Lecture 2: Simple Random Sample

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

General framework

- Suppose N is the population size. That is, there are N units in the **universe** or **finite population** of interest.
- The N units in the universe are denoted by an **index set** of labels:

$$\mathcal{U} = \{ 1, 2, 3, \ldots N \}$$

Note: Some texts will denote $\mathcal{U} = \{u_1, u_2, u_3, \ldots u_N\}$.

- From this universe (or population) a sample of n units is to be taken. Let $\mathcal{S}$ represent a sample of n units from $\mathcal{U}$.
- Associated with each of the N units is a measurable value related to the population characteristic of interest. Let y_i be the value associated with unit i, and the population of y-values is $\{y_1, y_2, \ldots, y_N\}$.
- Sampling designs that are based on planned randomness are called **probability samples**, and a probability $P(\mathcal{S})$ is assigned to every possible sample $\mathcal{S}$.

The probability that unit *i* will be included in a sample is denoted π_i and is called the **inclusion probability** for unit *i*.

- Common statistics of interest: Let $y_1, y_2, \ldots, y_n$ be a sample of y-values.
  - The **sample mean** is $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
  - The **sample variance** is $s^2 = \frac{1}{n-1}\left[(y_1-\overline{y})^2 + (y_2-\overline{y})^2 + \cdots + (y_n-\overline{y})^2\right]$

$$= \frac{1}{n-1}\sum_{i-1}^{n}(y_i-\overline{y})^2 = \frac{1}{n-1}\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]$$

  - The **sample standard deviation** s is $\sqrt{s^2}$.
- Common parameters of interest:
  - Notation: Let parameter t be the **population total** and parameter $\overline{y}_U$ be the **population mean** from a finite population of size N. Thus,

$$t = \sum_{i=1}^{N} y_i \qquad \overline{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i = t/N$$

  - The **population variance** parameter S^2 is defined as:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\overline{y}_U)^2$$
$$= \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{t^2}{N}\right) = \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - N\overline{y}_U^2\right)$$

  - The **population standard deviation** parameter S is defined as $S = \sqrt{S^2}$.

SRS: Simple random sampling without replacement

An SRS of size n is the probability sampling design for which a fixed number of n units are selected from a population of N units without replacement such that every possible sample of n units has equal probability of being selected.

$$\pi_i = \frac{n}{N}.$$

- Sample mean

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

- Unbiased estimate

$$E(\bar{y}) = \bar{y}_{\mathcal{U}}.$$

- Variance of sample mean

$$\mathrm{Var}(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N}),$$

where S^2 is the population variance, and $(1 - n/N)$ is called the finite population correction (fpc).

SRS

- Sample variance

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

- Unbiased estimate

$$E(s^2) = S^2.$$

$$\hat{\mathrm{Var}}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

- Standard error (SE)

$$\mathrm{SE}(\bar{y}) = \sqrt{(1 - n/N)s^2/n}.$$

- Coefficient of variation (CV)

$$\mathrm{CV}(\bar{y}) = \frac{\sqrt{\mathrm{Var}(\bar{y})}}{E(\bar{y})} = \sqrt{(1 - n/N)}\frac{S}{\sqrt{n}\bar{y}_{\mathcal{U}}},$$

and its estimate

$$\hat{\mathrm{CV}}(\bar{y}) = \frac{\mathrm{SE}(\bar{y})}{\bar{y}}.$$

SRS

- Total t

$$\hat{t} = N\bar{y}.$$

- Variance of $\hat{t}$

$$\mathrm{Var}(\hat{t}) = N^2 \mathrm{Var}(\bar{y}),$$

  and the estimate

$$\hat{\mathrm{Var}}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n}.$$

- Sampling weights

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

SRSWR

- Sample mean
$$\bar{y} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i.$$
- Unbiased estimate
$$E(\bar{y}) = \bar{y}_{\mathcal{U}}.$$
- Variance of sample mean
$$\mathrm{Var}(\bar{y}) = \frac{S^2}{n}(1-\frac{1}{N}),$$
where S^2 is the population variance.
- Biased estimate
$$E(s^2) = \frac{N-1}{N}S^2 = \sigma^2.$$

