Lecture 10. Sampling (Chapter 17)

Ping Yu

HKU Business School
The University of Hong Kong

Plan of This Lecture

- Stratified Sampling
- Other Sampling Methods
  - Cluster Sampling
  - Two-Phase Sampling
  - Nonprobabilistic Sampling Methods

Stratified Sampling

Stratified Sampling

- Due to resource constraints, you may only draw a random sample with a limited size n.
- Some subgroup of interest may have only a few samples, which limits the capability of inferences based on them.
- Although the estimator based on the few samples is still unbiased, unbiasedness requires repeated sampling among some of which the subgroup may be oversampled.
- You can also replace the original sample by a new one until the subgroup of interest has a large enough size, but this sampling process (which is definitely not random sampling) is hard to formalize such that the resulting sample is hard to analyze.
- An alternative better solution is to use stratified random sampling where particular identifiable characteristics are the subject of inquiry or particular subgroups are of special interest.
- Suppose that a population of N individuals can be partitioned into K groups (or strata), each with size N_j; stratified random sampling draws n_j independent random samples from each stratum j.
  - $N = \sum_{j=1}^{K} N_j$ and $n = \sum_{j=1}^{K} n_j$, where n_j need not be the same.

Analysis of Results from Stratified Random Sampling

- Let $\{\mu_j\}_{j=1}^K$ be the population means of K strata, and $\{\bar{x}_j\}_{j=1}^K$ be the corresponding sample means.
- Conditional on stratum j, since a random sample of size n_j from a population of size N_j was drawn without replacement, from Lecture 5,

$$\sigma^2_{\bar{x}_j} = Var\left(\bar{x}_j\right) = \frac{S_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j},$$

where $S_j^2 = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(x_{ji} - \mu_j\right)^2$ is the finite-population variance from stratum j, and

$$\hat{\sigma}^2_{\bar{x}_j} = \frac{s_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j}$$

is an unbiased estimator of $\sigma^2_{\bar{x}_j}$, where s_j^2 is the sample variance in stratum j.

- Note: The finite-population correction factor should be $\frac{N_j - n_j}{N_j}$ rather than $\frac{N_j - n_j}{N_j - 1}$ in the textbook because $E\left[s_j^2\right] = S_j^2$ rather than σ_j^2, but this does not matter in practice when N_j is large.

Estimation of the Population Mean

- Suppose we are interested in the population mean

  $$\mu = \frac{1}{N}\sum_{j=1}^{K} N_j \mu_j ;$$

  a natural unbiased estimator of μ is

  $$\bar{x}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j \bar{x}_j ,$$

  where the subscript *st* is from "**st**ratified", and N_j's are implicitly assumed known.
- Because the samples from different strata are independent, we conclude

  $$\hat{\sigma}^2_{\bar{x}_{st}} = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\bar{x}_j} = \frac{1}{N^2}\sum_{j=1}^{K} N_j \left(N_j - n_j\right) \frac{s_j^2}{n_j}$$

  is an unbiased estimator of $\sigma^2_{\bar{x}_{st}} = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2 \sigma^2_{\bar{x}_j} = \frac{1}{N^2}\sum_{j=1}^{K} N_j \left(N_j - n_j\right) \frac{S_j^2}{n_j}$.
- When n is large, we can construct the $(1-\alpha)$ CI for μ as

  $$\bar{x}_{st} \pm z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} .$$

Estimation of the Population Total

- If we are interested in the population total, $N\mu$, then an unbiased estimator is

$$N\bar{x}_{st} = \sum_{j=1}^{K} N_j \bar{x}_j.$$

- An unbiased estimator of $\sigma^2_{N\bar{x}_{st}}$ is

$$\hat{\sigma}^2_{N\bar{x}_{st}} = N^2 \hat{\sigma}^2_{\bar{x}_{st}} = \sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\bar{x}_j} = \sum_{j=1}^{K} N_j \left(N_j - n_j\right) \frac{s_j^2}{n_j}.$$

- When n is large, we can construct the $(1-\alpha)$ CI for $N\mu$ as

$$N\bar{x}_{st} \pm z_{\alpha/2} N \hat{\sigma}_{\bar{x}_{st}}.$$

Estimation of Population Proportion

- Let $\{p_j\}_{j=1}^K$ be the population proportion of a particular characteristic in the K strata, and $\{\hat{p}_j\}_{j=1}^K$ be the corresponding sample proportions.
  - Our analysis above can be applied if we define the population r.v. X as the indicator 1 (a particular characteristic).
- If we are interested in the population proportion (i.e., $E[X]$)

$$p = \frac{1}{N}\sum_{j=1}^K N_j p_j,$$

then a natural unbiased estimator is

$$\hat{p}_{st} = \frac{1}{N}\sum_{j=1}^K N_j \hat{p}_j.$$

- An unbiased estimator of $\sigma^2_{\hat{p}_{st}}$ is

$$\hat{\sigma}^2_{\hat{p}_{st}} = \frac{1}{N^2}\sum_{j=1}^K N_j^2 \hat{\sigma}^2_{\hat{p}_j} = \frac{1}{N^2}\sum_{j=1}^K N_j\left(N_j - n_j\right)\frac{\hat{p}_j\left(1-\hat{p}_j\right)}{n_j},$$

