Probability Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 11, 2018

Here (and throughout most of the course) we will assume: the sampling frame population is the target population (no undercoverage or overcoverage); no nonresponse or missing data; and no measurement error.

For most of the course we assume there is no *nonsampling error*, so that we can focus our attention on the study of *sampling error*.

In a **probability sample** from a population of N units, each of the 2^N possible samples has a known selection probability, and thus each unit in the population has a known inclusion probability.

Types of probability samples

1. **Simple random sample** without replacement (SRS).

   Imagine an urn containing N balls, well mixed. Select n of them at random.

   Let

$$Z_i = \begin{cases} 1 & \text{unit } i \text{ in sample} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i = 1, \ldots, N$$

   The probability of any particular sample, that is, any collection $(z_1, \ldots, z_N)$ of 0's and 1's, is

$$p(z_1, \ldots, z_N) = \begin{cases} \binom{N}{n}^{-1} & \sum_{i=1}^{N} z_i = n \\ 0 & \text{otherwise} \end{cases}$$

2. **Stratified random sample**

   The population is divided into subgroups, called *strata*. A separate, independent SRS is taken within each *stratum*.

3. **Cluster sampling**

   The population is divided into subgroups, called *clusters*. Then a SRS of clusters is drawn.

   - In *one-stage cluster sampling*, we take a complete census in the selected clusters;

   - in *two-stage cluster sampling* we take separate SRSs in the selected clusters.

4. **Systematic sampling**

Given a list of the units, choose a starting point at random, then sample that unit and every kth unit after it.

Example 1: Suppose the population is $\{1, 2, \ldots, 100\}$

- Suppose 10 strata: $\{1, \ldots, 10\}, \{11, \ldots, 20\} \ldots, \{91, \ldots, 100\}$.
- Suppose 20 clusters: $\{1, \ldots, 5\}, \{6, \ldots, 10\} \ldots, \{96, \ldots, 100\}$.

1. In a SRS of $n = 20$, any of the $\binom{100}{20}$ possible samples has the same probability of being the sample selected.

2. Stratified sample: Pick 2 units from stratum.

   There are $\binom{10}{2}^{10}$ possible samples, all equally likely.

3. Cluster sample: Take a SRS of 4 of the 20 clusters.

   There are $\binom{20}{4}$ possible samples, all equally likely.

4. Systematic sample: Pick one of $\{1, 2, 3, 4, 5\}$ at random, sample that unit and every 5th unit thereafter.

   There are 5 possible samples, equally likely.

Framework for probability sampling

The population is $\mathcal{U} = \{1, 2, \ldots, N\}$.

Consider a probability sampling method with m different possible samples, denote

$$\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$$

Each $\mathcal{S}_j$ is a subset of $\mathcal{U}$, and

$$P(\mathcal{S}_1) + \cdots + P(\mathcal{S}_m) = 1$$

For each unit i let

$$\pi_i = P(\text{unit } i \text{ in sample}) = \sum_{j:i \in \mathcal{S}_j} P(\mathcal{S}_j)$$

Suppose that associated with the ith unit of the population is a numeric value y_i.

Sometimes we will write that the population is $\{y_1, y_2, \ldots, y_N\}$.

Suppose we want to estimate the population mean

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

using the sample mean

$$\bar{y}_{\mathcal{S}} = \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{S}} y_i$$

Definition: The **sampling distribution** of the statistic $\bar{y}$ is specified by the possible values $\bar{y}$ can take, and the probabilities it takes those values.

For a finite population this is a discrete distribution!

$$P(\bar{y} = k) = \sum_{\mathcal{S}:\bar{y}_{\mathcal{S}}=k} P(\mathcal{S})$$

Definition: The **expected value** of $\bar{y}$ is the mean of this sampling distribution,

$$E(\bar{y}) = \sum_{\mathcal{S}} \bar{y}_{\mathcal{S}} P(\mathcal{S}) = \sum_{k} k P(\bar{y} = k)$$

Example 2: Consider the following population of $N = 6$ units

i	1	2	3	4	5	6
y_i	9	12	14	15	12	12

The population mean is $\bar{y}_U = 12.33$.

Consider a sampling scheme of three possible samples of size $n = 4$, with $\mathcal{S}_1 = \{1, 2, 3, 4\}$ and $\mathcal{S}_2 = \{2, 3, 4, 5\}$ and $\mathcal{S}_3 = \{3, 4, 5, 6\}$, and suppose that $P(\mathcal{S}_j) = 1/3$ for $j = 1, 2, 3$.

The inclusion probabilities are

$$\pi_1 = \pi_6 = \frac{1}{3} \quad \text{and} \quad \pi_2 = \pi_5 = \frac{2}{3} \quad \text{and} \quad \pi_3 = \pi_4 = 1$$

The possible values of the sample mean are

$$\bar{y}_{\mathcal{S}_1} = 12.50 \quad \text{and} \quad \bar{y}_{\mathcal{S}_2} = \bar{y}_{\mathcal{S}_3} = 13.25$$

The expected value of the sample mean is

$$E(\bar{y}) = 12.50\left(\frac{1}{3}\right) + 13.25\left(\frac{2}{3}\right) = 13.00$$

Definition: The **estimation bias** of $\bar{y}$ as an estimator of $\bar{y}_U$ is

$$\text{Bias}(\bar{y}) = E(\bar{y}) - \bar{y}_U$$

Example 2: Bias$(\bar{y}) = 0.67$.

Definition: The **variance** of the estimator $\bar{y}$ is

$$V(\bar{y}) = E\left\{[\bar{y} - E(\bar{y})]^2\right\} = \sum_{\mathcal{S}} [\bar{y}_{\mathcal{S}} - E(\bar{y})]^2 \, P(\mathcal{S})$$

and the **mean squared error** is

$$\text{MSE}(\bar{y}) = E\left[(\bar{y} - \bar{y}_U)^2\right] = \sum_{\mathcal{S}} [\bar{y}_{\mathcal{S}} - \bar{y}_U)]^2 \, P(\mathcal{S})$$

Note that

$$\begin{aligned}\text{MSE}(\bar{y}) &= E\left[(\bar{y}-\bar{y}_U)^2\right] \\ &= E\left\{[\bar{y}-E(\bar{y})+E(\bar{y})-\bar{y}_U]^2\right\} \\ &= E\left\{[\bar{y}-E(\bar{y})]^2\right\}+[E(\bar{y})-\bar{y}_U]^2 \\ &= V(\bar{y})+[\text{Bias}(\bar{y})]^2\end{aligned}$$

The cross-product term is zero since $E\,[\bar{y}-\bar{y}_U]=0$

Example 2:

$$V(\bar{y})=\tfrac{1}{3}(12.50-13.00)^2+\tfrac{2}{3}(13.25-13.00)^2=0.125$$

and

$$\begin{aligned}\text{MSE}(\bar{y}) &= \tfrac{1}{3}(12.50-12.33)^2+\tfrac{2}{3}(13.25-12.33)^2 \\ &= 0.5694 \\ &= 0.125+(0.67)^2\end{aligned}$$