$

Also we have

$$\mathrm{E}(X_i) = \mu \text{ and } \mathrm{Var}(X_i) = \sigma^2.$$

Expectation and Standard Error

We are interested in

1. $\mathrm{E}(\overline{X})$ as a measure of the centre of the sampling distribution.
2. $\mathrm{Se}(\overline{X})$ as a measure of the dispersion of the sampling distribution about $\mathrm{E}(\overline{X})$.

Note previously (slide 23) when we chose independent samples from population having common mean μ and σ^2 we obtained

$$\mathrm{E}\left(\overline{X}\right) = \mu \text{ and } \mathrm{Se}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}.$$

Theorem (Simple Random Sampling)

With simple random sampling

$$\mathrm{E}\left(\overline{X}\right) = \mu \;\; \textit{and} \; \mathrm{Se}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Proof of Theorem

It is easy to prove that $\mathrm{E}\left(\overline{X}\right) = \mu$. We shall sketch the proof

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right).$$

Note, we have

$$\begin{aligned}\mathrm{Var}\left(\overline{X}\right) &= \frac{1}{n^2}\mathrm{Var}\left(X_1 + \cdots + X_n\right) \\ &= \frac{1}{n^2}\left(n\mathrm{Var}\left(X_1\right) + \sum_{\substack{i,j=1\\ i\neq j}}^{n}\mathrm{Cov}\left(X_i, X_j\right)\right).\end{aligned}$$

Note, in the case of sampling with replacement we had

$$\mathrm{Cov}\left(X_i, X_j\right) = 0 \text{ for } i \neq j.$$

We hear we need the following lemma.

Proof of Theorem Cont.

Lemma

With simple random sampling we have

$$\mathrm{Cov}\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1} \text{ for } i \neq j.$$

Proof of the lemma. We have

$$\mathrm{Cov}\left(X_i, X_j\right) = \mathrm{E}\left(X_i X_j\right) - \mathrm{E}\left(X_i\right)\mathrm{E}\left(X_j\right) = \mathrm{E}\left(X_i X_j\right) - \mu^2.$$

Now

$$\mathrm{E}\left(X_i X_j\right) = \sum_{k,l=1}^{m} \zeta_k \zeta_l \mathrm{Pr}\left(X_j = \zeta_k, X_i = \zeta_l\right).$$

Using the conditional law of probability we have

$$\mathrm{Pr}\left(X_i = \zeta_k, X_j = \zeta_l\right) = \mathrm{Pr}\left(X_j = \zeta_l \mid X_i = \zeta_k\right)\mathrm{Pr}\left(X_i = \zeta_k\right).$$

Proof of Theorem Cont. Proof of Lemma

Thus, we can write

$$\mathrm{E}\left(X_i X_j\right) = \sum_{k=1}^{m} \zeta_k \Pr\left(X_i = \zeta_k\right) \sum_{l=1}^{m} \zeta_l \Pr\left(X_j = \zeta_l \mid X_i = \zeta_k\right).$$

Next we find

$$\Pr\left(X_j = \zeta_l \mid X_i = \zeta_k\right) = \begin{cases} \dfrac{n_l}{N-1} & k \neq l, \\ \dfrac{n_l - 1}{N-1} & k = l. \end{cases}$$

Therefore,

$$\sum_{l=1}^{m} \zeta_l \Pr\left(X_j = \zeta_l \mid X_i = \zeta_k\right) = \sum_{l \neq k}^{m} \zeta_l \frac{n_l}{N-1} + \zeta_k \frac{n_k - 1}{N-1}.$$

Proof of Theorem Cont. Proof of Lemma

Now we get

$$\mathrm{E}\left(X_i X_j\right) = \sum_{k=1}^{m} \zeta_k \Pr\left(X_i = \zeta_k\right) \left(\sum_{l \neq k}^{m} \zeta_l \frac{n_l}{N-1} + \zeta_k \frac{n_k - 1}{N-1} \right).$$

Given this with a few algebraic manipulations (**Exercise**) we obtain

$$\mathrm{E}\left(X_i X_j\right) = \mu^2 - \frac{\sigma^2}{N-1}$$

which give the proof for the lemma.

Proof of Theorem Cont.

Given the lemma, then we can write

$$\begin{aligned}
\operatorname{Var}\left(\overline{X}\right) &= \frac{1}{n^2}\left(n\operatorname{Var}\left(X_1\right) + \sum_{\substack{i,j=1\\ i\neq j}}^{n} \operatorname{Cov}\left(X_i, X_j\right)\right)\\
&= \frac{1}{n^2}\left(n\operatorname{Var}\left(X_1\right) - \sum_{\substack{i,j=1\\ i\neq j}}^{n} \frac{\sigma^2}{N-1}\right)\\
&= \frac{1}{n^2}\left(n\sigma^2 - \left(n^2 - n\right)\frac{\sigma^2}{N-1}\right),
\end{aligned}$$

which proves the theorem.

