# Confidence Intervals

Survey Sampling
Statistics 4234/5234
Fall 2018

September 18, 2018

## Confidence intervals (Sec 2.5)

Consider the population $\mathcal{U} = \{1, 2, \ldots, N\}$ with numerical values $\{y_1, y_2, \ldots, y_N\}$. As usual we denote the population mean and variance by

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

We will take a random sample according to the sampling scheme $P(\mathcal{S})$, then estimate the population parameter $\theta$ by the point estimate $\widehat{\theta}_{\mathcal{S}}$, and the interval estimate $\mathrm{CI}(\mathcal{S})$.

Definition: The interval estimator CI is a $100(1-\alpha)\%$ **confidence interval for** $\theta$ if

$$P(\theta \in \mathrm{CI}) = \sum_{\mathcal{S}} P(\mathcal{S})\, I_{\{\theta \in \mathrm{CI}(\mathcal{S})\}} \geq 1 - \alpha$$

Example: Consider the population of $N = 5$ units, with numerical values $\{20, 4, 10, 2, 12\}$. Suppose we wish to estimate the parameter $\theta = \bar{y}_U = 9.6$ using SRS of size $n = 2$ and

$$\text{CI}(\mathcal{S}) \quad \Leftarrow \quad \bar{y}_{\mathcal{S}} \pm 2\text{SE}_{\mathcal{S}}(\bar{y}_{\mathcal{S}})$$

where

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

Is this a 95% confidence interval? 85%? 75%?

Consider all 10 possible samples; compute the sample mean and standard deviation, and resulting confidence interval, for each.

You will find that the interval covers the value 9.6 for 8 of the 10 possible samples. Thus the interval defined above gives a $100(1-\alpha)\%$ confidence interval for and $\alpha \geq .20$.

In sampling from infinite populations (i.e., in every statistics course other than this one), we learned that if the sample size is sufficiently large, for many situations (for example maximum likelihood estimation) we have

$$\widehat{\theta} \dot{\sim} \text{Normal}\left[\theta,\ V\left(\widehat{\theta}\right)\right]$$

and thus

$$\widehat{\theta} \pm z_{\alpha/2}\text{SE}\left(\widehat{\theta}\right)$$

gives an *approximate* $100(1-\alpha)\%$ confidence interval for $\theta$.

There exists a simliar result for inference about the population mean (or total) in finite populations.

Propostion: If $N$ and $n$ and $N-n$ are sufficiently large, under simple random sampling,

$$\frac{\bar{y}-\bar{y}_U}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \dot{\sim} \text{Normal}\,(0,\ 1)$$

and thus an approximate $100(1-\alpha)\%$ confidence interval for $\bar{y}_U$ is (under SRS) given by

$$\bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

i.e.,

$$\bar{y} \pm z_{\alpha/2}\text{SE}\,(\bar{y})$$

A useful tool for assessing the appropriateness of the normal approximation in a particular problem is the *bootstrap*: Take repeated samples *with replacement* from your original sample; the distribution of the resulting sample means approximates the sampling distribution of $\bar{y}$.

## Sample size calculations (Sec 2.6)

Suppose we want our estimate $\bar{y}$, based on a SRS of size $n$ to be within $e$ of the population mean $\bar{y}_U$, with probability at least $1-\alpha$, where $e$ and $\alpha$ are specified. How large must $n$ be?

Well, we require that

$$P\left(|\bar{y} - \bar{y}_U| \leq e\right) \geq 1 - \alpha$$

and thus

$$P\left(\left|\frac{\bar{y}-\bar{y}_U}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\right| \leq \frac{e}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}}\right) \geq 1-\alpha$$

and thus

$$\frac{e}{\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}} \geq z_{\alpha/2}$$

We require that

$$e \geq z_{\alpha/2}\frac{S}{\sqrt{N}}\sqrt{1-\frac{n}{N}}$$

We can solve this in two stages.

First let

$$n_0 = \frac{n}{1 - \frac{n}{N}}$$

and solve

$$n_0 = \left( \frac{z_{\alpha/2} S}{e} \right)^2 .$$

Then

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

gives the necessary sample size for the desired precision and confidence.

Effectively, $n_0$ is the required sample size ignoring the fpc, and thus recommends the sample size required under simple random sampling with replacement. Since without replacement is more efficient, it will always be the case that $n \leq n_0$.

Example: The population size is $N = 500$; based on a pilot study, we estimate the population SD is about 0.85.

We want to estimate the population mean to within $\pm 0.30$ with probability at least .95.

Then $e = 0.30$ and $\alpha = .05$, so $z_{\alpha/2} = 1.96$; it is generally a good idea to build some conservatism into the standard deviation, we'll take $S = 1.20$ (no particular reason for this value, just what I choose).

Take

$$n_0 = \left(\frac{1.96 \times 1.20}{0.30}\right)^2 = 61.47$$

and

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{61.47}{1 + 61.47/500} = 54.74$$

so take a SRS of size $n = 55$.

In the case of estimating a population proportion, we have

$$S^2 = \frac{N}{N-1} p(1-p) \approx p(1-p) \leq 0.25$$

In some situations this may be overly conservative. If, for example, we are know that the population proportion is at least 0.65 and at most 0.85, we'd use

$$S^2 = (.65)(.35) = 0.2275$$

instead. Why?