Confidence Interval Estimation

Prof. dr. Siswanto Agus Wilopo, M.Sc., Sc.D.
Department of Biostatistics, Epidemiology and Population Health
Faculty of Medicine
Universitas Gadjah Mada
Learning Objectives

In this lecture, you learn:

- To construct and interpret confidence interval estimates for the mean and the proportion
- How to determine the sample size necessary to develop a confidence interval for the mean or proportion
Confidence Intervals

Content of this lecture

- Confidence Intervals for the Population Mean, $\mu$
  - when Population Standard Deviation $\sigma$ is Known
  - when Population Standard Deviation $\sigma$ is Unknown
- Confidence Intervals for the Population Proportion, $\pi$
- Determining the Required Sample Size
A point estimate is a single number,

a confidence interval provides additional information about variability

**Diagram:**
- **Point Estimate**
  - Lower Confidence Limit
  - Upper Confidence Limit
  - **Width of confidence interval**
### Point Estimates

We can estimate a Population Parameter … with a Sample Statistic (a Point Estimate)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Parameter</td>
<td>$\mu$</td>
<td>$\pi$</td>
</tr>
<tr>
<td>Sample Statistic</td>
<td>$\bar{X}$</td>
<td>$p$</td>
</tr>
</tbody>
</table>
Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?

- An interval estimate provides more information about a population characteristic than does a point estimate.

- Such interval estimates are called confidence intervals.
Confidence Interval Estimate

- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample
  - Based on observations from 1 sample
  - Gives information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Can never be 100% confident
Estimation Process

Population (mean, $\mu$, is unknown)

Random Sample

Mean $\bar{X} = 50$

Sample

I am 95% confident that $\mu$ is between 40 & 60.
General Formula

The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)
Confidence Level

- Confidence Level
  - Confidence for which the interval will contain the unknown population parameter
  - A percentage (less than 100%)
Confidence Level, \((1-\alpha)\)

- Suppose confidence level = 95%
- Also written \((1 - \alpha) = 0.95\)
- A relative frequency interpretation:
  - In the long run, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
  - No probability involved in a specific interval
Confidence Intervals

Population Mean

σ Known

σ Unknown

Population Proportion

Confidence Intervals

Biostatistics I: 2017-18

Chap 8-12
Confidence Interval for $\mu$ (\(\sigma\) Known)

- **Assumptions**
  - Population standard deviation \(\sigma\) is known
  - Population is normally distributed
  - If population is not normal, use large sample

- **Confidence interval estimate:**

\[
\bar{X} \pm Z \frac{\sigma}{\sqrt{n}}
\]

where \(\bar{X}\) is the point estimate

\(Z\) is the normal distribution critical value for a probability of \(\alpha/2\) in each tail

\(\sigma/\sqrt{n}\) is the standard error
Finding the Critical Value, $Z$

- Consider a 95% confidence interval: $Z = \pm 1.96$

$$1 - \alpha = 0.95$$

$$\alpha = 0.025$$

**Z units:**

- $Z = -1.96$
- $Z = 1.96$

**X units:**

- Point Estimate
- Lower Confidence Limit
- Upper Confidence Limit
### Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Confidence Coefficient, $1 - \alpha$</th>
<th>Z value</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%</td>
<td>0.80</td>
<td>1.28</td>
</tr>
<tr>
<td>90%</td>
<td>0.90</td>
<td>1.645</td>
</tr>
<tr>
<td>95%</td>
<td>0.95</td>
<td>1.96</td>
</tr>
<tr>
<td>98%</td>
<td>0.98</td>
<td>2.33</td>
</tr>
<tr>
<td>99%</td>
<td>0.99</td>
<td>2.58</td>
</tr>
<tr>
<td>99.8%</td>
<td>0.998</td>
<td>3.08</td>
</tr>
<tr>
<td>99.9%</td>
<td>0.999</td>
<td>3.27</td>
</tr>
</tbody>
</table>
Intervals and Level of Confidence

Confidence Intervals

Intervals extend from

\[ \bar{X} + Z \frac{\sigma}{\sqrt{n}} \]

to

\[ \bar{X} - Z \frac{\sigma}{\sqrt{n}} \]

\( (1-\alpha) \times 100\% \) of intervals constructed contain \( \mu \);
\( (\alpha) \times 100\% \) do not.
Example

- A sample of 11 baby born from a slum population has a mean birth weight of 2.20 kilogram. We know from past studied that the population standard deviation is 0.35 kilogram.

