Interval Estimates

Parameter estimates computed from a random sample \( x_1, x_2, \ldots, x_N \) 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 \( \hat{a} \).

An example is the z statistic:

\[
z(\hat{a}, a) = \frac{\hat{a} - a}{SD(\hat{a})}
\]

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

Example:

Suppose:

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

Then:

\[
z(\hat{a}, a) = \frac{m_x - E[x]}{SD(m_x)} = \frac{m_x - E[x]}{SD(x)} \frac{1}{\sqrt{N}}
\]
Deriving Interval Estimates

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

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

Suppose that $[z_L, z_U]$ is selected so that the probability that $z$ 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 $z$:

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

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

$$z_L = F^{-1}_z\left(\frac{\alpha}{2}\right) \quad z_U = F^{-1}_z\left(1 - \frac{\alpha}{2}\right)$$

From the Central Limit Theorem we know that $\hat{a}$ is normal and that $z$ has a unit normal distribution [i.e. $z \sim N(0,1)$] 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 $1-\alpha = 0.95$ then $z_L = -1.96$ and $z_U = +1.96$.

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

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

$$\hat{a} \geq a + z_L SD[\hat{a}] \quad \hat{a} \leq a + z_U SD[\hat{a}]$$

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

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

$$\hat{a} - z_U SD[\hat{a}] \leq a \leq \hat{a} - z_L SD[\hat{a}]$$
This gives the desired 1-\(\alpha\) confidence interval for \(a\):

\[
a_L = \hat{a} - z_U SD[\hat{a}] \quad a_U = \hat{a} - z_L SD[\hat{a}]
\]

We can obtain \(SD[\hat{a}]\) in two ways:

1. **Derive directly from the definition of the estimator**
   \(\hat{a} = \hat{a}(x_1, x_2, ..., x_N)\) (not always possible). This usually requires replacing population statistics [e.g. \(Std(x)\)] by sample statistics [e.g. \(s_x\)].

2. **Stochastic simulation**, using \(a = \hat{a}\) 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 \(SD[\hat{a}]\)

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

**Summary**

To derive a two-sided confidence interval:

1. Specify significance level \(\alpha\)
2. Compute estimate \(\hat{a}\) from the data
3. Compute \(SD[\hat{a}]\) 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 \(\alpha\), assuming an appropriate form for the CDF \(F_x(x)\)
5. Apply the two-sided confidence interval formula

**Example – Large-sample two-sided confidence interval for the population mean:**

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

Use result from Class 13 to derive \(SD[m_x]\) directly, replacing \(SD[x]\) by the sample standard deviation \(s_x\):

\[
SD[m_x] = \frac{SD[x]}{\sqrt{N}} \approx \frac{s_x}{\sqrt{N}}
\]
So the large sample (assume $z$ is normal) two-sided 95% confidence interval for the population mean is:

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

Suppose:

$$[x_1, x_2, ..., x_{10}] = [0.1, 2.9, 1.0, 1.4, 0.23, 0.54, 1.57, 8.0, 0.40, 1.6]$$

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

$$1.77 - 1.96 \frac{2.34}{\sqrt{10}} \leq E[x] \leq 1.77 + 1.96 \frac{2.34}{\sqrt{10}}$$

$$0.32 \leq E[x] \leq 3.23$$

Suppose we don’t know the probability distribution of $z$ (e.g. because the sample size is too small to justify using the Central Limit Theorem) but we know the distribution of $x$ (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.