Introduction to Inference
Confidence Intervals

- A Typical Inference Problem
- The 95% Confidence Interval
  - Definition
  - Calculating a 95% Confidence Interval
- Interpretation of Confidence Intervals
- The General Form of a Confidence Interval
- Finding $z^*$
- Factors Affecting CI Length

The purpose of statistical inference is to draw conclusions from data. We have previously used statistical methods to examine data. Now we are interested in substantiating or validating our conclusions by probability calculations.

A Typical Inference Problem

Suppose we want to find out about the mean lifetime $\mu$ of a certain brand of light bulbs.

Suppose that the true mean $\mu$ is unknown, but we know (perhaps from previous studies) that the SD $\sigma$ of the light bulb lifetime is 90 hours.

In order to estimate the population mean $\mu$ we:

- Take a SRS of 100 light bulbs.
- Calculate the mean lifetime in the sample to be 1100 hours.

Now, we know that if $Z \sim N(0, 1)$, then

$$P(-1.96 < Z < 1.96) = 0.95$$

And since $\frac{\bar{X} - \mu}{\sqrt{n}} \sim N(0, 1)$, we have that

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96\right) = 0.95$$

Thus,

$$P\left(\mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Rearranging terms, we have

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

What can we say about the population mean?

- $E(\bar{X}) = \mu$, $SD(\bar{X}) = 90/\sqrt{100} = 9$
- $\bar{X} \rightarrow \mu$ (Law of Large Numbers)

Recall from the previous lecture that by the Central Limit Theorem we have

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

The distribution is exact if the population distribution is normal, and approximately correct for large $n$ in other cases.

So $\bar{X} \sim N(\mu, 9)$
In other words, there is 95% probability that the random interval
\[
\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]
will cover \( \mu \).

In our example, \( \bar{x} = 1100, \sigma = 90, \) and \( n = 100. \)
Therefore, the 95% confidence interval for \( \mu \) is
\[
(1100 - 1.96 \times 9, 1100 + 1.96 \times 9) = (1082.36, 1117.64)
\]

**Calculating a 95% Confidence Interval**

For the time being, we’ll continue to assume that \( \sigma \) is known. To calculate a 95% confidence interval for the population mean \( \mu \)

1. Take a random sample of size \( n \) and calculate the sample mean \( \bar{x} \).
2. If \( n \) is large enough, \( \bar{x} \sim N \left( \mu, \frac{\sigma}{\sqrt{n}} \right) \) (by the CLT).
3. The confidence interval is given by
\[
\left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)
\]

**Example**

Consider estimating the speed of light using 64 measurements with sample mean \( \bar{x} = 298,054 \text{ km/s} \).
Assume we know (from previous experience) that the SD of measurements made using the same procedure is 60 km/s.
What is a 95% CI for the true speed of light?

Incorrect:
- There is a 95% probability that the true speed of light lies in the interval (298,039.3, 298,068.7).
- In 95% of all possible samples, the true speed of light lies in the interval (298,039.3, 298,068.7).

Correct:
- There is 95% confidence that the true speed of light lies in the interval (298,039.3, 298,068.7).
- There is 95% probability that the true speed of light lies in the random interval \( (\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) \).
- If we repeatedly draw samples and calculate confidence intervals using this procedure, 95% of these intervals will cover the true speed of light.
General Form of a Confidence Interval

In general, a CI for a parameter has the form

\[ \text{estimate} \pm \text{margin of error} \]

where the margin of error is determined by the confidence level \((1 - \alpha)\), the population SD \(\sigma\), and the sample size \(n\).

A \((1 - \alpha)\) confidence interval for a parameter \(\theta\) is an interval computed from a SRS by a method with probability \((1 - \alpha)\) of containing the true \(\theta\).

For a random sample of size \(n\) drawn from a population of unknown mean \(\mu\) and known SD \(\sigma\), a \((1 - \alpha)\) CI for \(\mu\) is

\[ \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \]

Here \(z^*\) is the critical value, selected so that a standard Normal density has area \((1 - \alpha)\) between \(-z^*\) and \(z^*\). The quantity \(z^*\sigma/\sqrt{n}\), then, is the margin of error.

If the population distribution is normal, the interval is exact. Otherwise, it is approximately correct for large \(n\).

Finding \(z^*\)

For a given confidence level \((1 - \alpha)\), how do we find \(z^*\)?

Let \(Z \sim N(0,1)\):

\[ P(-z^* \leq Z \leq z^*) = (1 - \alpha) \iff P(Z < -z^*) = \frac{\alpha}{2} \]

Thus, for a given confidence level \((1 - \alpha)\), we can look up the corresponding \(z^*\) value on the Normal table.

Common \(z^*\) values:

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>90</th>
<th>95</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>(z^*)</td>
<td>1.645</td>
<td>1.96</td>
<td>2.576</td>
</tr>
</tbody>
</table>

Factors Affecting CI Length

It is not hard to see that the length of a CI is given by

\[ \text{Length of the CI} = 2z^* \frac{\sigma}{\sqrt{n}} \]

What happens to the length of the CI if

- we decrease/increase the confidence level?
- we decrease/increase the sample size \(n\)?
- \(\sigma\) is larger/smaller?

If we know before conducting a study that want a specific confidence level and a specific margin of error \(m\), we need to adjust the sample size:

\[ n = \left( \frac{z^* \sigma}{m} \right)^2 \]

Note that this calculation requires the same assumptions we have been using all along: namely that the sample is an SRS, \(\bar{x}\) has normal distribution (or \(n\) is sufficiently large that it is approximately normal), and \(\sigma\) is known.