Lecture 3 - Part II
Confidence Intervals

Theophanis Tsandilas
Statistical inference

The process of deducing the parameters of an underlying probability distribution from a sample

Four broad types:

- Point estimation
- Interval estimation
- Hypothesis testing
- Prediction
Point estimates

How much informative is the following graph?

Mean Time (s)

Techniques

T1

T2

T3
Point estimates

A point estimate can be thought of as a « best guess » of the true population parameter.

Descriptive statistics such as the sample mean or the median are examples of point estimates.

**Question:** What are the point estimates of a population’s variance and standard deviation?
How much informative is the following graph?

A point estimate communicates no information about the uncertainty or quality of the estimate it provides.
Interval estimate

An interval estimate does not provide an exact value, but rather a range of values that the parameter might plausibly take.

Most common method: constructing a **confidence interval** (CI)
Confidence interval (CI)

It specifies a range of values that is expected to contain the true parameter value (but it may not)
Confidence interval (CI)

It specifies a range of values that is expected to contain the true parameter value (but it may not)

It is associated with a confidence level, usually expressed as a percentage

* e.g., 95% CI or 99% CI
Interpreting a confidence interval (CI)

What do 95% confidence intervals represent here?

![Graph showing 2019 general election voting intention with 95% confidence intervals for Conservative, Labour, Lib Dem, Brexit Party, and Other parties. The poll was conducted by Survation and The Economist, involving a telephone poll of 413 adults surveyed on November 21st-23rd. "Don't know" and refused responses were removed.](image-url)
Formal interpretation of CIs

Classical frequentists statistics view a probability as a statement about the frequency with which events occur in the long run.

Of the many 95% CIs that might be constructed, 95% are expected to contain the true population parameter. The other 5% may completely fail!
Understanding probabilities

Suppose that many research teams run independent experiments and construct a 95% CI to estimate a mean.

Question 1: What is the probability that the CI of a random experiment will include the true population value?

Question 2: Given an experiment (e.g., the one in red), what is the probability that its range includes the true population value?
CIs and probabilities

Formally speaking, a given CI does not specify a probability range. The true parameter is considered as fixed and the CI may or may not contain it.

We cannot claim that a 95% CI contains the true population parameter with a 95% probability or with 95% confidence.

Classical frequentists statistics do not allow for such probabilistic reasoning.
Informal interpretation of CIs

However, it is often reasonable to treat a CI as an expression of confidence or belief that it does contain the true value.

See [Baguley] and [Cumming and Finch, 2005]

**Attention:** This view has its critics.
Interpreting a confidence interval (CI)

What do 95% confidence intervals represent here?

---

2019 general election voting intention*, %

**Central estimate**

95% confidence interval

<table>
<thead>
<tr>
<th>Party</th>
<th>Central Estimate</th>
<th>Vote share, 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lib Dem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brexit Party</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sources: Survation; The Economist

*Telephone poll of 413 adults surveyed on November 21st-23rd. "Don't know" and refused removed.
Interpreting a confidence interval (CI)

What do 95% confidence intervals represent here?

Conservatives will gain between 35% and 45% of the votes with 95% probability.

There is a 95% probability that between 35% and 45% of the voters intend to vote for Conservatives.

We are 95% confident that between 35% and 45% of the voters intend to vote for Conservatives.

We expect that the true voting intention for Conservatives will be included by the CI of 95% of the polls.
Confidence level

A 100% CI will include the whole range of possible values.

A 0% CI reduces to a point estimate.

A 95% CI is the most common choice (by tradition).
alpha level

If C is the confidence level of a confidence interval, then:

\[ C = 100 \left( 1 - \alpha \right) \]

where \( \alpha \) (or alpha) represents the number of times that a C% CI is expected to fail:

If \( C = 95 \), then \( \alpha = .05 \)
Structure of a confidence interval

It is defined by two points that form its limits, i.e., its lower and upper bounds.

It can be **symmetrical**, where the point estimate lies in the center of the CI.

