Program L5

• Confidence intervals
  • Confidence interval around the mean, cont’d
  • Confidence intervals for small samples
  • Find out if a variable is Normally distributed
A study of 35 randomly chosen golfers showed that their average score on a particular course was 92. The standard deviation of the sample is 5.

A 95% confidence interval of the mean score for all golfers in the population can be created:

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

**Assumptions:**
- Simple random sample (SRS)
- Large sample (Rule of thumb: \( n > 30 \))
• 95% confidence interval for the mean
• Using more exact numbers, $z=1.96$ and not 2

$$[\bar{x} \pm 1.96 \text{ SE}_\bar{x}]$$
This tells us that we should use $z^* = 1.96$ to construct a 95% CI.
Confidence interval – Golf example

\[ X = \text{golf score} \]
\[ n = 35 \]
\[ \bar{x} = 92 \]
\[ s = 5 \]

\[
\left[ \bar{x} \pm z^* \frac{s}{\sqrt{n}} \right]
\]

= \[
\left[ 92 \pm 1.96 \frac{5}{\sqrt{35}} \right]
\]

\[ [92 \pm 1.7] \]

or, written as an interval:
\[ [90.3; 93.7] \]

**Interpretation:**

With 95% confidence, the interval 90.3 to 93.7 covers the average golf score in the population.
Confidence interval in SPSS

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>92,31</td>
<td>0,835</td>
</tr>
<tr>
<td>95% Confidence Interval for Mean</td>
<td>90,62</td>
<td>94,01</td>
</tr>
<tr>
<td>Lower Bound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper Bound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>92,29</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>92,00</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>24,398</td>
<td></td>
</tr>
<tr>
<td>Poäng</td>
<td>4,939</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>0,173</td>
<td>0,398</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0,492</td>
<td>0,778</td>
</tr>
</tbody>
</table>

Confidence interval for the mean

(SPSS uses more exact numbers than we did)

SPSS: Analyze >> Descriptive Statistics >> Explore.
20 possible confidence intervals

Look at this interval. It "missed" the population mean!

In the long run, 95% of all confidence intervals will capture the population parameter (C=95%)
• 95% confidence interval for the mean

\[ [\bar{x} \pm 1.96 \ SE_{\bar{x}}] \]

Margin of error \((m)\)
Confidence interval for the mean

- Level C confidence interval for the mean

\[ \left[ \bar{x} \pm z^* \cdot SE_{\bar{x}} \right] \]

Margin of error \((m)\)

C: 90-99% common

- 90% \( \Rightarrow z^* = 1.645 \)
- 95% \( \Rightarrow z^* = 1.96 \)
- 99% \( \Rightarrow z^* = 2.575 \)
Confidence interval for the mean

- Level C confidence interval for the mean

\[ [\bar{x} \pm z \times SE_{\bar{x}}] \]

**Assumptions:**
- Simple random sample (SRS)
- Large sample
  (Rule of thumb: \( n > 30 \))
Let’s use the study of the 35 randomly chosen golfers again.

A **99%** confidence interval of the mean score for all golfers in the population can be created:

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

**Assumptions:**
- Simple random sample (SRS)
- Large sample (Rule of thumb: \(n > 30\))
Confidence interval — Golf example

X = golf score
n = 35
\( \bar{x} = 92 \)
s = 5

\[
\left[ \bar{x} \pm z \times \frac{s}{\sqrt{n}} \right]
= \left[ 92 \pm 2.575 \times \frac{5}{\sqrt{35}} \right]
= \left[ 92 \pm 2.18 \right]
\]

or, written as an interval:
[89.8; 94.2]

**Interpretation:** With 99% confidence, the interval 89.8 to 94.2 covers the average golf score in the population.
Confidence interval in SPSS

### Descriptives

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>92,31</td>
<td>,835</td>
</tr>
<tr>
<td>99% Confidence Interval for Mean</td>
<td>90,04</td>
<td>94,59</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>92,29</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>92,00</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>24,398</td>
<td></td>
</tr>
<tr>
<td>Poäng Std. Deviation</td>
<td>4,939</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>,173</td>
<td>,398</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-.492</td>
<td>,778</td>
</tr>
</tbody>
</table>

Confidence interval for the mean

(SPSS uses more exact numbers than we did)

*SPSS: Analyze >> Descriptive Statistics >> Explore. Click “Statistics” to change confidence level (95% is default).*
Population variable and sampling distribution

Approximately normal sampling distribution

Non-normal population

Only for large samples
Small samples ok
But estimation of std introduces incertaintiy  ➞ t-distribution
If the sample is small and the standard deviation in the population is unknown (estimated by sample $s$), then $X$ is approximatively $t$-distributed.

Assumption: the variable $X$ is Normally distributed!
• Symmetrical distribution, centered around the mean (μ)

• The **degrees of freedom** (d.f.) decide the height of the distribution (d.f. depend on the sample size, d.f. = n -1)
Estimates of statistical parameters can be based upon different amounts of information or data.

The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df).

In general, the degrees of freedom is equal to the sample size minus the number of already estimated parameters.

When calculating the standard deviation, the sample mean is already estimated (and needed for the calculation). Thus $d.f. = n-1$
t-distribution with 1-10 d.f.
Normally distributed or not?

1) **Plot your data** (histogram)

Shows if the distribution is symmetric and bell shaped.

If roughly symmetric and bell-shaped:

2) **Normal quantile plot**
   (available in SPSS)
Normally distributed or not?

