Lecture 22. Survey Sampling: an Overview

March 25, 2013
Survey Sampling: What and Why

In surveys sampling we try to obtain information about a large population based on a relatively small sample of that population.

The main goal of survey sampling is to reduce the cost and the amount of work that it would take to explore the entire population.

First examples: Graunt (1662) and Laplace (1812) used survey sampling to estimate the population of London and France, respectively.

Mathematical Framework

Suppose that the target population is of size $N$ ($N$ is large) and a numerical value of interest $x_i$ (age, weight, income, etc) is associated with $i^{th}$ member of the population, $i = 1, \ldots, N$. Population parameters (quantities we are interested in):

- **Population mean**

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

- **Population variance**

  $$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$
There are several ways to sample from a population. We discussed two:

1. **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.

   If \( X_1, \ldots, X_n \) is the sample drawn from the population, then the sample mean is a natural estimate of the population mean \( \mu \):

   \[
   \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu
   \]

2. **Stratified Random Sampling**

   **Definition**
   In Stratified Random Sampling, the population is partitioned into subpopulations, or strata, which are then independently sampled using simple random sampling.

   If \( X_1^{(k)}, \ldots, X_{n_k}^{(k)} \) is the sample drawn from the \( k^{th} \) stratum, then the natural estimate of \( \mu \) is

   \[
   \bar{X}_n^* = \sum_{k=1}^{L} \omega_k \bar{X}_{n_k}^{(k)} \approx \mu
   \]
Statistical Properties of $\bar{X}_n$

Since $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, statistical properties of $\bar{X}_n$ are completely determined by statistical properties of $X_i$.

**Lemma**

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

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

Also

$$E[X_i] = \mu \quad \text{and} \quad \text{Var}[X_i] = \sigma^2$$

From this lemma, it follows immediately that $\bar{X}_n$ is an unbiased estimate of $\mu$:

$$E[\bar{X}_n] = \mu$$

Thus, on average $\bar{X}_n = \mu$. 
Statistical Properties of $\bar{X}_n$

The next important question is how variable $\bar{X}_n$ is.

As a measure of the dispersion of $\bar{X}_n$ about $\mu$, we use the standard deviation of $\bar{X}_n$, denoted as $\sigma_{\bar{X}_n} = \sqrt{\text{Var}[\bar{X}_n]}$.

**Theorem**

*The variance of $\bar{X}_n$ is given by*

\[
\text{Var}[\bar{X}_n] = \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right)
\]

Important observations:

- If $n << N$, then

\[
\text{Var}[\bar{X}_n] \approx \frac{\sigma^2}{n} \quad \sigma_{\bar{X}_n} \approx \frac{\sigma}{\sqrt{n}}
\]

\[
\left(1 - \frac{n-1}{N-1}\right)
\]

is called finite population correction. This factor arises because of dependence among $X_i$. 
Statistical Properties of $\bar{X}_n$

$$\sigma_{\bar{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$  \hspace{2cm} (1)

- To double the accuracy, the sample size must be **quadrupled**.
- If $\sigma$ is small (the population values are not very dispersed), then a small sample will be fairly accurate. But if $\sigma$ is large, then a larger sample will be required to obtain the same accuracy.
- We can’t use (1) in practice, since $\sigma$ is unknown. To use (1), $\sigma$ must be estimated from sample $X_1, \ldots, X_n$.

At first glance, it seems natural to use the following estimate

$$\hat{\sigma}^2_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \approx \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

However, this estimate is **biased**.
Statistical Properties of $\bar{X}_n$

**Theorem**

The expected value of $\hat{\sigma}^2_n$ is given by

$$E[\hat{\sigma}^2_n] = \sigma^2 \frac{Nn - N}{Nn - n}$$

In particular, $\hat{\sigma}^2_n$ tends to underestimate $\sigma^2$.

**Corollary**

- An unbiased estimate of $\sigma^2$ is

  $$\hat{\sigma}^2_{n,\text{unbiased}} = \frac{Nn - n}{Nn - N} \hat{\sigma}^2_n$$

- An unbiased estimate of $\nabla[\bar{X}_n]$ is

  $$s^2_{\bar{X}_n} = \frac{\hat{\sigma}^2_n}{n} \frac{Nn - n}{Nn - N} \left(1 - \frac{n - 1}{N - 1}\right)$$
Normal Approximation to the Distribution of $\bar{X}_n$

So, we know that the sample mean $\bar{X}_n$ is an unbiased estimate of $\mu$, and we know how to approximately find its standard deviation $\sigma_{\bar{X}_n} \approx s_{\bar{X}_n}$.