- Determine a 95% confidence interval for the true mean birth weight of this population.
Example (continued)

- A sample of 11 baby born from a slum population has a mean birth weight of 2.20 kilogram. We know from past studied that the population standard deviation is 0.35 kilogram.

- Solution:

\[ \bar{X} \pm Z \frac{\sigma}{\sqrt{n}} \]

\[ = 2.20 \pm 1.96 \left( \frac{0.35}{\sqrt{11}} \right) \]

\[ = 2.20 \pm 0.2068 \]

\[ 1.9932 \leq \mu \leq 2.4068 \]
Interpretation

- We are 95% confident that the true mean birth weight of slum population is between 1.9932 and 2.4068 kilogram.

- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.
Confidence Intervals

Confidence Intervals

Population Mean

σ Known

σ Unknown

Population Proportion

σ Known

σ Unknown
Confidence Interval for $\mu$ (\(\sigma\) Unknown)

- If the population standard deviation $\sigma$ is unknown, we can substitute the sample standard deviation, $S$.

- This introduces extra uncertainty, since $S$ is variable from sample to sample.

- So we use the t distribution instead of the normal distribution.
Confidence Interval for $\mu$ (\(\sigma\) Unknown)

(continued)

- **Assumptions**
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample

- **Use Student’s t Distribution**

- **Confidence Interval Estimate:**
  \[
  \bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}}
  \]

(where \(t\) is the critical value of the t distribution with \(n-1\) degrees of freedom and an area of \(\alpha/2\) in each tail)
Student’s t Distribution

- The t is a family of distributions
- The t value depends on degrees of freedom (d.f.)
  - Number of observations that are free to vary after sample mean has been calculated

\[ d.f. = n - 1 \]
Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is $X_3$?

If the mean of these three values is 8.0, then $X_3$ must be 9 (i.e., $X_3$ is not free to vary)

Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)
Student’s t Distribution

Note: $t \xrightarrow{} Z$ as $n$ increases

t-distributions are bell-shaped and symmetric, but have ‘fatter’ tails than the normal

Standard Normal
($t$ with $df = \infty$)

$t$ ($df = 13$)

$t$ ($df = 5$)
Student’s t Table

<table>
<thead>
<tr>
<th>df</th>
<th>.25</th>
<th>.10</th>
<th>.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
</tr>
<tr>
<td>2</td>
<td>0.817</td>
<td>1.886</td>
<td><strong>2.920</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.765</td>
<td>1.638</td>
<td>2.353</td>
</tr>
</tbody>
</table>

Let: \( n = 3 \)
\( df = n - 1 = 2 \)
\( \alpha = 0.10 \)
\( \alpha/2 = 0.05 \)

The body of the table contains t values, not probabilities.
**t distribution values**

With comparison to the Z value

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>t (10 d.f.)</th>
<th>t (20 d.f.)</th>
<th>t (30 d.f.)</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>1.372</td>
<td>1.325</td>
<td>1.310</td>
<td>1.28</td>
</tr>
<tr>
<td>0.90</td>
<td>1.812</td>
<td>1.725</td>
<td>1.697</td>
<td>1.645</td>
</tr>
<tr>
<td>0.95</td>
<td>2.228</td>
<td>2.086</td>
<td>2.042</td>
<td>1.96</td>
</tr>
<tr>
<td>0.99</td>
<td>3.169</td>
<td>2.845</td>
<td>2.750</td>
<td>2.58</td>
</tr>
</tbody>
</table>

Note: $t \rightarrow Z$ as $n$ increases
Example

A random sample of $n = 25$ taken from a normal population has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for $\mu$

- d.f. = $n - 1 = 24$, so

$$t_{\alpha/2, n-1} = t_{0.025, 24} = 2.0639$$

The confidence interval is

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 \leq \mu \leq 53.302$$
Confidence Intervals

Population Mean
- $\sigma$ Known
- $\sigma$ Unknown

Population Proportion

Confidence Intervals
An interval estimate for the population proportion \( \pi \) can be calculated by adding and subtracting an allowance for uncertainty to the sample proportion \( p \)
Confidence Intervals for the Population Proportion, $\pi$

