Simple Random Sampling

Survey Sampling
Statistics 4234/5234
Fall 2018

September 13, 2018

Consider a population that has N units. We will take a random sample consisting of n draws from this population.

In **simple random sampling with replacement** (SRSwR) we take n independent samples of size 1.

In **simple random sampling without replacement** (SRS), there are $\binom{N}{n}$ possible samples, and each is equally likely. Thus the probability of any particular set of n units is

$$P(\mathcal{S}) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

Under SRS the probability that the ith unit is included in the sample is

$$\pi_i = \frac{\text{number of samples including unit } i}{\text{total number of possible samples}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Denote the population values by $\{y_1, y_2, \ldots, y_N\}$.

The population mean is

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

and the population variance is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2$$

and the population standard deviation is $S = \sqrt{S^2}$.

Suppose we take a simple random sample of size n and compute the sample mean

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

Proposition: $E(\bar{y}) = \bar{y}_U$ and $V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$.

Proof: Define the random variables

$$Z_i = \begin{cases} 1 & \text{unit } i \text{ is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, N$.

Then

$$E(Z_i) = E(Z_i^2) = P(Z_i = 1) = \pi_i$$

and

$$V(Z_i) = E(Z_i^2) - (EZ_i)^2 = \pi_i - \pi_i^2 = \pi_i(1 - \pi_i)$$

and thus

$$E(Z_i) = \frac{n}{N} \quad \text{and} \quad V(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

Also, for $j \neq i$ we have

$$
\begin{aligned}
E(Z_iZ_j) = P(Z_iZ_j = 1) &= P(\text{both } i \text{ and } j \text{ in sample}) \\
&= \frac{\text{number samples include units } i \text{ and } j \text{ both}}{\text{total number of possible samples}} \\
&= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}
\end{aligned}
$$

and thus

$$
\begin{aligned}
\text{Cov}(Z_i, Z_j) &= E(Z_iZ_j) - (EZ_i)(EZ_j) \\
&= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\
&= -\frac{1}{N-1}\left(\frac{n}{N}\right)\left(1 - \frac{n}{N}\right)
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
E(\bar{y}) &= E\left(\frac{1}{n}\sum_{i\in\mathcal{S}} y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{n}\frac{n}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_U
\end{aligned}
$$

and

$$
\begin{aligned}
V(\bar{y}) &= V\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + 2\sum_{i=1}^{N-1}\sum_{j=1+1}^{N} y_i y_j \text{Cov}(Z_i, Z_j)\right] \\
&= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{2}{N-1}\sum_{i=1}^{N-1}\sum_{j=1+1}^{N} y_i y_j\right] \\
&= \frac{1}{nN}\left(1-\frac{n}{N}\right)\left(\frac{1}{N-1}\right) N \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 \quad \text{see Sec 2.8} \\
&= \frac{S^2}{n}\left(1-\frac{n}{N}\right)
\end{aligned}
$$

The factor $\left(1-\frac{n}{N}\right)$ is called the **finite population correction**.

- If $n = N$ then $V(\bar{y}) = 0$.

- If $n = 1$ then $V(\bar{y}) = \frac{S^2}{1}\left(1-\frac{1}{n}\right) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$

Of course in practice the population variance is unknown, and thus estimated by the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2$$

The estimated variance of $\bar{y}$ is

$$\widehat{V}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

and the standard error is

$$\text{SE}(\bar{y}) = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

Proposition: $E(s^2) = S^2$

Proof: (See Section 2.8)

$$
\begin{aligned}
E\left[\sum_{i\in\mathcal{S}}(y_i-\bar{y})^2\right] &= E\left\{\sum_{i\in\mathcal{S}}\left[(y_i-\bar{y}_U)-(\bar{y}-\bar{y}_U)\right]^2\right\} \\
&= E\left[\sum_{i=1}^{N} Z_i(y_i-\bar{y}_U)^2\right] - nE\left[(\bar{y}-\bar{y}_U)^2\right] \\
&= \frac{n}{N}(N-1)\sum_{i=1}^{N}(y_i-\bar{y}_U)^2 - n\frac{S^2}{n}\left(1-\frac{n}{N}\right) \\
&= \frac{n}{N}(N-1)S^2 - S^2\left(\frac{N-n}{N}\right) = (n-1)S^2
\end{aligned}
$$

Two final points about estimation under SRS

1. Want to estimate $t = \sum_{i=1}^{N} y_i = N\bar{y}_U$?

   Use the estimator $\hat{t} = N\bar{y}$. Follows immediately from the work above that

$$E\left(\hat{t}\right) = NE(\bar{y}) = N\bar{y}_U = t$$

   and

$$V\left(\hat{t}\right) = N^2 V(\bar{y}) = N^2 \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

   and so the standard error of the unbiased estimator $\hat{t}$ is

$$\mathrm{SE}(\hat{t}) = N\,\mathrm{SE}(\bar{y}) = N\frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

2. If we want to estimate the proportion of a population which possess some trait of interest, we let

$$y_i = \begin{cases} 1 & \text{unit } i \text{ has that trait} \\ 0 & \text{otherwise} \end{cases}$$

and proceed as above.

In this situation we will often employ specialized notation: the population proportion is commonly denoted by $\bar{y}_U = p$, and the population variance reduces to

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - p)^2 = \frac{N}{N-1}p(1-p)$$

We estimate the population mean (proportion) by the sample mean (proportion) $\bar{y} = \widehat{p}$, and find

$$V(\widehat{p}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

The sample variance reduces to

$$s^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} (y_i - \widehat{p})^2 = \frac{n}{n-1} \widehat{p}(1 - \widehat{p})$$

Then

$$\widehat{V}\,(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n-1}\left(1 - \frac{n}{N}\right) \quad \text{and} \quad \mathsf{SE}\,(\widehat{p}) = \sqrt{\widehat{V}\,(\widehat{p})}$$

Sampling weights (Sec 2.4)

Define the **sampling weight** of unit i to be the reciprocal of the inclusion probability

$$w_i = \frac{1}{\pi_i}$$

We interpret w_i as the number of population units represented by unit i.

In SRS $w_i = 1/\pi_i = N/n$ for each i. Thus each unit in the sample represents N/n units, itself plus $N/n - 1$ of the unsampled units.

Definition: A sampling design for which every unit has the same sampling weight is called a *self-weighting* sample.

Thus SRS is a self-weighting method.

Also for SRS, we can write our estimates of t and $\bar{y}_U$ as

$$\hat{t} = \sum_{i \in \mathcal{S}} w_i y_i$$

and

$$\bar{y} = \frac{\displaystyle\sum_{i \in \mathcal{S}} w_i y_i}{\displaystyle\sum_{i \in \mathcal{S}} w_i}$$

Sampling weights will become useful later in the course, when we consider sampling schemes with unequal selection probabilities.