where $\hat{\sigma}^2_{\hat{p}_j} = \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}\cdot\frac{N_j-n_j}{N_j}$.
- When n is large, we can construct the $(1-\alpha)$ CI for p as

$$\hat{p}_{st} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}_{st}}.$$

Allocation of Sample Effort Among Strata: Proportional Allocation

- Given a fixed n, how to allocate n_j such that $\sum_{j=1}^{K} n_j = n$?
- If little or nothing is known beforehand about the population and if there are no strong requirements for the production of information about sparsely populated individual strata, a natural choice, often employed in practice, is proportional allocation: the proportion of sample members from any stratum is the same as the proportion of population members in that stratum, i.e.,

$$n_j = \frac{N_j}{N} \cdot n.$$

- Drawback: produce relatively few samples in strata of interest such that inferences on the population parameters of these particular strata are imprecise.

Optimal Allocation

- If the sole objective of a survey is to estimate as precisely as possible μ or $N\mu$, then it can be shown that the most precise estimators can be obtained by optimal allocation:
$$n_j = \frac{N_j S_j}{\sum_{j=1}^K N_j S_j} \cdot n. \qquad (1)$$
- Why? Since $\sigma^2_{N\bar{x}_{st}} = \sum_{j=1}^K N_j^2 \left(\frac{S_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j} \right)$, we try to solve the following minimization problem,
$$\min_{\{n_j\}_{j=1}^K} \sum_{j=1}^K N_j^2 \left(\frac{S_j^2}{n_j} \cdot \frac{N_j - n_j}{N_j} \right)$$
$$\text{subject to} \sum_{j=1}^K n_j = n,$$
whose solution is (1).
- Because $S_j^2 = \frac{N}{N-1}\sigma_j^2$, the optimal $n_j = \frac{N_j \sigma_j}{\sum_{j=1}^K N_j \sigma_j} \cdot n$ in the textbook.
- Compared with the proportional allocation, it allocates more samples to strata with higher population variances. Intuitively, more samples are required to bring down the variance of sample mean when the population variance is high.

continue

- Drawback: σ_j's are unknown or even no estimates of them exist.
  - We can use two-phase sampling to handle this drawback. Specifically, prior to conducting the bulk of the survey, first carry out an initial pilot study in which a relatively small proportion of the sample members are contacted and estimate σ_j's based on this pilot sample. Although time consuming, this sampling approach has at least three further advantages besides obtaining an estimate of σ_j's:
  1. The investigator is able, at modest cost, to try out the proposed questionnaire to ensure that the various questions can be thoroughly understood.
  2. The pilot study may also suggest additional questions whose potential importance have previously been overlooked.
  3. The pilot study can also provide an estimate of the likely rate of nonresponse (which should not be too high, or modify the method of soliciting responses).
- If we are interested in the population proportion, the optimal allocation is

$$n_j = \frac{N_j\sqrt{p_j\left(1-p_j\right)}}{\sum_{j=1}^{K} N_j\sqrt{p_j\left(1-p_j\right)}}\cdot n. \quad (2)$$

  - Compared with the proportional allocation, it allocates more samples to strata with p_j close to 0.5.
  - Like (1), (2) involves the unknown proportion p_j, the very quantity that the survey is designed to estimate.

Determining Sample Sizes for Specified Degree of Precision

- Recall that

$$\sigma^2_{\bar{x}_{st}} = \frac{1}{N^2}\sum_{j=1}^K N_j^2 \sigma^2_{\bar{x}_j} = \frac{1}{N^2}\sum_{j=1}^K N_j \left(N_j - n_j\right)\frac{S_j^2}{n_j} = \frac{1}{N^2}\sum_{j=1}^K N_j^2 \frac{\sigma_j^2}{n_j}\frac{N_j - n_j}{N_j - 1}, \quad (3)$$

  which is valid for any allocation of n_j.
- To achieve the desired variance for proportional allocation, we can let $n_j = \frac{N_j}{N} \cdot n$ in (3) and solve out n as a function of $\sigma^2_{\bar{x}_{st}}$ as

$$n = \frac{\sum_{j=1}^K N_j \sigma_j^2}{N\sigma^2_{\bar{x}_{st}} + \frac{1}{N}\sum_{j=1}^K N_j \sigma_j^2}. \quad (4)$$

  - In terms of S_j^2, $n = \frac{\sum_{j=1}^K (N_j - 1) S_j^2}{N\sigma^2_{\bar{x}_{st}} + \frac{1}{N}\sum_{j=1}^K (N_j - 1) S_j^2}$.
- Similarly, in optimal allocation, let $n_j = \frac{N_j \sigma_j}{\sum_{j=1}^K N_j \sigma_j} \cdot n$ in (3) and solve out n as a function of $\sigma^2_{\bar{x}_{st}}$ as

$$n = \frac{\frac{1}{N}\left(\sum_{j=1}^K N_j \sigma_j\right)^2}{N\sigma^2_{\bar{x}_{st}} + \frac{1}{N}\sum_{j=1}^K N_j \sigma_j^2}. \quad (5)$$

  - If σ_j^2 are equal for all j, then (5) reduces to (4).