Remark

Note, in the

$$\operatorname{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

we see that if $f = \frac{n}{N}$ is small, then we have

$$\operatorname{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \approx \frac{\sigma^2}{n}.$$

Estimating Population Variance

Having looked that the sampling distribution for $\overline{X}$, let us now investigate the estimator

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Here the main results is the following

Theorem

With simple random sampling we have

$$\mathrm{E}\left(\sigma_x^2\right) = \sigma^2 \frac{n-1}{n}\frac{N}{N-1}.$$

Therefore,

$$\frac{n}{n-1}\frac{N-1}{N}\sigma_x^2$$

is an unbiased estimator for σ^2.

Corollary

Recall we had

$$\operatorname{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

and in the previous slide we saw that

$$\frac{n}{n-1}\frac{N-1}{N}\sigma_x^2$$

is an unbiased estimator for σ^2, so

$$\frac{\sigma_x^2}{n-1}\left(\frac{N-n}{N}\right)$$

is an unbiased estimator for $\operatorname{Var}\left(\overline{X}\right)$.

Example

For example, in the case were the responses are binary $\overline{X}$ is the proportion of the sample that possesses the characteristic of interest denoted by $\hat{p}$, and we have

$$\operatorname{Var}\left(\hat{p}\right) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right).$$

Exercise

A simple random sample of 50 of the 393 hospitals was taken. From this sample, $\overline{X} = 938.5$ (population value $= 814.6$) and $s = 614.53$ population value $\sigma = 590$. Find an estimate of the variance of $\overline{X}$ and an estimate of the variance of X.

Summary

The following table summarises the the properties of estimators in the simple sampling method.

Parameter	Estimate	Variance Estimate	Estimated Variance
μ (location)	$\overline{X}$	$\frac{\sigma^2}{n}\frac{N-n}{N-1}$	$\frac{\sigma_x^2}{n-1}(1-f)$
p (proportion)	$\hat{p}$	$\frac{p(1-p)}{n}\frac{N-n}{N-1}$	$\frac{\hat{p}(1-\hat{p})}{n-1}(1-f)$
τ (total)	$T = N\overline{X}$	$N^2\frac{\sigma^2}{n}\frac{N-n}{N-1}$	$N^2\frac{\sigma_x^2}{n-1}(1-f)$
σ^2 (spread)	$\frac{n(N-1)}{(n-1)N}\sigma_x^2$		

Summary

Today we reviewed the following concepts...

- Estimation, bias, and standard error for estimators with simple random sampling.
- Read more on John A. Rice (1987) Mathematical Statistics and Data Analysis, Chapter 7 [Ric06].

Week 5-6
Estimation of Ratio
and
Stratified Sampling

Intended Learning Outcomes

In this chapter you learn about

1. Estimation of ratio with simple sampling
2. Stratified sampling

Estimation of Ratio

- In this section, we consider the estimation of a ratio. For each member of a population, two values, x and y are recorded.
- The ratio of interest is

$$r = \frac{\sum_i y_i}{\sum_i x_i} = \frac{\mu_y}{\mu_x}.$$

- For example, if y is weekly food expenditure and x is number of inhabitants, then r is weekly food cost per inhabitant.
- Note, the ratio above is different to

$$\sum_i^N \frac{y_i}{x_i}.$$

Ration Estimation

- Suppose that a sample is drawn consisting of the pairs (X_i, Y_i); the natural estimate of r is

$$R = \frac{\overline{Y}}{\overline{X}}.$$

- In this section we find $\mathrm{E}(R)$ and $\mathrm{Var}(R)$.
- We cannot do this in closed form. We will therefore employ the approximate methods.

Approximation Methods

Suppose $Z = g(X, Y)$. We can expand g around $\mu = (\mu_X, \mu_Y)$ to get

$$
\begin{aligned}
Z \approx & g(\mu) + (X - \mu_X)\frac{\partial g}{\partial X}(\mu) + (Y - \mu_Y)\frac{\partial g}{\partial Y}(\mu) \\
& + \frac{1}{2}(X - \mu_X)^2\frac{\partial^2 g}{\partial X^2}(\mu) + \frac{1}{2}(Y - \mu_Y)^2\frac{\partial^2 g}{\partial Y^2}(\mu) \\
& + \frac{1}{2}(X - \mu_X)(Y - \mu_Y)\frac{\partial^2 g}{\partial X \partial Y}(\mu).
\end{aligned}
$$

