Chapter 9: Confidence Intervals

• Statistical Estimation
  
  Point Estimation
  
  Interval Estimation

• Confidence Intervals
  
  Two-sided Confidence Intervals
  
  One-sided Confidence Intervals

• Student’s t Distribution
Statistical Estimation

Point Estimation: using the data to calculate a single estimate of the parameter of interest. For example, we often use the sample mean $\bar{x}$ to estimate the population mean $\mu$.

Interval Estimation: provides a range of values (an interval) that may contain the unknown parameter (such as the population mean $\mu$).
Confidence Intervals: an interval that contains the unknown parameter (such as the population mean $\mu$) with certain degree of confidence.

Example: Consider the distribution of serum cholesterol levels for all males in the US who are hypertensive and who smoke. This distribution has an unknown mean $\mu$ and a standard deviation 46 mg/100ml. Suppose we draw a random sample of 12 individual from this population and find that the mean cholesterol level is $\bar{x} = 217mg/100ml$.

$\bar{x} = 217mg/100ml$ is a point estimate of the unknown mean cholesterol level $\mu$ in the population.

However, because of the sampling variability, it is important to construct an interval estimate of $\mu$ to account for the sampling variability. A 95% confidence interval for $\mu$ is

$$
\left( 217 - 1.96 \frac{46}{\sqrt{12}}, \quad 217 + 1.96 \frac{46}{\sqrt{12}} \right).
$$

or

$$(191, \quad 243).$$

A 99% confidence interval for $\mu$ is

$$
\left( 217 - 2.58 \frac{46}{\sqrt{12}}, \quad 217 + 2.58 \frac{46}{\sqrt{12}} \right).
$$

or

$$(183, \quad 251).$$
Confidence Intervals

Under the normality assumption

\[ P \left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95. \]

In general, by the CLT, for reasonably large sample size \( n \), the above equation is still approximately true. Thus a 95% confidence interval for \( \mu \) when \( \sigma \) is known is

\[ \left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right). \]

Let \( z_{\alpha/2} \) be the value that cuts off an area of \( \alpha/2 \) in the upper tail of the standard normal distribution. A \( 1 - \alpha \) confidence interval for the population mean \( \mu \) is

\[ \left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \]
Confidence Intervals: What do they mean?

In repeated sampling, from a normally distributed population with a known standard deviation, $100(1 - \alpha)$ percent of all intervals of the form

$$
\left( \bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)
$$

will in the long run cover the population mean $\mu$.

See the simulations in R.
Confidence Intervals

In general, a confidence interval of an unknown quantity is

\[
\text{point estimate} \pm (\text{reliability coefficient}) \times (\text{standard error}).
\]

Sometimes, we call

\[
\text{margin of error} = (\text{reliability coefficient}) \times (\text{standard error})
\]

\[
= \text{half of the length of the confidence interval.}
\]
Sample size calculation based on specified length of CI

In the cholesterol level example, the 95% confidence interval is (191, 243). Its length is \(243 - 191 = 52\). How large a sample would we need to reduce its length to 20?

Recall that the 95% confidence interval is 
\[
\left(217 - 1.96 \frac{46}{\sqrt{n}}, \quad 217 + 1.96 \frac{46}{\sqrt{n}}\right).
\]

The length of this confidence interval is \(2 \times 1.96 \times 46/\sqrt{n}\). So to find the required sample size \(n\), we can solve the equation
\[
2 \times 1.96 \frac{46}{\sqrt{n}} = 20.
\]

We find
\[
n = \left(\frac{1.96 \times 46}{10}\right)^2 = 81.3 \approx 82.
\]
One-sided confidence interval

Sometimes, we are interested in an upper limit for the population mean \( \mu \) or a lower limit for \( \mu \). In such cases, one-sided confidence intervals are appropriate.

