Lecture 4
Confidence intervals for non-normal distributions
Bootstrapping
Null Hypothesis testing
Introduction to significance tests

Theophanis Tsandilas
Complex designs

Experimental designs can be more complex in different aspects:

- Control for multiple **factors** (*independent variables*)
- Study more than two levels per factor
- Combine between-participants and repeated-measures designs
Example

A research team is interested in assessing the effect of a Geometry course on students’ IQ performance. They randomly create two groups of students (Control vs. Geometry). Each student takes three IQ tests over three weeks.

Control Group

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>102</td>
<td>94</td>
<td>90</td>
<td>104</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Week 2</td>
<td>105</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>99</td>
<td>105</td>
<td>100</td>
<td>93</td>
<td>103</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>101</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>102</td>
<td>98</td>
<td>104</td>
<td>93</td>
</tr>
</tbody>
</table>

Geometry Class Group

<table>
<thead>
<tr>
<th></th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
<th>P17</th>
<th>P18</th>
<th>P19</th>
<th>P20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>105</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>103</td>
<td>104</td>
<td>103</td>
<td>103</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>Week 2</td>
<td>106</td>
<td>100</td>
<td>93</td>
<td>100</td>
<td>100</td>
<td>105</td>
<td>100</td>
<td>95</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>103</td>
<td>101</td>
<td>98</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>101</td>
<td>95</td>
<td>104</td>
<td>97</td>
</tr>
</tbody>
</table>
Example

This is a **mixed design** that combines:

- a between-participants factor (Geometry Course) with two levels: **No** (*Control*) and **Yes** (*Geometry*)
- a within-participants (repeated-measures) factor (Week) with three levels: **Week 1**, **Week 2**, and **Week 3**.
Example

**Question 1**: What’s the mean difference in IQ scores between students who took and student who did not take the geometry course?

**Control Group**

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>102</td>
<td>94</td>
<td>90</td>
<td>104</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Week 2</td>
<td>105</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>99</td>
<td>105</td>
<td>100</td>
<td>93</td>
<td>103</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>101</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>102</td>
<td>98</td>
<td>104</td>
<td>93</td>
</tr>
<tr>
<td>Mean:</td>
<td>102.7</td>
<td>97.0</td>
<td>92.0</td>
<td>102.0</td>
<td>97.3</td>
<td>103.3</td>
<td>99.7</td>
<td>95.7</td>
<td>101.0</td>
<td>95.7</td>
</tr>
</tbody>
</table>

**Geometry Class Group**

<table>
<thead>
<tr>
<th></th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
<th>P17</th>
<th>P18</th>
<th>P19</th>
<th>P20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>105</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>103</td>
<td>104</td>
<td>103</td>
<td>103</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>Week 2</td>
<td>106</td>
<td>100</td>
<td>93</td>
<td>100</td>
<td>100</td>
<td>105</td>
<td>100</td>
<td>95</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>103</td>
<td>101</td>
<td>98</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>101</td>
<td>95</td>
<td>104</td>
<td>97</td>
</tr>
<tr>
<td>Mean:</td>
<td>104.7</td>
<td>99.7</td>
<td>97.0</td>
<td>101.0</td>
<td>100.3</td>
<td>103.0</td>
<td>101.3</td>
<td>97.7</td>
<td>100.0</td>
<td>98.0</td>
</tr>
</tbody>
</table>
Example

**Question 1:** What’s the mean difference in IQ scores between students who took and student who did not take the geometry course?

<table>
<thead>
<tr>
<th>Control Group</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td>102</td>
<td>94</td>
<td>90</td>
<td>104</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td>105</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>99</td>
<td>105</td>
<td>100</td>
<td>93</td>
<td>103</td>
<td>98</td>
</tr>
<tr>
<td><strong>Week 3</strong></td>
<td>101</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>102</td>
<td>98</td>
<td>103</td>
<td>93</td>
</tr>
<tr>
<td><strong>Mean:</strong></td>
<td>102.7</td>
<td>97.0</td>
<td>92.0</td>
<td>102.0</td>
<td>97.3</td>
<td>103.3</td>
<td>99.7</td>
<td>95.7</td>
<td>101.0</td>
<td>95.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geometry Class Group</th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
<th>P17</th>
<th>P18</th>
<th>P19</th>
<th>P20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td>105</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>103</td>
<td>104</td>
<td>103</td>
<td>103</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td>106</td>
<td>100</td>
<td>93</td>
<td>100</td>
<td>100</td>
<td>105</td>
<td>100</td>
<td>95</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td><strong>Week 3</strong></td>
<td>103</td>
<td>101</td>
<td>98</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>101</td>
<td>95</td>
<td>104</td>
<td>97</td>
</tr>
<tr>
<td><strong>Mean:</strong></td>
<td>104.7</td>
<td>99.7</td>
<td>97.0</td>
<td>101.0</td>
<td>100.3</td>
<td>103.0</td>
<td>101.3</td>
<td>97.7</td>
<td>100.0</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Taking the mean is the most common way of aggregating data over an independent variable. However, it is not the only way. Depending on their research goals, the researchers may decide to use the median, or a trimmed mean, e.g., to reduce the influence of extreme values.
Results