In such case we can have

$$
\mathrm{E}(Z) \approx g(\mu) + \frac{1}{2}\sigma_x^2\frac{\partial^2 g}{\partial X^2}(\mu) + \frac{1}{2}\sigma_y^2\frac{\partial^2 g}{\partial Y^2}(\mu) + \sigma_{xy}^2\frac{\partial^2 g}{\partial X \partial Y}(\mu),
$$

and we may approximate $\mathrm{Var}(Z)$ by

$$
\mathrm{Var}(Z) \approx \sigma_x^2\left(\frac{\partial g}{\partial X}(\mu)\right)^2 + \sigma_y^2\left(\frac{\partial g}{\partial Y}(\mu)\right)^2 + 2\sigma_{xy}^2\frac{\partial g}{\partial X}(\mu)\frac{\partial g}{\partial Y}(\mu).
$$

E(R) and Var(R)

Given the estimate in the previous page, letting

$$Z = g(X, Y) = \frac{Y}{X}$$

we can prove the following theorems.

Theorem (Expectation of Ratio)

With simple random sampling,

$$\mathrm{E}(R) = r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r\sigma_x^2 - \rho\sigma_x\sigma_y\right).$$

Theorem (Variance of Ratio)

With simple random sampling,

$$\mathrm{Var}(R) = \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right).$$

Example

Suppose that 100 people who recently bought houses are surveyed, and the monthly mortgage payment and gross income of each buyer are determined. Let y denote the mortgage payment and x the gross income. Suppose that

$$\overline{X} = 3100, \ \overline{Y} = 868, \ s_y = 250, \ s_x = 1200, \ \hat{\rho} = 0.85, \ R = 0.28.$$

Noting that

$$\hat{\rho} = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

we have that

$$\begin{aligned} \mathrm{E}(R) &= 0.28 + \frac{1}{100}\left(1 - \frac{99}{N-1}\right)\frac{1}{3100^2}\left(0.28 \times 1200^2 - 0.85 \times 1200 \times 250\right) \\ &= 0.28 + \left(1 - \frac{99}{N-1}\right) \times 0.00015 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(R) &= \frac{1}{100}\left(1 - \frac{99}{N-1}\right)\frac{1}{3100^2}\left(0.28^2 \times 1200^2 + 250^2 - 2 \times 0.28 \times 0.85 \times 1200 \times 250\right) \\ &= \left(1 - \frac{99}{N-1}\right) \times 0.000034. \end{aligned}$$

Stratified Random Sampling

- In stratified random sampling, the population is partitioned into subpopulations, or **strata**, which are then independently sampled.
- The results from the strata are then combined to estimate population parameters, such as the mean.
- For example, in samples of human populations, geographical areas often form natural strata.

Stratified Random Sampling

- The use of a stratified random sample guarantees a prescribed number of observations from each subpopulation, whereas the use of a simple random sample can result in underrepresentation of some subpopulations.
- The stratified sample mean can be considerably more precise than the mean of a simple random sample.
- A simple random sample is taken within each stratum, the results will follow easily from the derivations of earlier sections.

Properties of Stratified Estimates

Suppose there are L strata and let the number of population elements in stratum 1 be denoted by N_1, the number in stratum 2 be N_2 etc... Then we have

$$N = N_1 + \cdots + N_L.$$

Denote by μ_l and σ_l^2 the mean and variance of the l^{th} stratum. We have

$$\mu = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} x_{il} = \frac{1}{N} \sum_{l=1}^{L} N_l \mu_l = \sum_{l=1}^{L} W_l \mu_l,$$

where

$$W_l = \frac{N_l}{N}.$$

Stratified Estimates

Within each stratum, a simple random sample of size n_l is taken. The sample mean in stratum l is denoted by

$$\overline{X}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} X_{il}.$$

An estimate for μ is

$$\overline{X}_s = \sum_{l=1}^{L} \frac{N_l}{N} \overline{X}_l.$$

Exercise

Prove $\overline{X}_s$ is an unbiased estimator of μ.

Variance of $\overline{X}_s$

Note, samples from different strata are independent of one another so we have the following.

Theorem

The variance of the stratified sample mean is given by

$$\operatorname{Var}\left(\overline{X}_s\right) = \sum_{l=1}^{L} W_l^2 \left(\frac{\sigma_l^2}{n_l}\right)\left(\frac{N_l - n_l}{N_l - 1}\right).$$

Proof.

Exercise. □

Note, ignoring the finite population correction we get

$$\operatorname{Var}\left(\overline{X}_s\right) \approx \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l}.$$

Exercise

Assume that the number of beds in 4 hospitals is known but that the number of discharges is not. Let stratum A consist of the 98 smallest hospitals, stratum B of the 98 next larger, stratum C of the 98 next larger, and stratum D of the 99 largest. The following table shows the results of this stratification of hospitals by size.