Mean = 156.68
Std. Dev. = 14.826
N = 99
Normal quantile plot

Follows a reasonably straight line if Normally distributed

Deviations from the line indicate non-Normality
Detrended Normal quantile plot

Detrended Normal Q-Q Plot of Height

Should be no real clustering of points, with most collecting around the zero line
Examples of non-Normality

The points are not collecting around the zero line.

Skewed
Examples of non-Normality

Skewed with outliers

Large deviations from the line

Not clustered around the zero line
Examples of Normality

Symmetrical

Reasonably straight line

Most points clustered around the zero line
For large samples, the Normal distribution can be used to calculate confidence intervals (approximate values will be provided).

For small samples, drawn from a Normal distribution, the $t$-distribution can be used to calculate confidence intervals. Note: The $t$-distribution is correct to use for all sample sizes, when the standard deviation is estimated from the sample.

If you have a small sample, and cannot assume that the variable comes from a Normal distribution, then confidence intervals cannot be used to draw conclusions about the population!
Confidence interval for the mean

- Level C confidence interval for the mean

\[
\left[ \bar{x} \pm z \times \text{SE}_x \right]
\]

**Assumptions:**
- Simple random sample (SRS)
- Large sample (Rule of thumb: \( n > 30 \))

\[ t \text{-distribution can also be used:} \]
\[
\left[ \bar{x} \pm t \times \text{SE}_x \right]
\]

Value from \( t \)-distribution for \( n-1 \) d.f. (Table D)
Working with the t-distribution: Table D

Table D gives the proportion of a t-distributed population which is *above* a specific critical value (t*)

Note! The numbers within the table are t*
Table D also gives the proportion of a t-distributed population which is *between* $t^*$ and $-t^*$.

Note! The numbers within the table are $t^*$ (a minus in front of the number gives $-t^*$)
t-distribution
Confidence interval for the mean

- Level C confidence interval for the mean

\[ \bar{x} \pm z \times SE_{\bar{x}} \]

**Assumptions:**
- Simple random sample (SRS)
- Large sample (Rule of thumb: \( n > 30 \))

\( t \)-distribution can also be used:

\[ \bar{x} \pm t \times SE_{\bar{x}} \]

**Assumptions:**
- Simple random sample (SRS)
- Sample drawn from a Normal distribution

where \( SE_{\bar{x}} = \frac{s}{\sqrt{n}} \)
The structure of a confidence interval

Add the separate parts together

Point estimate: $\bar{x}$

$\pm$

Number of standard deviations: $z^*$ or $t^*$

$\times$

Standard deviation: $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

**Assumptions:**
- Simple random sample
- Large sample (>30)

**Assumptions:**
- Simple random sample (SRS)
- Sample drawn from a Normal distr.
Confidence interval – BMI example

For a sample of twelve patients, eating a certain medicine known to increase body weight, BMI has been calculated.

\[ X = \text{BMI} \quad \bar{x} = 30.4 \quad s = 2.6 \]

Let’s calculate a 95% confidence interval for the average BMI in this population.
Confidence interval – BMI example

Follows a reasonably straight line

No real clustering, and most points are collecting around the zero line
Recap: Confidence interval for the mean

- Level C confidence interval for the mean

\[ \bar{x} \pm z \times SE_{\bar{x}} \]

**Assumptions:**
- Simple random sample (SRS)
- Large sample (Rule of thumb: \( n > 30 \))

\[ t - \text{distribution can also be used:} \]

\[ \bar{x} \pm t \times SE_{\bar{x}} \]

**Assumptions:**
- Simple random sample (SRS)
- Sample drawn from a Normal distribution
Confidence interval – BMI example

For a sample of twelve patients, eating a certain medicine known to increase body weight, BMI has been calculated.

\[ X = \text{BMI} \]

Let’s calculate a 95% confidence interval for the average BMI in this population.

\[ \bar{x} = 30.4 \quad s = 2.6 \]

The variable BMI can be assumed to be Normally distributed.

\[
\left[ \bar{x} \pm t^* \frac{s}{\sqrt{n}} \right]
\]

\[ t^* = 2.201 \]
This tells us that 95% of a t-distribution with 11 d.f. is located within 2.201 standard deviations around the mean.

Therefore we should use $t^* = 2.201$ to construct a 95% CI.
Confidence interval – BMI example

\[
\left[ \bar{x} \pm t * \frac{s}{\sqrt{n}} \right] = \left[ 30.4 \pm 2.201 \frac{2.6}{\sqrt{12}} \right] = \left[ 30.4 \pm 1.7 \right]
\]

or, written as an interval: \([28.7; 32.1]\)

**Interpretation:**
The average BMI in this patient population is with 95% confidence covered by the interval 28.7 to 32.1 kg/m² (BMI over 25 is considered overweight)
We can have the same confidence interval calculated by SPSS (based on the \( t \)-distribution by default):

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>30,3973</td>
<td>0,76080</td>
</tr>
<tr>
<td>95% Confidence Interval for Mean Lower Bound</td>
<td>28,7228</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upper Bound</td>
<td>32,0718</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>30,5324</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>30,5704</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>6,946</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>2,63550</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>24,04</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>34,32</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>10,28</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>3,05</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>-1,091</td>
<td>0,637</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2,368</td>
<td>1,232</td>
</tr>
</tbody>
</table>

In SPSS:
Analyze >> Descriptive Statistics >> Explore