1.017/1.010 Class 15
Confidence Intervals

Interval Estimates

Parameter estimates computed from a random sample $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ can vary around the unknown true value. For any given estimate, we seek a **two-sided confidence interval** that is likely to include the true value a:

$$a_L \leq a \leq a_U$$

$[a_L, a_U]$ is called an **interval estimate**.

Standardized Statistics

Interval estimate is often derived from a **standardized statistic**. This is a random variable that depends on both the unknown true value a and its estimate $\hat{\boldsymbol{a}}$.

An example is the **z statistic**:

$$\boldsymbol{z}(\hat{\boldsymbol{a}}, a) = \frac{\hat{\boldsymbol{a}} - a}{SD(\hat{\boldsymbol{a}})}$$

If the estimate is unbiased $E[\boldsymbol{z}] = 0$ and $Var[\boldsymbol{z}] = 1$, for any $\boldsymbol{x}$ or $\hat{\boldsymbol{a}}$ probability distribution with finite moments (prove).

Example:

Suppose:

$a = E[\boldsymbol{x}]$ = mean of the $\boldsymbol{x}$ probability distribution
$\hat{\boldsymbol{a}} = \boldsymbol{m_x}$ = sample mean (of random sample outcome $x_1, x_2, ..., x_N$).

Then:

$$\boldsymbol{z}(\hat{\boldsymbol{a}}, a) = \frac{\boldsymbol{m}_x - E[\boldsymbol{x}]}{SD(\boldsymbol{m}_x)} = \frac{\boldsymbol{m}_x - E[\boldsymbol{x}]}{\frac{SD(x)}{\sqrt{N}}}$$

Deriving Interval Estimates

If we know the probability distribution of the standardized statistic $\boldsymbol{z}$ we can derive an interval estimate. Specify the probability 1-α that $\boldsymbol{z}$ falls in the interval $[z_L, z_U]$ for a **given value** of a (e.g. 0.95):

$$P[z_L \le \boldsymbol{z}(\hat{\boldsymbol{a}}, a) \le z_U] = 1 - \alpha$$

Suppose that $[z_L, z_U]$ is selected so that the probability that $\boldsymbol{z}$ lies above the interval is the same as the probability that it lies below the interval. This gives a **two-sided** interval $[z_L, z_U]$ for $\boldsymbol{z}$:

$$P[\boldsymbol{z}(\hat{\boldsymbol{a}}, a) \le z_L] = F_z(z_L) = \frac{\alpha}{2}$$

$$P[\boldsymbol{z}(\hat{\boldsymbol{a}}, a) \ge z_U] = 1 - F_z(z_U) = \frac{\alpha}{2}$$

$$z_L = F_Z^{-1}\left[\frac{\alpha}{2}\right] \qquad z_U = F_Z^{-1}\left[1 - \frac{\alpha}{2}\right]$$

From the Central Limit Theorem we know that $\hat{\boldsymbol{a}}$ is normal and that $\boldsymbol{z}$ has a **unit normal** distribution [i.e. $\boldsymbol{z} \sim N(0,1)$] for **large sample sizes**. In this case the CDF $F_z(z)$ and its inverse can be evaluated from standard normal distribution tables or with the MATLAB functions `normcdf` and `norminv`.

If 1-α = 0.95 then z_L = -1.96 and z_U = +1.96.

Substitute the definition of $\boldsymbol{z}$ to obtain the corresponding two-sided interval for $\hat{\boldsymbol{a}}$:

$$P[z_L \le \frac{\hat{\boldsymbol{a}} - a}{SD[\hat{\boldsymbol{a}}]} \le z_U] = P[\ a + z_L SD[\hat{\boldsymbol{a}}] \le \hat{\boldsymbol{a}} \le a + z_U SD[\hat{\boldsymbol{a}}]\ ] = 1 - \alpha$$

$$\hat{a} \ge a + z_L SD[\hat{\boldsymbol{a}}] \qquad \hat{a} \le a + z_U SD[\hat{\boldsymbol{a}}]$$

Probability is 1-α that actual sample estimate $\hat{\boldsymbol{a}}$ lies in this interval.