Ideally, we would like to know the **entire distribution** of $\bar{X}_n$ (sampling distribution) since it would tell us everything about the accuracy of the estimation $\bar{X}_n \approx \mu$.

It can be shown that if $n$ is large, but still small relative to $N$, then $\bar{X}_n$ is approximately normally distributed

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma_{\bar{X}_n}^2) \quad \sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$

From this result, it is easy to find the probability that the error made in estimating $\mu$ by $\bar{X}_n$ is less than $\varepsilon > 0$:

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \varepsilon) \approx 2\Phi\left(\frac{\varepsilon}{\sigma_{\bar{X}_n}}\right) - 1$$
Confidence Intervals

Let $\alpha \in [0, 1]$

**Definition**

A $100(1 - \alpha)\%$ confidence interval for a population parameter $\theta$ is a random interval calculated from the sample, which contains $\theta$ with probability $1 - \alpha$.

**Interpretation:**

If we were to take many random samples and construct a confidence interval from each sample, then about $100(1 - \alpha)\%$ of these intervals would contain $\theta$.

**Theorem**

An (approximate) $100(1 - \alpha)\%$ confidence interval for $\mu$ is

$$(\overline{X}_n - z_{\frac{\alpha}{2}} \sigma_{\overline{X}_n}, \overline{X}_n + z_{\frac{\alpha}{2}} \sigma_{\overline{X}_n})$$

That is the probability that $\mu$ lies in that interval is approximately $1 - \alpha$

$$\Pr(\overline{X}_n - z_{\frac{\alpha}{2}} \sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}} \sigma_{\overline{X}_n}) \approx 1 - \alpha$$
Estimation of a Ratio

Suppose that for each member of a population, two values are measured:

\[ i^{th} \text{ member } \sim (x_i, y_i) \]

We are interested in the following ratio:

\[ r = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\mu_y}{\mu_x} \]

Let \( \begin{pmatrix} X_1 & \ldots & X_n \\ Y_1 & \ldots & Y_n \end{pmatrix} \) be a simple random sample from a population.

Then the natural estimate of \( r \) is

\[ R_n = \frac{\overline{Y}_n}{\overline{X}_n} \]

To obtain expressions for \( \mathbb{E}[R_n] \) and \( \mathbb{V}[R_n] \) we use the \( \delta \)-method.
The δ-method

The δ-method is developed to address the following problem:

**Problem**

Suppose that $X$ and $Y$ are random variables, and that $\mu_X, \mu_Y, \sigma^2_X, \sigma^2_Y$, and $\sigma_{XY} = \text{Cov}(X, Y)$ are known. The problem is to find $\mu_Z$ and $\sigma^2_Z$, where $Z = f(X, Y)$.

Using the Taylor series expansion to the first order:

$$Z = f(X, Y) \approx f(\mu) + (X - \mu_X) \frac{\partial f}{\partial x}(\mu) + (Y - \mu_Y) \frac{\partial f}{\partial y}(\mu), \quad \mu = (\mu_X, \mu_Y)$$

Therefore,

$$\mu_Z \approx f(\mu)$$

$$\sigma^2_Z \approx \sigma^2_X \left( \frac{\partial f}{\partial x}(\mu) \right)^2 + \sigma^2_Y \left( \frac{\partial f}{\partial y}(\mu) \right)^2 + 2\sigma_{XY} \frac{\partial f}{\partial x}(\mu) \frac{\partial f}{\partial y}(\mu)$$

To obtain a better approximation for $\mu_Z$, we can use the Taylor series expansion to the second order.
Approximations of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$

Using the $\delta$-method, we obtain

**Theorem**

The expectation and variance of $R_n$ are given by

\[
\begin{align*}
\mathbb{E}[R_n] &\approx r + \frac{1}{n} \left(1 - \frac{n-1}{N-1}\right) \frac{1}{\mu_x^2} (r\sigma_x^2 - \sigma_{xy}) & (2) \\
\mathbb{V}[R_n] &\approx \frac{1}{n} \left(1 - \frac{n-1}{N-1}\right) \frac{1}{\mu_x^2} (r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}) & (3)
\end{align*}
\]

In applications, population parameters $\mu_x, \sigma_x, \sigma_y, \sigma_{xy}$ are unknown. To compute the estimated values of $\mathbb{E}[R_n]$ and $\mathbb{V}[R_n]$, we use (2) and (3) together with

- $r \approx R_n$, $\mu_x \approx \overline{X}_n$
- $\sigma_x^2 \approx \hat{\sigma}_x^{2, \text{unbiased}} = \frac{N-1}{Nn-N} \sum_{i=1}^n (X_i - \overline{X}_n)^2$
- $\sigma_y^2 \approx \hat{\sigma}_y^{2, \text{unbiased}} = \frac{N-1}{Nn-N} \sum_{i=1}^n (Y_i - \overline{Y}_n)^2$
- $\sigma_{xy} \approx \frac{N-1}{Nn-N} \sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$
Stratified Random Sampling

In Stratified Random Sampling, a population is partitioned into strata, which are then independently sampled using simple random sampling.

