

Lecture 2 Sampling Theory

Last Lecture Recap

Introductory Remarks (Ch 1)

Finite Populations and Samples

Important Basic Sampling Designs

Simple Random Sampling

Estimation of Population Mean and Total

Review of Some Properties of the Sample Mean

The sample y_1, y_2, \dots, y_n

The sample mean $\bar{y} = \sum_{i=1}^n y_i / n$

$$E(\bar{y}) = \mu$$

$$\text{Var}(\bar{y}) = \sigma^2 / n$$

$$SE(\bar{y}) = \sigma / \sqrt{n}$$

Replace σ by s in estimates.

Modifications Needed in Finite Popns

CAPTURE-RECAPTURE MODELS

LINCOLN-PETERSEN MODEL

N - Population size

n₁ - No. of marked animals in the population

n₂ - Sample size

m₂ - No. of marked animals in the sample

Sample

Population

$$(m_2/n_2) \approx (n_1/N)$$

CAPTURE-RECAPTURE MODELS

LINCOLN-PETERSEN MODEL

$$(m_2/n_2) \approx (n_1/N)$$

$$\hat{N} = n_1 n_2 / m_2$$

THE SAMPLING IS MODEL BASED: MODEL ASSUMPTIONS

1. Closure-Population is constant
2. Equal Catchability-All animals equally catchable in each sample and there is no trap response.
3. Zero Mark Loss (Marking is definitive)

Note-We will see that Model based sampling is used often

All Sampling Design Topics

Basic Sampling Designs

Simple Random Sampling (2-5)

Unequal Probability Sampling Estimators

Hansen-Hurwitz (6)

Horvitz-Thompson (6)

Use of Auxiliary Information (increases precision)

Ratio Estimators (7)

Regression Estimators (8)

Advanced Sampling Designs

Stratified Random Sampling (11)

Systematic Random Sampling (12)

Cluster Sampling (12)

Multi-Stage Sampling (13)

Double Sampling (14)

Ch 1. Introductory Information and Definitions

- ◆ Basic Ideas
- ◆ Sampling Units
- ◆ Sampling and Non Sampling Errors
- ◆ Models in Sampling

Ch 1. Introductory Information and Definitions

- ◆ Basic Ideas
- ◆ Sampling Units
- ◆ Sampling and Non Sampling Errors
- ◆ Models in Sampling

Sampling Design Definitions

◆ Population

$U = \{1, 2, 3, \dots, N\}$ Population of N elements

The Universe is a list of all members of the population and is called a frame.

◆ Examples

Survey using list of all NC Licensed Hunters

Survey using list of all NC Licensed Dentists

Survey using list of all NC Registered Voters

Sampling Design Definitions

- ◆ **Population**- consists of N sampling units with N Known and Finite. Sampling units are labeled $1, 2, 3, \dots, N$. To begin assume that the sampling units are all of the same size. Think of them as respondents in a survey.
- ◆ **Sample** –consists of n sampling units. Drawn without replacement by some probabilistic method like simple random or stratified random sampling.

Sampling Design Definitions

- ◆ Variables of interest- e.g survey response like income (continuous) or opinion (usually binary) are measured on the individual sampling units (y_i)
- ◆ Population parameters like Population Mean (μ) or Population Total (τ) can be estimated from the sample values.

Probability Sampling

We need to use probability based sampling methods because they allow us to be able to study the properties of the samples we collect.

A probability based sampling procedure is one where we know the probability of drawing each sample.

We discourage use of convenience samples or purposive samples because there we do not know their properties.

Assumptions of Initial Theory

- ◆ **Sampling Frame is Complete.** For example, no names are missing and none are duplicated.
- ◆ **When the sample is drawn the y 's can be measured without error.** That is there is no nonresponse or inaccurate response if this a sample from a survey
- ◆ **Implication**
If sample size n goes to N this means that we have a perfect census! In this case the true parameters are known.

Ch 1. Introductory Information and Definitions

- ◆ Basic Ideas
- ◆ Sampling Units
- ◆ Sampling and Non Sampling Errors
- ◆ Models in Sampling

Sampling Units

- ◆ Sometimes clear cut what the sampling unit is- i.e. a licensed hunter.
- ◆ Sometimes less clear. Suppose one is sampling area as in an agricultural survey then one would have plot units and have to decide how large to make them. Similarly with volume in a water quality survey.

Ch 1. Introductory Information and Definitions

- ◆ Basic Ideas
- ◆ Sampling Units
- ◆ Sampling and Non Sampling Errors
- ◆ Models in Sampling

Sampling and Nonsampling Errors

- ◆ Sampling Error

Due to the fact not all the population is sampled for cost reasons. Measured by the standard error of the estimator

- ◆ Non Sampling Errors

 - Frame Errors

 - Nonresponse Errors

 - Response Errors

Lets Discuss and make a list of reasons for each kind.

Ch 1. Introductory Information and Definitions

- ◆ Basic Ideas
- ◆ Sampling Units
- ◆ Sampling and Non Sampling Errors
- ◆ Models in Sampling

Models in Sampling

- ◆ We already saw the capture-recapture sampling problem which used a model.
- ◆ Later we will consider ratio and regression estimators which use regression models you have seen in some of your other classes

Ch 2. Simple Random Sampling

We begin by considering simple random sampling without replacement which is the simplest probability sampling method. In this case each of the distinct possible samples has the same probability.