Stratum	N_l	W_l	μ_l	σ_l
A	98	0.249	182	103
B	98	0.249	526	204
C	98	0.249	956	243
D	99	0.251	1592	419

Suppose we take samples of size $\frac{n}{4}$, for some n from each hospital. Find

$$\operatorname{Var}\left(\overline{X}_s\right).$$

Summary

Today we reviewed the following concepts...

- Estimation of ratio.
- Stratified sampling.

Week 7
Methods of Allocation for Stratified Sampling

Intended Learning Outcomes

In this chapter you learn about

1. Minimising variance with stratified sampling.
2. Proportional sampling and comparisons.

Methods of Allocation

Recall in the previous week for stratified sampling, ignoring the finite population correction, we had

$$\operatorname{Var}\left(\overline{X}_s\right) \approx \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l}, \text{ where } W_l = \frac{N_l}{N}.$$

Suppose we want to survey n samples. We aim to choose $n_1, ..., n_L$ in order to have the smallest possible value for

$$\operatorname{Var}\left(\overline{X}_s\right),$$

subject to $n_1 + \cdots + n_L = n$.

Methods of Allocation: Calculus of Variation

Give a function $f(x_1, ..., x_n)$ to be maximised subject to constrains $g(x_1, ..., x_n) = 0$, we use the Lagrange multiplier and function

$$\mathcal{L}(x_1, ..., x_n) = f - \lambda g,$$

and we solve for points $(x_1, ..., x_n)$ with $\nabla f = \lambda \nabla g$ and $g = 0$.

Theorem

The sample sizes $n_1, ..., n_L$ *that minimize* $\operatorname{Var}\left(\overline{X}_s\right)$ *subject to the constraint* $n_1 + \cdots + n_L = n$ *are given by*

$$n_l = n \frac{W_l \sigma_l}{\sum_k W_k \sigma_k}$$

where $l = 1, ..., L$.

Proof

We have a Lagrange function

$$\mathcal{L} = \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l} + \lambda \left(\sum_{l=1}^{L} n_l - n \right).$$

Then we have

$$\frac{\partial \mathcal{L}}{\partial n_l} = -\frac{W_l^2 \sigma_l^2}{n_l^2} + \lambda = 0 \text{ for } l = 1, ..., L.$$

This gives

$$n_l = \frac{W_l \sigma_l}{\sqrt{\lambda}}.$$

And we have

$$n = \frac{1}{\sqrt{\lambda}} \sum_{l=1}^{L} W_l \sigma_l,$$

which gives us the results we need.

Remark

Note substituting

$$n_l = n \frac{W_l \sigma_l}{\sum_k W_k \sigma_k}$$

in

$$\operatorname{Var}\left(\overline{X}_s\right) \approx \sum_{l=1}^{L} \frac{W_l^2 \sigma_l^2}{n_l}, \text{ where } W_l = \frac{N_l}{N},$$

we find

$$\operatorname{Var}\left(\overline{X}_{so}\right) \approx \frac{1}{n} \sum_{l=1}^{L} W_l \sigma_l \sum_{k=1}^{L} W_k \sigma_k = \frac{1}{n} \left(\sum_{l=1}^{L} W_l \sigma_l \right)^2 .$$

Example

For the exercise on page 70 find the optimal n_l for each $l = A, B, C, D$.

Proportional Allocation

Using the optimal allocation can be difficult as we dont always have information about σ_l.

An alternative is to sample proportionally

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \cdots = \frac{n_L}{N_L}$$

which hold if we set

$$n_l = n\frac{N_l}{N} = nW_l.$$

In this case we have

$$\overline{X}_{sp} = \sum_{l=1}^{L} W_l \overline{X}_l = \sum_{l=1}^{L} \frac{W_l}{n_l} \sum_{i=1}^{n_l} X_{il} = \frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n_l} X_{il}.$$

Proportional Sampling

Theorem

With stratified sampling based on proportional allocation, ignoring the finite population correction we have

$$\operatorname{Var}\left(\overline{X}_{sp}\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \sigma_l^2 .$$

We can compare $\operatorname{Var}\left(\overline{X}_{so}\right)$ and $\operatorname{Var}\left(\overline{X}_{sp}\right)$.

Theorem

With stratified random sampling we have

$$\operatorname{Var}\left(\overline{X}_{sp}\right) - \operatorname{Var}\left(\overline{X}_{so}\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \left(\sigma_l - \overline{\sigma}\right)^2 ,$$

where $\overline{\sigma} = \sum_{l=1}^{L} W_l \sigma_l$.