If \( X_{1}^{(k)}, \ldots, X_{n_{k}}^{(k)} \) is the sample drawn from the \( k^{\text{th}} \) stratum, then the estimate of \( \mu \) is

\[
\overline{X}_{n}^{*} = \sum_{k=1}^{L} \omega_{k} \overline{X}_{n_{k}}^{(k)} \approx \mu,
\]

where \( \omega_{k} = \frac{N_{k}}{N} \) is the fraction of the population in the \( k^{\text{th}} \) stratum.

- \( \overline{X}_{n}^{*} \) is an unbiased estimate of \( \mu \)

\[
\mathbb{E}[\overline{X}_{n}^{*}] = \mu
\]

- The variance of \( \overline{X}_{n}^{*} \) is

\[
\text{Var}[\overline{X}_{n}^{*}] = \sum_{k=1}^{L} \omega_{k}^{2} \frac{\sigma_{k}^{2}}{n_{k}} \left( 1 - \frac{n_{k} - 1}{N_{k} - 1} \right) \approx \sum_{k=1}^{L} \omega_{k}^{2} \frac{\sigma_{k}^{2}}{n_{k}}
\]
Question:
Suppose that the resources of a survey allow only a total of \( n \) units to be sampled. How to choose \( n_1, \ldots, n_L \) to minimize \( \text{Var}[\bar{X}_n^*] \) subject to constraint \( \sum n_k = n \)?

Optimization problem:
\[
\text{Var}[\bar{X}_n^*] \rightarrow \min \quad \text{s.t.} \sum_{k=1}^{L} n_k = n
\] (4)

Theorem
- The sample sizes \( n_1, \ldots, n_L \) that solve the optimization problem (4) are given by
  \[
  \hat{n}_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j} \quad k = 1, \ldots, L
  \]

- The variance of the optimal stratified estimate is
  \[
  \text{Var}[\bar{X}_{n,\text{opt}}^*] = \frac{1}{n} \left( \sum_{k=1}^{L} \omega_k \sigma_k \right)^2
  \]
Proportional Allocation

There are two main disadvantages of Neyman allocation:

1. Optimal allocations $\hat{n}_k$ depends on $\sigma_k$ which generally will not be known.
2. If a survey measures several values for each population member, then it is usually impossible to find an allocation that is simultaneously optimal for all values.

A simple and popular alternative method of allocation is proportional allocation: to choose $n_1, \ldots, n_L$ such that

\[
\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}
\]

This holds if

\[
\tilde{n}_k = n\frac{N_k}{N} = n\omega_k \quad k = 1, \ldots, L
\]  \hspace{1cm} (5)

Theorem

The variance of $\overline{X}_{n,p}^*$ is given by

\[
\text{Var}[\overline{X}_{n,p}^*] = \frac{1}{n} \sum_{k=1}^{L} \omega_k \sigma_k^2
\]
Neyman vs Proportional and Simple vs Stratified

By definition, Neyman allocation is always better than proportional allocation.

Question: When is it substantially better?

\[ \mathbb{V}[\bar{X}_{n,p}^*] - \mathbb{V}[\bar{X}_{n,opt}^*] = \frac{1}{n} \sum_{k=1}^{L} \omega_k (\sigma_k - \bar{\sigma})^2, \quad \bar{\sigma} = \sum_{k=1}^{L} \omega_k \sigma_k \]

- if the variances \( \sigma_k \) of the strata are all the same, then proportional allocation is as efficient as Neyman allocation, \( \mathbb{V}[\bar{X}_{n,p}^*] = \mathbb{V}[\bar{X}_{n,opt}^*] \)
- the more variable \( \sigma_k \), the more efficient the Neyman allocation scheme

Question: What is more efficient: simple random sampling or stratified random sampling with proportional allocation?

\[ \mathbb{V}[\bar{X}_n] - \mathbb{V}[\bar{X}_{n,p}^*] = \frac{1}{n} \sum_{k=1}^{L} \omega_k (\mu_k - \mu)^2 \]

Thus, stratified random sampling with proportional allocation always gives a smaller variance than simple random sampling does (providing that the finite population correction is ignored, \( (n - 1)/(N - 1) \approx 0 \)).