9. Introduction to Estimation. Confidence Interval for a Population Mean

Two major problems

Statistical inference is the process by which we acquire information about populations from samples. There are two procedures for making inferences:

- Estimation
- Hypotheses testing

Concepts of estimation

The objective of estimation is to determine the value of a population parameter on the basis of a sample statistic. There are two types of estimators

- Point Estimator
- Interval estimator

Point estimator

A point estimator draws inference about a population by estimating the value of unknown parameter using a single value or a point.

An *unbiased* estimator of a population parameter is an estimator whose expected value is equal to that parameter.

An unbiased estimator is said to be *consistent* if difference between estimator and the parameter grows smaller as sample size grows larger.

If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to be *relatively efficient*.

Point estimator: three main examples

- **The sample mean $\overline{X}$ is the most commonly used unbiased and consistent estimator of population mean μ. If data come from a normal population than it is also most efficient.**

- **The sample variance s^2 is the most commonly used unbiased and consistent estimator of population variance σ^2.**

- **The sample proportion $\widehat{p}$ is the most commonly used unbiased and consistent estimator of population proportion p.**

Confidence interval estimator

A *confidence interval (CI) estimator* draws inferences about population by estimating the value of an unknown parameter using an interval.

It is an interval calculated from the observations (which differs in principle from sample to sample) that frequently includes the parameter of interest if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the *confidence level*.

CI for population mean μ when sample size is large ($n \geq 30$)

Consider a sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2. When n is large, by the Central Limit Theorem, the sample mean $\bar{X}$ is approximately normally distributed with mean μ and variance σ^2/n. Using this fact one can show that for any $0 < \alpha < 1$

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

Since by the Law of Large Numbers $s^2 \approx \sigma^2$ we also get that

$$P\left(\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) \approx 1 - \alpha$$

CI for population mean μ when sample size is large ($n \geq 30$)

Thus the $100(1-\alpha)\%$ *confidence interval for population mean* μ *when* $n \geq 30$ is the following interval:

$$\left[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

The confidence interval is often represented in this form:

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

The $\pm$ part of the formula, $z_{\alpha/2}\frac{s}{\sqrt{n}}$, is called the *margin of error*.

Note also that the point estimator of μ, $\bar{X}$, is the middle point of the CI.

The probability $1-\alpha$ is called the *confidence level*.

Interpreting the confidence interval

Before the data are collected it is OK to say that the probability of covering population mean μ by the $100(1-\alpha)\%$ CI is equal to $1-\alpha$.

However, once the sample mean and the margin error are computed (based on a specific sample) we cannot talk about probability any longer because population mean μ is a number. That is why we use the word "confidence".

We usually say "*with confidence level* $100(1-\alpha)$ *%, population mean* μ *lies in the confidence interval*".

z-CI for μ: example

Example 1. The caffeine content (in milligrams) of a random sample of 50 cups of black coffee dispensed by a new machine is measured. The mean and standard deviation are 100 milligrams and 7.1 milligrams, respectively. Construct a 98% CI for the true mean caffeine content per cup dispensed by the machine.

Solution. The sample size is $n = 50 > 30$; the confidence level is 98%, therefore,

$$\alpha = .02, \alpha/2 = .01, z_{.01} = 2.33, \bar{X} = 100, s = 7.1$$

Thus, the 98% CI is given by

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 100 \pm 2.33 \frac{7.1}{\sqrt{50}} = 100 \pm 2.34$$

So, we can claim that with confidence level 98%, the mean caffeine content per cup lies in the confidence interval 100 ± 2.34.

Sample size determination

We can control the width of the CI by changing the sample size: the larger the sample size, the smaller the margin of error. However, the large sample size will cost more. So sometimes *before* the data are collected we can try to estimate what sample size n can provide us with *targeted* confidence level $1-\alpha$ and margin of error W.

This leads us to the following equation:

$$z_{\alpha/2}\frac{s}{\sqrt{n}} \approx W$$

Solving this equation with respect to n we get this estimate for a required sample size:

$$n \approx \left(\frac{z_{\alpha/2}s}{W}\right)^2$$

Sample size determination

Since α and W are given, the only issue with the formula is that we also need to know *sample* variance s^2, because we only *plan* to draw a sample.

There are two ways how we can address the problem:

- We can use the value given by the *range approximation*:

$$s \approx Range/4$$

- Or we can use so-called *historical value* of s based on past experience. For instance, we can draw a small *trial* sample and use its sample variance.

Sample size determination: example

Example 2. A research project for an insurance company wishes to investigate the mean value of the personal property held by urban apartment renters. A previous study suggested that the sample standard deviation should be roughly \$10000. A 95% confidence interval with width of \$1000 (a plus or minus of \$500) is desired. How large a sample must be taken to obtain such a confidence interval?

Solution. Since $W = 500$, $\alpha = .05$ and we have info on the standard deviation, the required sample size is equal to

$$\boldsymbol{n \approx \left(\frac{z_{\alpha/2} s}{W}\right)^2 = \left(\frac{1.96 \times 10000}{500}\right)^2 \approx 1537}$$

CI for population mean μ when sample size is small ($n < 30$)

If the sample size is small (less than 30) the confidence interval is given by

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,n-1}$ is the critical value of t-distribution with $n - 1$ degrees of freedom.

A random sample is assumed to be taken from a *normal* population.

t-CI for μ: example

Example 3. A furniture mover calculates the actual weight of shipment as a proportion of estimated weight for a sample of 25 recent jobs. The sample mean is 1.13 and the sample standard deviation is .16. Calculate a 95% CI for the population mean. Assume that data are taken from a normal population.

Solution. The sample size is $n = 25 < 30$, the confidence level is 95%, therefore,

$$\alpha = .05, t_{.025,24} = 2.064, \bar{X} = 1.13, s = .16$$

Thus, the 95% CI is given by

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}} = 1.13 \pm 2.064 \frac{.16}{\sqrt{25}} = 1.13 \pm .066$$

CI for μ: exercises

Exercise 1. Sixty pieces of a plastic are randomly selected, and the breaking strength of each piece is recorded in pounds per square inch. Suppose that: $\overline{X} = 26$ and $s = 1.5$ pounds per square inch. Find a 99% confidence interval for the mean breaking strength .

CI for μ: exercises

Exercise 2. An electrical company tested a new type of oil to be used in its transformers. Twenty-five readings of dielectric strength were obtained. Dielectric strength is the potential (in kilovolts per centimeter of thickness) necessary to cause a disruptive discharge of electricity through an insulator. The results of the test gave: $\bar{X} = 77$ kV, $s = 8$ kV. Find a 95% confidence interval for the mean dielectric strength of the oil. Assume that data are taken from a normal population.

CI for μ: exercises

Exercise 3. A production manager noticed that the mean time to complete a job was 160 minutes. The manager made some changes in the production process in an attempt to reduce the mean time to finish the job. A stem-and-leaf plot of a sample of 11 times is as follows:

13 | 9

14 | 25

15 | 01356

16 | 24

17 | 0

Note: 14|5 = 145 minutes

The sample mean and standard deviation are 153.36 and 9.47, respectively. Construct a 95% confidence interval for the mean time.

CI for μ: exercises

Exercise 4. **How many households in a large town should be randomly sampled to estimate the mean number of dollars spent per household (per week) on food supplies to within \$10 with 90% confidence? Assume a standard deviation of \$50.**