**Example:** Consider the distribution of hemoglobin levels for the population of children under 6 who have been exposed to high levels of lead. Suppose that this distribution has \( \text{sd } \sigma = 0.85g\text{100ml} \). Because children who have lead poisoning tend to have much lower levels of hemoglobin than children who do not, we are interested in an upper confidence limit for \( \mu \), the mean of the hemoglobin levels in this population.

Suppose that we have a random sample of 74 children from this population. The sample mean \( \bar{x} = 10.6g\text{100ml} \). We construct a 95% upper confidence limit. The idea is to find \( c \) such that

\[
P \left( \mu \leq \bar{X} + c \frac{\sigma}{\sqrt{n}} \right) = 0.95.
\]

That is

\[
P \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq -c \right) = 0.95.
\]

Thus \( c = 1.645 \). The upper confidence limit is

\[
10.6 + 1.645 \times \frac{0.85}{\sqrt{74}} = 10.8.
\]
Student’s t-distribution

So far we have assumed that $\sigma$ is known. However, in reality, both $\mu$ and $\sigma$ are usually unknown. All we have is the data. Let $x_1, \ldots, x_n$ be the observations. Let the sample mean and sample variance be

$$
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
$$

The confidence intervals can be constructed based on the following $t$-statistic:

$$
T = \frac{\bar{x} - \mu}{s/\sqrt{n}}.
$$

We can compare the $t$-statistic with the $z$-statistic:

$$
Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.
$$

The difference is

- In $Z$, we use $\sigma$ (when $\sigma$ is known).
- In $T$, we use $s$ (when $\sigma$ is unknown).
Suppose the data is from the normal distribution $N(\mu, \sigma^2)$. Then $T$ has a t-distribution with $n - 1$ degrees of freedom. This is often denoted as

$$T \sim t_{n-1}.$$ 

This result was first obtained by W. S. Gosset in the paper “The Probable Error of a Mean,” Biometrika, 6 (1908), 1-25. Gosset used the pseudonym “Student”. So this distribution is called student’s t distribution, or in short, t-distribution.
Example: A sample of 16 ten-year-old girls gave a mean weight of 71.5 and a standard deviation of 12 pounds. Assuming normality, find the 90, 95, and 99 percent confidence intervals for the population
mean weight $\mu$.

Let $t_{\alpha/2}(n - 1)$ the value that cuts off the upper area of $\alpha/2$ in a
t-distribution with $n - 1$ degrees of freedom. The general form of the
certainty interval based on the t-distribution is

$\left( \bar{x} - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}} \right)$.

The 90, 95, and 99 percent confidence intervals are

$\left( 71.5 - 1.75 \frac{12}{4}, 71.5 + 1.75 \frac{12}{4} \right)$,

$\left( 71.5 - 2.13 \frac{12}{4}, 71.5 + 2.13 \frac{12}{4} \right)$,

$\left( 71.5 - 2.95 \frac{12}{4}, 71.5 + 2.95 \frac{12}{4} \right)$,

or

$(66.25, 76.75)$,

$(65.11, 77.89)$,

$(62.65, 80.35)$,

respectively.
Let $\bar{X}$ be the sample mean and $S^2$ be the sample variance of a random sample from a $N(\mu, \sigma^2)$ distribution. Denote

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$ 

Then

$$T \sim t_{n-1}.$$ 

This result was first obtained by W. S. Gosset in the paper “The Probable Error of a Mean,” Biometrika, 6 (1908), 1-25. Gosset used the pseudonym “Student”. So this distribution is called student’s $t$ distribution, or in short, $t$-distribution.
Example: Lloyd and Mailloux [1988, Analysis of S-100 Protein Positive Folliculo-Stelate Cells in Rat Pituitary Tissues, American Journal of Pathology, 133, 338-348] reported the following data on the pituitary gland weight in a sample of four Wistar Furth Rats: mean = 9.0 mg, standard error of the mean = 1.0.

(a) What was the sample standard deviation?
(b) Construct a 95% confidence interval for the mean pituitary weight of a population of similar rats.