Math 408 - Mathematical Statistics

Lecture 12. Introduction to Survey Sampling

February 15, 2013

Agenda

- Goals of Survey Sampling
- Population Parameters
- Simple Random Sampling
- Estimation of the population mean
- Summary

Survey Sampling

Sample surveys are use to obtain information about a large population.
The purpose of **survey sampling** is to reduce the cost and the amount of work that it would take to survey the entire population.

By a small sample
we may judge of the whole piece

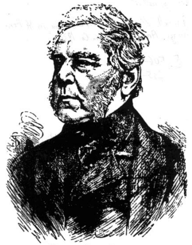
Miguel de Cervantes
"Don Quixote"

Familiar Examples of Survey Sampling:

- the cook in the kitchen taking a spoonful of soup to determine its taste
- the brewer needing only a sip of beer to test its quality

History of Survey Sampling

The first known attempt to make statements about a population using only information about part of it was made by the English merchant John Graunt. In his famous tract (Graunt, 1662) he describes a method to estimate the population of London based on partial information. John Graunt has frequently been merited as the founder of demography.


Captain John Graunt

The second time a survey-like method was applied was more than a century later. Pierre Simon Laplace realized that it was important to have some indication of the accuracy of the estimate of the French population (Laplace, 1812).

Recommended Reading: "The rise of survey sampling," by J. Bethlehem (2009).

Survey Sampling: Population Parameters

Suppose that the target population is of size N (N is very large) and a numerical value of interest x_i is associated with i^{th} member of the population, $i = 1, \ldots, N$.

Examples:

- x_i = age, weight, etc.
- $x_i = 1$ if some characteristic is present, and $x_i = 0$ otherwise.

There are two "standard" parameters of population that we are typically interested:

Definition

- **Population mean**

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Population variance**

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Simple Random Sampling

Important Remark:

Note that μ and σ^2 are not random. They are some fixed unknown parameters. We want to estimate them by picking n out of N members of the population and constructing estimates of μ and σ^2 based only on these n members.

The most elementary form of sampling from a population is **simple random sampling**.

Definition

In Simple Random Sampling, each member is chosen entirely by chance and, therefore, each member has an equal chance of being included in the sample; each particular sample of size n has the same probability of occurrence.

Let $X_1, \ldots, X_n$ be the sample drawn from the population.

Important Remark: Each X_i is a random variable:

- X_i is the value of the i^{th} element of the sample that was randomly chosen from the population
- x_i is the value of the i^{th} member of the population

Estimate

We will consider the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an **estimate** of the population mean μ. Since X_i are random, $\overline{X}_n$ is also random. Distribution of $\overline{X}_n$ is called its sampling distribution.The sampling distribution of $\overline{X}_n$ determines how accurately $\overline{X}_n$ estimates μ: the more tightly the sampling distribution is centered on μ, the better the estimate.

Our goal: is to investigate the sampling distribution of $\overline{X}_n$

Since $\overline{X}_n$ depends on X_i, let us start with examining the distribution of a single sample element X_i.

Basic Lemma

Lemma

Denote the distinct values assumed by the population members by $\xi_1, \dots, \xi_m$, $m \leq N$, *and denote the number of population members that have the value* ξ_i *by* n_i. *Then* X_i *is a discrete random variable with probability mass function*

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N} \tag{1}$$

Also

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2 \tag{2}$$

$\overline{X}_n$ is an unbiased estimator of μ

Theorem

With simple random sampling,

$$\mathbb{E}[\overline{X}_n] = \mu \tag{3}$$

This result can be interpreted as follows: "on average" $\overline{X}_n = \mu$

Definition

Suppose we want to estimate a parameter θ by a function $\hat{\theta}$ of the sample $X_1, \ldots, X_n$,

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

The estimator $\hat{\theta}$ is called **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$

Thus, $\overline{X}_n$ is an unbiased estimator of μ

Summary

- Sample surveys are used to obtain information about a large population
- Population parameters: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- We use sample mean $\overline{X}_n$ to estimate the population mean μ.
  - μ is unknown fixed parameter
  - $\overline{X}_n$ is random
- Properties of the sample element X_i:

$$\mathbb{P}(X_i = \xi_j) = \frac{n_j}{N} \qquad \mathbb{E}[X_i] = \mu \qquad \mathbb{V}[X_i] = \sigma^2$$

- $\overline{X}_n$ is an unbiased estimator of μ

$$\mathbb{E}[\overline{X}_n] = \mu$$

- Our next goal is to study the sampling distribution of $\overline{X}_n$.