Proportion

- Sample proportion
$$\hat{p} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i.$$
- Unbiased estimate
$$E(\hat{p}) = p.$$
- Variance of sample proportion
$$\mathrm{Var}(\hat{p}) = \frac{N-n}{N-1} \frac{p(1-p)}{n}.$$
- Estimated variance
$$\hat{\mathrm{Var}}(\hat{p}) = (1 - n/N) \frac{\hat{p}(1-\hat{p})}{n-1}.$$

Confidence interval

- Confidence interval-large sample

$$[\bar{y} - z_{\alpha/2} SE(\bar{y}), \bar{y} + z_{\alpha/2} SE(\bar{y})].$$

Recall

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

- Confidence interval

$$[\bar{y} - t_{\alpha/2, n-1} SE(\bar{y}), \bar{y} + t_{\alpha/2, n-1} SE(\bar{y})].$$

- Sample size for normal approximation

$$n_{\min} = 28 + 25 \left(\sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^3 / NS^3 \right)^2 .$$

In practice, use s to estimate S, and

$$\sum_{i \in \mathcal{S}} (y_i - \bar{y}_{\mathcal{U}})^3 / n \to \sum_{i=1}^{N} (y_i - \bar{y}_{\mathcal{U}})^3 / N.$$

Confidence interval for proportion

- The probability that a SRS of size n will have exactly j sampling units possessing the attribute (successes) is

$$P(Y=j)=\frac{\binom{t}{j}\binom{N-t}{n-j}}{\binom{N}{n}}.$$

t are one's in the population but unknown.

- Normal approximation

$$\hat{p}\sim N(p,\mathrm{Var}(\hat{p})),$$

i.e., $100(1-\alpha)\%$ confidence interval for p is:

$$\hat{p}\pm z_{\alpha/2}SE(\hat{p}),$$

where

$$SE(\hat{p})=\sqrt{(1-n/N)\frac{\hat{p}(1-\hat{p})}{n-1}}.$$

Sample size: $np\geq 5$ and $n(1-p)\geq 5$.

Example

The U.S. government conducts a Census of Agriculture every five years, collecting data on all farms (defined as any place from which $1000 or more of agricultural products were produced and sold) in the 50 states.[2] The Census of Agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agricultural measures for each of the $N = 3078$ counties and county-equivalents in the United States. The file agpop.dat contains the 1982, 1987, and 1992 information on the number of farms, acreage devoted to farms, number of farms with fewer than 9 acres, and number of farms with more than 1000 acres for the population.

Example

We substitute the sample values $s = 344{,}551.9$ and $\sum_{i \in S} (y_i - \bar{y})^3/n = 1.05036 \times 10^{17}$ for the population quantities S and $\sum_{i=1}^{N} (y_i - \bar{y}_U)^3/N$ in (2.23), giving an estimated minimum sample size of

$$n_{\min} = 28 + 25\left[\frac{1.05036 \times 10^{17}}{(344{,}551.9)^3}\right]^2 \approx 193.$$

For this example, our sample of size 300 appears to be sufficiently large for the sampling distribution of $\bar{y}$ to be approximately normal.