Proof for the Second Theorem

Note we have

$$\operatorname{Var}\left(\overline{X}_{sp}\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \sigma_l^2,$$

and

$$\operatorname{Var}\left(\overline{X}_{so}\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \sigma_l \sum_{k=1}^{L} W_k \sigma_k = \frac{1}{n}\overline{\sigma}^2.$$

Now $\operatorname{Var}\left(\overline{X}_{sp}\right) - \operatorname{Var}\left(\overline{X}_{so}\right)$ can be written as

$$\frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l^2 - \overline{\sigma}^2\right) = \frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l^2 - \overline{\sigma}^2 - \overline{\sigma}^2 + \overline{\sigma}^2 \sum_{l=1}^{L} W_l\right)$$

$$= \frac{1}{n}\left(\sum_{l=1}^{L} W_l \sigma_l^2 - 2\overline{\sigma}\sum_{l=1}^{L} \sigma_l W_l + \overline{\sigma}^2 \sum_{l=1}^{L} W_l\right)$$

$$= \frac{1}{n}\left(\sum_{l=1}^{L} W_l \left(\sigma_l^2 - 2\overline{\sigma}\sigma_l + \overline{\sigma}^2\right)\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \left(\sigma_l - \overline{\sigma}\right)^2.$$

Comparison with Simple Random Sampling

Example

For the exercise on page 70 calculate how much better optimal allocation is than proportional allocation for the population of hospitals.

We can also compare with simple random sampling. The variance under simple random sampling is, neglecting the finite population correction,

$$\mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n},$$

where

$$\sigma^2 = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} \left(x_{il} - \mu\right)^2 .$$

Comparison with Simple Random Sampling

Note we can write

$$\begin{aligned}(x_{il}-\mu)^2 &= (x_{il}-\mu_l+\mu_l-\mu)^2\\ &= (x_{il}-\mu_l)^2+2\,(x_{il}-\mu_l)\,(\mu_l-\mu)+(\mu_l-\mu)^2\,,\end{aligned}$$

so we have

$$\begin{aligned}\sum_{i=1}^{N_l}(x_{il}-\mu)^2 &= \sum_{i=1}^{N_l}(x_{il}-\mu_l)^2+2\,(x_{il}-\mu_l)\,(\mu_l-\mu)+(\mu_l-\mu)^2\\ &= \sum_{i=1}^{N_l}(x_{il}-\mu_l)^2+N_l\,(\mu_l-\mu)^2\end{aligned}$$

so

$$\sigma^2=\frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N_l}(x_{il}-\mu_l)^2+W_l\,(\mu_l-\mu)^2=\sum_{l=1}^{L}W_l\left(\sigma_l^2+(\mu_l-\mu)^2\right).$$

Comparison with Simple Random Sampling

With this we get

$$\operatorname{Var}\left(\overline{X}\right) - \operatorname{Var}\left(\overline{X}_{so}\right) = \frac{1}{n}\sum_{l=1}^{L} W_l \left(\mu_l - \mu\right)^2 .$$

Finally, if there is C_0 denote an overhead cost in implementing a survey by stratified sampling and C_l the cost per unit of survey-cum-selection so that the total cost is $C = C_0 + \sum_{l=1}^{L} C_l n_l$, we can take

$$n_l = n\frac{W_l \sigma_l}{\sqrt{C_l}} \left(\sum_k \frac{W_k \sigma_k}{\sqrt{C_l}}\right)^{-1} .$$

This rule prescribes a higher scale of sampling per stratum of relatively bigger size, higher stratum-variability taking account of the relatively lesser cost reflected through the square root of the survey-sampling cost per stratum.

Summary

Today we reviewed the following concepts...

- Various methods of allocation for stratified sampling.
- Comparison between them.

Week 8
Cluster Sampling

Intended Learning Outcomes

In this chapter you learn about

1. Cluster sampling methods.
2. Estimation of parameters.

Motivation

- Cluster sampling is opposite to stratified sampling.
- A population may be divided into natural clusters.
- A sample of clusters are then chosen.
- Each of the chosen clusters can then be studied.
- Cluster sampling may be used when the sampling frame of the ultimate units is not readily available or expensive to construct.
- This method of sampling is much cheaper, easier, and operationally convenient.
- Cluster sampling is generally less efficient than a sampling scheme, which selects units directly.

Example

Example

Suppose 2000 students are to be sampled from all students in higher education in the UK. A list of students may not be available. However, a list of higher education institutions can be constructed. We can take a sample these institutions and study their students.

Simple one-stage cluster sample: List all the clusters in the population, and from the list, select the clusters usually with simple random sampling. All units (elements) in the sampled clusters are selected for the survey.