...or **asymmetrical**, where the point estimate is not at the center of the CI.
Symmetrical CIs

The intervals can be described by the point estimate plus or minus half of the interval, e.g., 165 ± 6 cm

This half width of the interval is known as the margin of error (MOE)
Width of a CI

Depends on the confidence level:
- 99% CIs are wider than 95% CIs

It also depends on the size of the sample:
- small samples produce wide CIs
Sampling distributions and CIs

To derive the C% CI of a statistic (e.g., the mean), we first need to approximate the sampling distribution of this statistic.

Why?

Because the sampling distribution provides the probability distribution of all possible values of the statistic. Our goal is to identify the C% of these possible values.
Consider the **sampling distribution** of the mean for a normally distributed population (M = 100, SD = 10)

The sampling distribution becomes narrower as more samples are added. Thus, CIs should also become narrower.
Intuition of how CIs work
(explained for means)

Consider the sampling distribution of the mean \((n = 20)\). Let’s take the range between its 10\(^{th}\) and its 90\(^{th}\) percentile.

**Question 1:** How many of the distribution’s values does this range contain?

Imagine that we repeat an experiment by drawing random samples \((n = 20)\) a large number of times. Every time, we calculate the mean of the sample.

**Question 2:** On average, how many of these sample means does the above range contain?
Intuition of how CIs work (explained for means)

On average, how many sample means will be included in this range?
Intuition of how CIs work
(explained for means)

On average, how many 80% CIs will include the true mean?
Standard error (CE) of a statistic

There is a clear connection between the standard deviation of the sampling distribution of a statistic and a CI.

Standard deviation of the sampling distribution

= Standard error (SE) of the statistic
Remember (normal distributions)…

Suppose this represents the sampling distribution of the mean.

How could we derive the 95% confidence interval of the mean given the standard error?
Example

Standard error of the mean (SEM) for a normal population ($M = 100$, $SD = 10$), when $n = 10$, $30$, and $100$

- $n = 10$, SEM = 3.16, 95% CI = ?
- $n = 30$, SEM = 1.24, 95% CI = ?
- $n = 100$, SEM = 1.00, 95% CI = ?
SEM calculation

The **standard error of the mean** \( (\sigma_{\hat{\mu}}) \) derives from the standard deviation \( (\sigma) \) of the original population and the sample size \( n \):

\[
\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n} \quad \text{(variances)}
\]

\[
\Rightarrow \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \quad \text{(standard deviations)}
\]
From standard errors to CIs

The generic form of a symmetric CI is:

$$CI = \hat{\mu} \pm \phi_{a/2} \times \sigma_{\hat{\mu}}$$

where $\phi_{a/2}$ is the $a/2 \times 100$ percentile of a symmetric standardized distribution.
Normal sampling distribution

If the **sampling distribution** is normal. Then, the CI can be constructed as follows:

\[
CI = \hat{\mu} \pm z_{a/2} \times \sigma \hat{\mu}
\]

For a 95% CI, we take the 2.5\(^{th}\) percentile of the standard normal distribution.

standard normal (z) distribution
Normal sampling distribution

standard normal (z) distribution

non-standardized normal distribution

\[ z_{0.025} \]

\[ z_{0.975} = -z_{0.025} \]

\[ \pm \sigma_{\hat{\mu}} \times z_{0.025} \]
Calculate the 2.5 percentile of the z distribution:

```r
> qnorm(.025)
[1] -1.959964
```

Calculate the 97.5 percentile of the z distribution:

```r
> qnorm(.975)
[1] 1.959964
```

Calculate the 2.5 and 97.5 percentiles of the normal distribution with M = 100 and SD = 4:

```r
> qnorm(.975, 100, 4)
[1] 107.8399

> qnorm(.025, 100, 4)
[1] 92.16014
```
Example

Imagine that a researcher takes a random sample of 30 people and keeps them awake for 24 hours prior to taking an IQ test. The researcher finds an IQ score equal to $M = 94.6$.

A common assumption is that the standard deviation of IQ scores in populations is $SD = 15$. Based on this assumption, construct a 95% CI of the mean.

$$CI = \hat{\mu} \pm z_{\alpha/2} \times \sigma_{\hat{\mu}} \implies CI = 94.6 \pm 5.4$$