For the data in Example 2.5, an approximate 95% CI for $\bar{y}_U$, using $t_{\alpha/2,299} = 1.968$, is

$$\begin{aligned}&[297{,}897 - (1.968)(18{,}898.434),\ 297{,}897 + (1.968)(18{,}898.434)]\\ &\qquad = [260{,}706,\ 335{,}088].\end{aligned}$$

For the population total t, an approximate 95% CI is

$$\begin{aligned}&[916{,}927{,}110 - 1.968(58{,}169{,}381),\ 916{,}927{,}110 + 1.968(58{,}169{,}381)]\\ &\qquad = [8.02 \times 10^8, 1.03 \times 10^9].\end{aligned}$$

For estimating proportions, the usual criterion that the sample size is large enough to use the normal distribution if both $np \geq 5$ and $n(1-p) \geq 5$ is a useful guideline. A 95% CI for the proportion of counties with fewer than 200,000 acres in farms is

$$0.51 \pm 1.968(0.0275), \text{ or } [0.456, 0.564].$$

To find a 95% CI for the total number of counties with fewer than 200,000 acres in farms, we only need to multiply all quantities by N, so the point estimate is $3078(0.51) = 1570$, with standard error $3078 \times \text{SE}(\hat{p}) = 84.65$ and 95% CI [1403, 1736].

Example

The SURVEYMEANS Procedure

Data Summary

Number of Observations	300
Sum of Weights	3078

Class Level Information

Class Variable	Levels	Values
lt200k	2	0 1

Statistics

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Sum
acres92	297897	18898	260706	335088	916927110
lt200k=0	0.490000	0.027465	0.435951	0.544049	1508.220000
lt200k=1	0.510000	0.027465	0.455951	0.564049	1569.780000

Statistics

Variable	Std Dev	Lower 95% CL for Sum	Upper 95% CL for Sum
acres92	58169381	802453859	1031400361
lt200k=0	84.537220	1341.856696	1674.583304
lt200k=1	84.537220	1403.416696	1736.143304

The weight for every observation in this sample is $w_i = 3078/300$; note that the sum of the weights is 3078 $(= N)$. ■

Sample size esitmation

Follow these steps to estimate the sample size:

- Precision: How much error is tolerable? For example,

$$P(|\bar{y} - \bar{y}_{\mathcal{U}}| \leq r) = 1 - \alpha.$$

  e is called the **margin of error** in many surveys. For many surveys of people in which a proportion is measured, $e = 0.03$ and $\alpha = 0.05$. Sometimes

$$P\left(|\frac{\bar{y} - \bar{y}_{\mathcal{U}}}{\bar{y}_{\mathcal{U}}}| \leq e\right) = 1 - \alpha,\ \bar{y}_{\mathcal{U}} \neq 0.$$

- Equation: Find an equation relating the sample size n and your expectations of the sample. To obtain absolute precision e, e.g,

$$e = z_{\alpha/2}\sqrt{(1 - n/N)}\frac{S}{\sqrt{n}}.$$

- Solution: Estimate and solve for n.

$$n = \frac{n_0}{1 + n_0/N} = \frac{z_{\alpha/2}^2 S^2}{e^2 + z_{\alpha/2}^2 S^2/N},$$

  where $n_0 = (z_{\alpha/2}S/e)^2$, sample size for SRSWR.
- Adjustment: Make any possible adjustment.

- If the main responses of interest is a proportion,

$$S^2 = p(1-p)$$

which attains its maximal value when $p = .5$.

- One major problem S^2 is unknown! Follow the following ways:
  - A Pilot Study: A small sample size pilot study can be conducted prior to the primary study to provide an estimate of S^2.
  - Previous Studies: Other similar studies may have been conducted elsewhere and appear in the professional journals. Measures of variability from earlier studies may provide an estimate of S^2.
  - Guess?: If nothing else is available, guess the variance. Sometimes a hypothesized distribution of the data will give us information about the variance.

Example

Before taking the sample of size 300 in Example 2.5, we took a pilot sample of size 30 from the population. One county in the pilot sample of size 30 was missing the value of *acres92*; the sample standard deviation of the remaining 29 observations was 519,085. Using this value, and a desired margin of error of 60,000,

$$n_0 = \cancel{(1.96)^2} \frac{519{,}085^2}{60{,}000^2} = 288.$$

We took a sample of size 300 in case the estimated standard deviation from the pilot sample is too low. Also, we ignored the fpc in the sample size calculations; in most populations, the fpc will have little effect on the sample size.