Simple two-stage cluster sample: List all the clusters in the population. First, select the clusters, usually by simple random sampling (SRS). The units (elements) in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

Equal Cluster Sizes

- Suppose N clusters, each cluster consists of m individuals, hence Nm individuals in all.
- A random sample of n clusters is chosen. The sampling fraction is $f = \frac{n}{N}$.
- We may write the values by y_{ij} for cluster $i = 1, ..., n$ and elements $j = 1, ..., m$. The sample mean is

$$\overline{y}_{cl} = \frac{1}{mn} \sum_i \sum_j y_{ij} = \frac{1}{n} \sum_i \overline{y}_i.$$

- The variance of $\overline{y}_{cl}$ is

$$\frac{(1 - f)\, S_{cl}^2}{n},$$

which can be estimated by

$$\frac{1 - f}{n(n-1)} \sum_i \left(\overline{y}_i - \overline{y}_{cl}\right)^2 .$$

Exercise

To estimate the average number of newspapers purchased per household in a city, 2000 households were listed in 400 geographical clusters of 5 households each and a simple random sample of 4 clusters was selected given in the table below

Clusters	Number of Newspapers Purchased per Household				
1	1	2	1	3	3
2	1	3	2	2	3
3	2	1	1	1	1
4	1	1	3	2	1

calculate $\mathrm{Var}\left(\overline{y}_{cl}\right)$.

Unequal Cluster Sizes

- Suppose N clusters, which each cluster consists of m_i individuals, hence $M = \sum_i m_i$ individuals in all.
- A random sample of n clusters is chosen. The sampling fraction is $f = \dfrac{n}{N}$. Set $\overline{M} = \dfrac{M}{N}$.
- We may write the values by y_{ij} for cluster $i = 1, ..., n$ and elements $j = 1, ..., m_i$.
- Now

$$\overline{y}_{ucl} = \frac{1}{\sum_k m_k} \sum_i \sum_j y_{ij} = \frac{1}{\sum_i m_i} \sum_i m_j \overline{y}_i.$$

- We also have a ratio estimator for

$$\mathrm{Var}\left(\overline{y}_{ucl}\right) = \frac{(1-f) \sum_i \left(y_i - m_i \overline{y}\right)^2}{n \overline{M}^2 (n-1)},$$

which is biased unless cluster sizes are equal.

Exercise

To estimate the average number of newspapers purchased per household in a city, 2000 households were listed in 400 geographical clusters of households each and a simple random sample of 4 clusters was selected given in the table below

Clusters	Number of Newspapers Purchased per Household						
1	1	2	1	3			
2	1	3	2	2	3		
3	2	1	1				
4	1	1	3	2	1	1	1

calculate $\mathrm{Var}\left(\overline{y}_{ucl}\right)$.

Remarks and Summary

Remark

- Strata should be as homogeneous as possible within, but different as much as possible from one another.
- Clusters should be as heterogeneous as possible within, but very much like each other.

Today we reviewed the following concepts...

- Cluster sampling and estimation.
- Equal and unequal cluster sizes.

Week 9
Systematic Sampling

Intended Learning Outcomes

In this chapter you learn about

1. Systematic sampling methods.

2. Estimation of parameters.

Motivation

- Simple random sampling is the ”conceptually” easiest way of sampling as well as being at the basis of the sampling theory.
- However, it can be expensive and sometimes not feasible.
- Systematic sampling is a method which is easier to implement and is widely used in practice.
- The method ensures that each unit has equal probability of being selected.
- The method is used if more units are registered as time progresses.

Linear Systematic Sampling

- Suppose we have N units in our population and we want to select a sample of size n.
- We further assume that $k = \frac{N}{n}$ is an integer.
- Select a random unit r from the first 1 to k units. The is called the **random start**.
- Select every k^{th} unit after r, i.e.,

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

- In this was have k possible systematic samples (recall with Random Simple Sampling we have $\binom{N}{n}$ samples), so the chance selecting a sample is $\frac{1}{k}$.
- The integer k is known as the **sampling interval**.

Linear Systematic Sampling

The systematic sample has a better spread over the population. Depending on ordering, large part will not fail to be represented in the sample. Systematic sampling fails in case of too many blanks.

Example

As part of a cost-containment and quality-of-care review program, a sample of inpatient medical records is selected on an ongoing basis for a detailed audit. The total number of records in the population is not likely to be known in advance of the sampling since the records are to be sampled on an ongoing basis, and as a result, it would not be possible to use simple random sampling to choose the records. However, it may be possible to guess the approximate number of records that would be available for selection per time period and to select a sample of one in every k records as they become available.

Population Parameters Estimators

- We can write y_{ij} for the observation from unit $i + (j-1)k^{\text{th}}$ so $i = 1, ..., k$ and $j = 1, ...n$.
- The sample mean for random start i is then

$$\overline{y}_{sy} = \overline{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}.$$

- The mean of the population is given by

$$\overline{Y} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_i.$$

Theorem

The random variable $\overline{y}_{sy}$ is an unbiased estimator of $\overline{Y}$.

