Math 408 - Mathematical Statistics

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 μ:

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

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 μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)} \approx \mu$$

Statistical Properties of $\overline{X}_n$

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

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

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

Also

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

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

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

Thus, on average $\overline{X}_n = \mu$.

Statistical Properties of $\overline{X}_n$

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

As a measure of the dispersion of $\overline{X}_n$ about μ, we use the standard deviation of $\overline{X}_n$, denoted as $\sigma_{\overline{X}_n} = \sqrt{\mathbb{V}[\overline{X}_n]}$.

Theorem

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

$$\boxed{\mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)}$$

Important observations:

- If $n << N$, then

$$\mathbb{V}[\overline{X}_n] \approx \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}}$$

$\left(1 - \frac{n-1}{N-1}\right)$ is called finite population correction. This factor arises because of dependence among X_i.

Statistical Properties of $\overline{X}_n$

$$\sigma_{\overline{X}_n} \approx \frac{\sigma}{\sqrt{n}} \tag{1}$$

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

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

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

However, this estimate is biased.

Statistical Properties of $\overline{X}_n$

Theorem

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

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

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

Corollary

- *An unbiased estimate of σ^2 is*

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

- *An unbiased estimate of $\mathbb{V}[\overline{X}_n]$ is*

$$s^2_{\overline{X}_n} = \frac{\hat{\sigma}_n^2}{n}\frac{Nn - n}{Nn - N}\left(1 - \frac{n-1}{N-1}\right)$$

Normal Approximation to the Distribution of $\overline{X}_n$

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

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

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

$$\overline{X}_n \dot{\sim} \mathcal{N}(\mu, \sigma^2_{\overline{X}_n}) \qquad \sigma_{\overline{X}_n} = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n-1}{N-1}}$$

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

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

Confidence Intervals

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

Definition

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

Theorem

An (approximate) $100(1-\alpha)\%$ *confidence interval for* μ *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 μ *lies in that interval is approximately* $1-\alpha$

$$\boxed{\mathbb{P}(\overline{X}_n - z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\sigma_{\overline{X}_n}) \approx 1-\alpha}$$

Estimation of a Ratio

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

$$i^{\text{th}} \text{ member} \rightsquigarrow (x_i, y_i)$$

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 $\begin{pmatrix} X_1 & \dots & X_n \\ Y_1 & \dots & Y_n \end{pmatrix}$ 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 δ-**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 $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$, and $\sigma_{XY} = Cov(X, Y)$ are known. The problem is to find μ_Z and σ_Z^2, where $Z = f(X, Y)$.

Using the Taylor series expansion to the first order:

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

Therefore,

$$\boxed{\mu_Z \approx f(\mu)} \qquad \boxed{\sigma_Z^2 \approx \sigma_X^2 \left(\frac{\partial f}{\partial x}(\mu)\right)^2 + \sigma_Y^2 \left(\frac{\partial f}{\partial y}(\mu)\right)^2 + 2\sigma_{XY}\frac{\partial f}{\partial x}(\mu)\frac{\partial f}{\partial y}(\mu)}$$

To obtain a better approximation for μ_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 δ-method, we obtain

Theorem

The expectation and variance of R_n are given by

$$\boxed{\mathbb{E}[R_n] \approx r + \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r\sigma_x^2 - \sigma_{xy})} \tag{2}$$

$$\boxed{\mathbb{V}[R_n] \approx \frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)\frac{1}{\mu_x^2}(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy})} \tag{3}$$

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

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

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^{th} stratum, then the estimate of μ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)} \approx \mu,$$

where $\omega_k = N_k/N$ is the fraction of the population in the k^{th} stratum.

- $\overline{X}_n^*$ is an unbiased estimate of μ

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

- The variance of $\overline{X}_n^*$ is

$$\mathbb{V}[\overline{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k} \left(1 - \frac{n_k - 1}{N_k - 1}\right) \approx \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k}$$

Neyman (=Optimal) Allocation Scheme

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 $\mathbb{V}[\overline{X}^*_n]$ subject to constraint $\sum n_k = n$?

Optimization problem:

$$\mathbb{V}[\overline{X}^*_n] \to \min \qquad \text{s.t.} \sum_{k=1}^{L} n_k = n \tag{4}$$

Theorem

- *The sample sizes $n_1, \ldots, n_L$ that solve the optimization problem (4) are given by*

$$\hat{n}_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j} \qquad k = 1, \ldots, L$$

- *The variance of the optimal stratified estimate is*

$$\mathbb{V}[\overline{X}^*_{n,opt}] = \frac{1}{n} \left(\sum_{k=1}^{L} \omega_k \sigma_k \right)^2$$

Proportional Allocation

There are two main disadvantages of Neyman allocation:

1. Optimal allocations $\hat{n}_k$ depends on σ_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

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

This holds if

$$\tilde{n}_k = n\frac{N_k}{N} = n\omega_k \qquad k = 1, \ldots, L \tag{5}$$

Theorem

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

$$\mathbb{V}[\overline{X}^*_{n,p}] = \frac{1}{n}\sum_{k=1}^{L}\omega_k\sigma_k^2$$

Neyman vs Proportional and Simple vs Stratified

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

Question: When is it substantially better?

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

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

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

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

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, $(n-1)/(N-1) \approx 0$).