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will estimate this with sample data:

$$\sqrt{\frac{p(1-p)}{n}}$$
Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

\[ p \pm Z \sqrt{\frac{p(1-p)}{n}} \]

- where
  - \( Z \) is the standard normal value for the level of confidence desired
  - \( p \) is the sample proportion
  - \( n \) is the sample size
Example

- A random sample of 100 men shows that 25 are snoring during sleeping.
- Form a 95% confidence interval for the true proportion of men snoring during sleeping.
Example

- A random sample of 100 men shows that 25 are snoring during sleeping.
- Form a 95% confidence interval for the true proportion of a snoring during sleeping.

\[ p \pm Z \sqrt{\frac{p(1-p)}{n}} \]

\[ = \frac{25}{100} \pm 1.96 \sqrt{0.25(0.75)/100} \]

\[ = 0.25 \pm 1.96 (0.0433) \]

\[ 0.1651 \leq \pi \leq 0.3349 \]
Interpretation

- We are 95% confident that the true percentage of a snoring during sleeping in the population is between 16.51% and 33.49%.

- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.
Determining Sample Size using a confidence interval

- For the Mean
- For the Proportion
Sampling Error

- The required sample size needed to estimate a population parameter to within a selected margin of error (e) using a specified level of confidence (1 - \( \alpha \)) can be computed.

- The margin of error is also called sampling error:
  - the amount of imprecision in the estimate of the population parameter
  - the amount added and subtracted to the point estimate to form the confidence interval.
Determining Sample Size

For the Mean

Sampling error (margin of error)

\[ \bar{X} \pm Z \frac{\sigma}{\sqrt{n}} \]

\[ e = Z \frac{\sigma}{\sqrt{n}} \]
Determining Sample Size

For the Mean

\[ e = Z \frac{\sigma}{\sqrt{n}} \]

Now solve for \( n \) to get

\[ n = \frac{Z^2 \sigma^2}{e^2} \]
To determine the required sample size for the mean, you must know:

- The desired level of confidence \((1 - \alpha)\), which determines the critical Z value
- The acceptable sampling error, \(e\)
- The standard deviation, \(\sigma\)
Required Sample Size Example

If \( \sigma = 45 \), what sample size is needed to estimate the mean within \( \pm 5 \) with 90% confidence?

\[
n = \frac{Z^2 \sigma^2}{e^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19
\]

So the required sample size is \( n = 220 \)

(Always round up)
If σ is unknown

- If unknown, σ can be estimated when using the required sample size formula
  - Use a value for σ that is expected to be at least as large as the true σ
  - Select a pilot sample and estimate σ with the sample standard deviation, S
Determining Sample Size

For the Proportion

\[ e = Z \sqrt{\frac{\pi (1 - \pi)}{n}} \]

Now solve for \( n \) to get

\[ n = \frac{Z^2 \pi (1 - \pi)}{e^2} \]
Determining Sample Size

To determine the required sample size for the proportion, you must know:

- The desired level of confidence \((1 - \alpha)\), which determines the critical Z value
- The acceptable sampling error, \(e\)
- The true proportion of “successes”, \(\pi\)
  - \(\pi\) can be estimated with a pilot sample, if necessary (or conservatively use \(\pi = 0.5\))
How large a sample would be necessary to estimate the true proportion defective in a large population within ±3%, with 95% confidence?

(Assume a pilot sample yields $p = 0.12$)
Solution:

For 95% confidence, use \( Z = 1.96 \)
\( e = 0.03 \)
\( p = 0.12 \), so use this to estimate \( \pi \)

\[
\begin{align*}
n &= \frac{Z^2 \pi (1-\pi)}{e^2} \\
   &= \frac{(1.96)^2(0.12)(1-0.12)}{(0.03)^2} \\
   &= 450.74
\end{align*}
\]

So use \( n = 451 \)
Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate
- The level of confidence should always be reported
- The sample size should be reported
- An interpretation of the confidence interval estimate should also be provided
Chapter Summary

- Introduced the concept of confidence intervals
- Discussed point estimates
- Developed confidence interval estimates
- Created confidence interval estimates for the mean (\( \sigma \) known)
- Determined confidence interval estimates for the mean (\( \sigma \) unknown)
- Created confidence interval estimates for the proportion
Chapter Summary

(continued)

- Determined required sample size for mean and proportion settings
- Addressed ethical issues for confidence interval estimation