Variance of $\overline{y}_i$

Note we have

$$\begin{aligned}\operatorname{Var}(\overline{y}_{st}) &= \frac{1}{k}\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n^2}\left(\sum_{j=1}^{n} y_{ij} - \overline{Y}\right)^2 \\ &= \cdots \\ &= \frac{\sigma_y^2}{n}\left(1+(n-1)\rho\right),\end{aligned}$$

where

$$\rho = \frac{1}{kn(n-1)\sigma_y^2}\left(\sum_{i=1}^{k}\sum_{\substack{j=1\\ j\neq l}}^{n}\sum_{l=1}^{n}(y_{ij}-\overline{Y})(y_{il}-\overline{Y})\right).$$

Remarks and Summary

- It turns out that we cannot get an unbiased estimator of $\mathrm{Var}\left(\overline{y}_i\right)$ from a single systematics sample.
- In practice we estimate $\mathrm{Var}\left(\overline{y}_i\right)$ by

$$(1-f)\frac{s_y^2}{n}.$$

Today we reviewed the following concepts...

- Systematics sampling procedures.
- Estimation of parameters.

Week 10
Systematic Sampling Comparison

Intended Learning Outcomes

In this chapter you learn about

1. Comparison of systematic sampling with simple random sampling and stratified procedure.
2. Relative efficiencies.

Recap

- Recall we can write y_{ij} for the observation from unit $i+(j-1)k^{\text{th}}$ so $i=1,...,k$ and $j=1,...n$.
- The sample mean for random start i is then

$$\overline{y}_{sy} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \text{ and } \mathrm{Var}\left(\overline{y}_{st}\right) = \frac{\sigma_y^2}{n}\left(1+(n-1)\rho\right).$$

- Recall also with

$$\mathrm{Var}\left(\overline{y}_{sw}\right) = \frac{\sigma_y^2}{n} \text{ and } \mathrm{Var}\left(\overline{y}_{swo}\right) = \frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right).$$

- Note depending on the sign of ρ we get better or worse estimator for $\overline{y}_{st}$.
- However ρ cannot be too small as a negative number.

Comparison with SRSWOR $\text{Var}\left(\overline{y}_{sy}\right)$

Theorem

We have

$$\text{Var}\left(\overline{y}_{swo}\right) - \text{Var}\left(\overline{y}_{sy}\right) = \frac{n-1}{n}\left(S_{sy}^2 - S^2\right)$$

where

$$S_{sy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{y}_i\right)^2$$

and

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{Y}\right)^2 .$$

For the relative efficiency of $\overline{y}_{swo}$ and $\overline{y}_{swo}$ we find

$$\frac{\text{Var}\left(\overline{y}_{swo}\right)}{\text{Var}\left(\overline{y}_{sy}\right)} = \frac{n(k-1)}{(nk-1)\left(1+\rho(n-1)\right)}.$$

Different Form of Var $(\overline{y}_{st})$

Note we have

$$
\begin{aligned}
\sigma_y^2 &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{Y}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i+\overline{y}_i-\overline{Y}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+2\left(y_{ij}-\overline{y}_i\right)\left(\overline{y}_i-\overline{Y}\right)+\left(\overline{y}_i-\overline{Y}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+\frac{2}{N}\sum_{i=1}^{k}\left(\overline{y}_i-\overline{Y}\right)\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)+\frac{n}{N}\sum_{i=1}^{k}\left(\overline{y}_i-\overline{Y}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+\frac{1}{k}\sum_{i=1}^{k}\left(\overline{y}_i-\overline{Y}\right)^2 \\
&= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\overline{y}_i\right)^2+\mathrm{Var}\left(\overline{y}_{st}\right).
\end{aligned}
$$

Comparison with SRSWOR $\mathrm{Var}\left(\overline{y}_{sy}\right)$

So we find

$$\mathrm{Var}\left(\overline{y}_{sy}\right) = \sigma_y^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_i\right)^2 .$$

Now We have

$$\mathrm{Var}\left(\overline{y}_{swo}\right) - \mathrm{Var}\left(\overline{y}_{sy}\right) = \frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right) - \sigma_y^2 + \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_i\right)^2 .$$

Comparison with Stratified Sampling

Considering the set up of stratified sample in the set up of systematic sample, we have

- Number of strata n
- Size of strata k (row size)
- Sample size to be drawn from each stratum is 1.

In this case we find

$$\operatorname{Var}\left(\overline{y}_{st}\right) = \frac{N-n}{Nn} S_{st}^2$$

where

$$S_{st}^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{y}_j\right)^2 .$$

In this case we find

$$\operatorname{Var}\left(\overline{y}_{sy}\right) - \operatorname{Var}\left(\overline{y}_{st}\right) = \frac{N-n}{Nn} \left(n-1\right) \rho' S_{st}^2.$$

Remarks and Summary

Today we reviewed the following concepts...