Now suppose that this relatively likely event occurs when the outcome of a particular experiment (i.e. the $\hat{\boldsymbol{a}}$ obtained from a particular random sample) is $\hat{a}$. Then the true a must satisfy the following inequality:

$$\hat{a} - z_U SD[\hat{\boldsymbol{a}}] \le a \le \hat{a} - z_L SD[\hat{\boldsymbol{a}}]$$

This gives the desired 1-α **confidence interval** for a:

$$a_L = \hat{a} - z_U SD[\hat{\boldsymbol{a}}] \qquad a_U = \hat{a} - z_L SD[\hat{\boldsymbol{a}}]$$

We can obtain $SD[\hat{\boldsymbol{a}}]$ in two ways:

1. Derive directly from the **definition of the estimator** $\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ (not always possible). This usually requires replacing population statistics [e.g. $Std(x)$] by sample statistics [e.g. s_x]

2. **Stochastic simulation**, using a = $\hat{\boldsymbol{a}}$ in random number generator (usually possible but not exact). This generally requires an assumption about the form of the underlying distribution $F_x(x)$.

The confidence interval is wider for larger $SD[\hat{\boldsymbol{a}}]$

The confidence interval is wider for larger 1-α (e.g. 99%)

Summary

To derive a two-sided confidence interval:

1. Specify significance level α
2. Compute estimate $\hat{a}$ from the data
3. Compute $Std[\hat{a}]$ in one of two ways:
    - If possible, relate $Std(\hat{a})$ to $Std(x)$ and use the approximation $Std(x) \approx s_x$ (i.e assume the unknown population standard deviation is equal to the sample standard deviation computed from data).
    - Otherwise, derive $Std(\hat{a})$ using stochastic simulation
4. Compute z_L and z_U from and the specified α, assuming an appropriate form for the CDF $F_x(x)$
5. Apply the two-sided confidence interval formula

<u>Example – Large-sample two-sided confidence interval for the population mean:</u>

Consider the sample mean $\boldsymbol{m_x}$, used to estimate the population mean $E[\boldsymbol{x}]$.
In this case, $a = E[\boldsymbol{x}]$ and $\hat{\boldsymbol{a}} = \boldsymbol{m_x}$

Use result from Class 13 to derive $SD[\boldsymbol{m_x}]$ directly, replacing $SD[\boldsymbol{x}]$ by the **sample standard deviation** s_x :

$$SD[\boldsymbol{m}_x] = \frac{SD[\boldsymbol{x}]}{\sqrt{N}} \approx \frac{s_x}{\sqrt{N}}$$

So the large sample (assume z is normal) two-sided 95% confidence interval for the population mean is:

$$m_x - 1.96\frac{s_x}{\sqrt{N}} \le E[\boldsymbol{x}] \le m_x + 1.96\frac{s_x}{\sqrt{N}}$$

Suppose:

$[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{10}] = [\ 0.1 \quad 2.9 \quad 1.0 \quad 1.4 \quad 0.23 \quad 0.54 \quad 1.57 \quad 8.0 \quad 0.40 \quad 1.6]$

Then $m_x = 1.77$, $s_x = 2.34$, $N = 10$ and:

$$1.77 - 1.96\frac{2.34}{\sqrt{10}} \le E[\boldsymbol{x}] \le 1.77 + 1.96\frac{2.34}{\sqrt{10}}$$
$$0.32 \le E[\boldsymbol{x}] \le 3.23$$

Suppose we don't know the probability distribution of $\boldsymbol{z}$ (e.g. because the sample size is too small to justify using the Central Limit Theorem) but we know the distribution of $\boldsymbol{x}$ (except for a few unknown parameters).

Then we can approximate $F_z(z)$ with a virtual experiment, replacing unknown parameters with estimates computed from the random sample. Once this is done we can plot $F_z(z)$ vs z and identify from the plot the F_z^{-1} values needed to derive z_L and z_U. Usually we make a large sample (normal) assumption.