It is analogous to using a completely random design in experimental design.

It is most useful if our sampling units are homogeneous.

Now Consider the Simplest Possible Problems

Simple Random Sampling: Estimation

Now we consider the estimation of the finite population parameters.

1. Estimate the Population Mean (μ)
2. Estimate the Population Total (τ)

Simple Random Sampling: Estimation of the Population Mean

Represent the Population $\{y_1, y_2, \dots, y_N\}$

Population Mean (Finite Population) - Parameter

$$\mu = \frac{\sum_{i=1}^N y_i}{N} \quad (\tau = \sum_{i=1}^N y_i, \text{later})$$

Represent the Sample by $\{y_1, y_2, \dots, y_n\}$

Sample Mean - Estimate of the Parameter

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

\bar{y} is an estimate of μ .

Simple Random Sampling: Properties of Sample Mean as an Estimate of the Popn Mean

First

\bar{y} is an Unbiased Estimate of μ .

Second

What is the SE of \bar{y} ?

Simple Random Sampling: Standard Error of Sample Mean

Standard Result from say ST 311

Applies to Infinite Populations

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

where

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

Simple Random Sampling: Standard Error of Sample Mean

Standard Result

$$\text{SE of } \bar{y} = \frac{s}{\sqrt{n}}$$

Finite Population Result

$$\text{SE of } \bar{y} = \frac{s}{\sqrt{n}} \left(\sqrt{\frac{N-n}{N}} \right)$$

Note : The second term is called the finite population correction factor.

Simple Random Sampling: Another Problem-Estimate Population Total

Population Total (Parameter)

This is a very important new parameter in finite populations

$$\tau = \sum_{i=1}^N y_i = N\mu$$

Population Total Estimate

$$\hat{\tau} = N\bar{y}$$

$$\text{SE}(\hat{\tau}) = N \text{SE}(\bar{y})$$

$$= \frac{s}{\sqrt{n}} (\sqrt{N(N-n)})$$

Simple Random Sampling

- ◆ This is a summary of the key results. I will illustrate and expand in class.
- ◆ Will conclude class with a very simple example

Simple Random Sampling

Cottontail Rabbit Example

Finite Population:

There is a set of $N = 1000$ sampling units in the population.

$n = 100$ sampling units (plots) in the sample.

The sampling units are chosen by simple random sampling without replacement.

There are 1600 total animals counted in the 100 plots in the sample.

$$\hat{\mu} = \bar{y} = \sum_{i=1}^n y_i / n = 1600 / 100 = 16,$$

Note : Suppose that $s^2 = 40$

The population total estimate is:

$$\hat{\tau} = N \bar{y} = 1000 \times 16 = 16000$$

We are assuming all animals are detected for the moment.

Simple Random Sampling

Cottontail Rabbit Example P64 Williams et al.(2002)

$$\hat{\tau} = N \bar{y} = 1000 \times 16 = 16,000$$

$$\begin{aligned} \text{Var}(\hat{\tau}) &= N^2 \text{Var}(\bar{y}) = N(N-n) \frac{s^2}{n} \\ &= 1000(1000-100) \frac{40}{100} = 360,000 \end{aligned}$$

$$SE(\hat{\tau}) = \sqrt{\text{Var}(\hat{\tau})} = 600$$

$$95\% \text{ CI } \hat{\tau} \pm 1.96 \times SE(\hat{\tau})$$

$$16,000 \pm 1.96 \times 600$$

$$16,000 \pm 1,176$$

Text Book Example

Please read very carefully the textbook example on P 16-17 Section 2.3.

Simple Random Sampling

- Is this a Reasonable Design to Use is a Key Question? Populations often have a lot of heterogeneity in which case Stratified Random Sampling is better. Future Lecture.
- Also sometimes in spatial sampling Systematic Random Sampling may be better. Briefly now and also in future Lectures.
- Sample Size Calculation – How large a sample do we need to take is another key design issue?? Future Lecture

Systematic Random Sampling : Aside

Systematic random sampling sometimes gives a better spatial coverage than simple random sampling. Here is an example.

- Think of sampling along a transect of length 100 meters where you start at a random point in first 10 m (7 meters from Excel) and then every 10th meter. The systematic random sample will be

7,17,27,37,47,57,67,77,87,97

- I also chose a completely random sample of n=10 using Excel

18,20,33,59,63,85,90,91,92,96

- Notice clumping along transect, random does not mean uniform!!!

NOTE: Only major danger of systematic random sampling would be if there is some cyclical pattern of response along the transect. This does sometimes happen

Brief Summary so far: Simple Random Sampling

Brief Summary so far: Simple Random Sampling

N sampling units in population, **n** in the sample. Good if the population is **homogeneous**.

The key result is the **finite population correction factor** on the variance and standard error of the sample mean due to the sampling being without replacement.

Applications to surveys of human popns, area surveys etc

Modify to **Systematic Random Sampling** sometimes and we consider later

Later we will move onto considering how to sample heterogeneous populations using **stratified random sampling** to improve the precision over simple random sampling.

Homework Set 1 Due Next Tuesday

- ◆ Q2 Ch2.
- ◆ Q3 Ch2.
- ◆ Q2 Ch3.
- ◆ Q1 Ch 5.