- Comparison of systematic sampling with stratified and simple random sampling with/without replacement.

Week 11
Nonsampling Errors

Intended Learning Outcomes

In this chapter you learn about

1. Sampling vs Nonsampling errors.
2. Sources of nonsampling errors.
3. Controlling nonsampling errors.
4. Treatment of nonsampling errors.

Introduction: Sampling Errors

- We have learnt about enumeration and sample surveys in order to estimate unknown population parameters.
- None of these methods produce an exact estimate.
- In sample survey we draw inference by observing part of the population.
- The error introduced in this way is called **sampling errors**, which is not present in census.

Introduction: Nonsampling Errors

- When we collect information from a unit in a sample survey or complete enumeration, the information regarding the value of the characteristic under study is not free from error.
- The combination of all the kinds of errors, other than the sampling error, for which one cannot obtain the true value of the parameter by conducting a survey is known as the **nonsampling errors.**
- The nonsampling error is present in both the sample survey and complete enumeration.
- In general, the sampling errors decrease as the sample size increases whereas non-sampling error increases as the sample size increases.

Sources of nonsampling errors

The sources of nonsampling errors are numerous.

1. **Inappropriate sampling frame:** Completely specify the coverage of the survey and prepare a list of all the units of the target population.
2. **Response error/measurement error:** Response errors constitute wrong values obtained from the respondents. Response errors occur when the design of questionnaire is inappropriate.
3. **Error in data processing:** Errors can also be committed while entering data into the computer.
4. **Nonresponse error:** Nonresponse means failure to gather information on all or some of the items in a schedule or questionnaire. They mainly occur for two reasons: noncoverage and nonresponse. If all the items of information of a questionnaire or schedule are missing, we call it unit nonresponse, and if information on some of the items is obtained and on the rest is missing, then we call this item

Controlling Nonsampling Errors I

Some measures to take.

- The sampling frame should be accurate.
- Every unit of the population should be easily identifiable.
- Definitions of each of the items of investigations such as households and family should be provided unambiguously.
- Investigators should be provided with proper training, explaining the meaning and possible answers to each question they might ask.
- There should be proper supervision of enumerators to ensure collection of quality data. The questionnaire should be designed properly.
- The order of questioning should follow a logical sequence with easy and nonsensitive questions coming first to make the respondent comfortable.

Controlling Nonsampling Errors II

More measures to take.

- Related questions should be grouped together. The length of the questionnaire should be as short as possible.
- Questionnaires should have a method of consistency check whenever possible.
- Before conducting a survey, a pilot survey should be undertaken to test the questionnaire, time taken to administer it, cost of the proposed survey, administration of the survey, etc.
- The analysis of the questionnaire will help in checking the accuracy of the tables and other anticipated results.
- After conducting the survey, the filled questionnaires should be scrutinized properly. At least 5% of the collected data should be recollected again by different investigators to check the quality of data collected.

Treatment of Nonsampling Errors

- In almost all large-scale surveys, nonresponse is inevitable.
- Nonresponse creates bias in the estimates.

Poststratification

The set of responded and nonresponded units will be denoted, respectively, by s_1 and s_2. A simple random subsample s_2^* of size m is surveyed using more intensive method, which is obviously expensive.

Treatment of Nonsampling Errors

Use of Response Probabilities

Probabilistic models are proposed to describe the unknown response distributions.

Politz and Simmons Method

method of finding response probabilities of the sampled households selected by simple random sampling with replacement. More details on Survey Sampling Theory and Applications, R. Arnab, Chapter 15 [Arn17].

Treatment of Nonsampling Errors

Imputation

Imputation is used for item nonresponse. We assign one or more values to a missing item to reduce the bias and control variance due to nonresponse. Finally, a single estimate and its standard error are obtained by combining all these separate estimates.

- **Deductive imputation:** Here, the missing value is imputed through a consistency check - establishing relationship with other available items.
- **Substitution:** In this case, the missing information is replaced by a unit nearest to it.
- **Cold deck imputation:** The missing data are replaced using records from a recent past survey.
- **Hot deck imputation:** The sampled units are divided into classes using prior information.
- **Mean, Ratio and, Regression imputations.**

Summary

Today we reviewed the following concepts...

- Sampling va Nonsampling errors.
- Methods of controlling and treatment of Nonsampling errors.

References

[Arn17] R. Arnab. *Survey Sampling Theory and Applications.* Elsevier Science, 2017.

[Ric06] John A. Rice. *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition, 2006.

See You Next Time

Please Do Not Forget To

- Ask any **questions** now or through my contact details.
- Drop me **comments** and **feedback** relating to any aspects of the course.
- **My office hours** are on Mondays 15:00-16:00 & Fridays 11:00-12:00.
  Alternative: open door policy or email to make an appointment.

Thank You!