Other Sampling Methods

Cluster Sampling

- Suppose we wants to survey a population spread over a wide geographical area, such as a large city or a state.
- Either random sampling or stratified sampling has two problems:

1. Either sampling requires accurate listing of the population members, which is often impossible.
2. Even if the list is available, the sample members usually thinly spread over a large area, so personal interviews are quite costly.
   - Using a mail questionnaire to reduce cost may result in a high rate of nonresponse, so personal interviews are still preferred.

- To solve these two problems, cluster sampling is proposed.
- This approach is attractive when a population can be subdivided into M relatively small, geographically compact[1] units called clusters. For example, a city might be subdivided into political wards or residential blocks, which is possible even when a complete listing of residents or households is unavailable.
- Then a simple random sample of m of these clusters is selected, and information is obtained from every member (although no need) of the sampled clusters.[2]

[1] This implies that personal interviews are relatively inexpensive.

[2] This implies that the population listing is not necessary.

Estimators for Cluster Sampling

- Let n_i be the numbers of population members in cluster i, $\bar{x}_i$ be the sample mean of cluster i and $\hat{p}_i$ the proportion possessing an attribute of interest in cluster i. We want to estimate the population mean μ and population proportion p.
- The unbiased estimator of μ is

$$\bar{x}_c = \frac{\sum_{i=1}^m n_i \bar{x}_i}{\sum_{i=1}^m n_i} = \frac{\sum_{i=1}^m x_i^t}{m\bar{n}},$$

where the subscript c is from "**c**luster", $x_i^t = \sum_{j=1}^{n_i} x_{ij}$ is the total sum of cluster i, and $\bar{n} = \frac{1}{m}\sum_{i=1}^m n_i$ is the average number of members in the sampled clusters.

- Since

$$Var\left(\frac{\sum_{i=1}^m x_i^t}{m}\right) = \frac{M-m}{M}\frac{S_{x_i^t}^2}{m},$$

where $S_{x_i^t}^2 = \frac{1}{M-1}\sum_{i=1}^M \left(x_i^t - \mu_{x_i^t}\right)^2$, and $\mu_{x_i^t} = E\left[x_i^t\right] = n_i\mu$ can be unbiasedly estimated by $n_i\bar{x}_c$, the unbiased estimator of $\sigma_{\bar{x}_c}^2 = Var\left(\bar{x}_c\right)$ is

$$\hat{\sigma}_{\bar{x}_c}^2 = \frac{M-m}{Mm\bar{n}^2}s_{x_i^t}^2 := \frac{M-m}{Mm\bar{n}^2}\frac{\sum_{i=1}^m \left(x_i^t - n_i\bar{x}_c\right)^2}{m-1} = \frac{M-m}{Mm\bar{n}^2}\frac{\sum_{i=1}^m n_i^2\left(\bar{x}_i - \bar{x}_c\right)^2}{m-1}.$$

continue

- Similarly, the unbiased estimator of p is
$$\hat{p}_c = \frac{\sum_{i=1}^m n_i \hat{p}_i}{\sum_{i=1}^m n_i},$$
and the unbiased estimator of $\sigma_{\hat{p}_c}^2 = Var\left(\hat{p}_c\right)$ is
$$\hat{\sigma}_{\hat{p}_c}^2 = \frac{M-m}{Mm\bar{n}^2} \frac{\sum_{i=1}^m n_i^2 \left(\hat{p}_i - \hat{p}_c\right)^2}{m-1}.$$
- Implicitly, we assume that M and $\{n_i\}_{i=1}^m$ are known, but we need not know N and all $\{n_i\}_{i=1}^M$.
- When n is large, we can construct the $(1-\alpha)$ CIs for μ and p as
$$\bar{x}_c \pm z_{\alpha/2} \hat{\sigma}_{\bar{x}_c} \text{ and } \hat{p}_c \pm z_{\alpha/2} \hat{\sigma}_{\hat{p}_c},$$
respectively.
- Drawback: compared with stratified sampling where a sample is taken from every stratum, cluster sampling takes a random sample of clusters so that some clusters will have no members in the sample. Since members in a cluster are fairly homogeneous, important subgroups of the population are either not represented at all or grossly underrepresented in the final sample, which results in imprecise sample estimates.

Nonprobabilistic Sampling Methods

- In the previous sampling schemes, the probability that any particular sample will be drawn is specified such that statistical inferences can be conducted.
- In practice, nonprobabilistic methods are used as a matter of convenience.
- Quota sampling by polling organizations is an example.
- Interviewers are assigned to a particular locale and instructed to contact specified numbers of people of certain age, race, and gender characteristics.
- These assigned quotas represent what are thought to be appropriate proportions for the population at large.
- However, once the quotas are determined, interviewers are granted flexibility in the choice of sample members. Their choice is typically not random.
- Drawback: since the sample is not chosen using probabilistic methods, there is no valid way to determine the reliability of the resulting estimates.