IQ Score

95% CIs

Control

Geometry

Diff IQ Score

95% CI
Storing data on CSV files

- group, participant, week, score
- "control", 1, 1, 102
- "control", 1, 2, 105
- "control", 1, 3, 101
- "control", 2, 1, 94
- "control", 2, 2, 97
- "control", 2, 3, 100
- "control", 3, 1, 90
- "control", 3, 2, 93
- ...
- "geometry", 11, 1, 105
- "geometry", 11, 2, 106
- "geometry", 11, 3, 103
- "geometry", 12, 1, 98
- "geometry", 12, 2, 100
- "geometry", 12, 3, 101
- "geometry", 13, 1, 100
- "geometry", 13, 2, 93
- "geometry", 13, 3, 98
- "geometry", 14, 1, 97
- "geometry", 14, 2, 100
- "geometry", 14, 3, 106
...

Each raw represents a unique observation
#read the data from a csv file into a data frame
data <- read.csv(file="IQ-tests.csv", header=TRUE, sep="",""

# aggregate the scores by group and participant
data.aggr <- aggregate(score~group+participant, data, FUN = mean)

# split the aggregated data into two groups
data.control <- data_aggr[data_aggr$group=="control",]
data.geometry <- data_aggr[data_aggr$group=="geometry",]

# calculate the means
mean.control <- mean(data.control$score)
mean.geometry <- mean(data.geometry$score)
Some bad practices

A bad practice is to construct and graph confidence intervals over the full set of data.

This approach is not correct as it treats all observations as independent, i.e., as observations from different participants.
Some bad practices

A common (bad) practice is to construct and graph confidence intervals over the full set of data.

Recommendation: Be always clear about what reported CIs represent and how they were calculated.
Example

Question 2: How does student performance evolve over time, from Week 1 to Week 3?

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>102</td>
<td>94</td>
<td>90</td>
<td>104</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>Week 2</td>
<td>105</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>99</td>
<td>105</td>
<td>100</td>
<td>93</td>
<td>103</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>101</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>102</td>
<td>98</td>
<td>104</td>
<td>93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
<th>P17</th>
<th>P18</th>
<th>P19</th>
<th>P20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometry Class Group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>105</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>103</td>
<td>104</td>
<td>103</td>
<td>103</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>Week 2</td>
<td>106</td>
<td>100</td>
<td>93</td>
<td>100</td>
<td>100</td>
<td>105</td>
<td>100</td>
<td>95</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>103</td>
<td>101</td>
<td>98</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>101</td>
<td>95</td>
<td>104</td>
<td>97</td>
</tr>
</tbody>
</table>
Example: Approach

We examine how IQ scores evolve for all 20 students.

We forget the student groups for now and focus on the repeated-measures variable, the Week.
Example: Results

IQ Score

Week 1  Week 2  Week 3

-2  -1  0  1  2  3

Diff IQ Score

Weeks 2 - 1  Weeks 3 - 2  Weeks 3 - 1
## Control Group

<table>
<thead>
<tr>
<th>Week 1</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>94</td>
<td>90</td>
<td>104</td>
<td>95</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>105</td>
<td>99</td>
<td>93</td>
<td>98</td>
<td>99</td>
<td>105</td>
<td>100</td>
<td>97</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>Week 3</td>
<td>101</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>98</td>
<td>105</td>
<td>102</td>
<td>98</td>
<td>104</td>
<td>93</td>
</tr>
<tr>
<td>Week 3 -1</td>
<td>-1</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>-3</td>
</tr>
</tbody>
</table>

## Geometry Class Group

<table>
<thead>
<tr>
<th>Week 1</th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
<th>P14</th>
<th>P15</th>
<th>P16</th>
<th>P17</th>
<th>P18</th>
<th>P19</th>
<th>P20</th>
</tr>
</thead>
<tbody>
<tr>
<td>105</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>103</td>
<td>104</td>
<td>103</td>
<td>103</td>
<td>100</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>106</td>
<td>100</td>
<td>93</td>
<td>106</td>
<td>100</td>
<td>100</td>
<td>101</td>
<td>100</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>Week 3</td>
<td>103</td>
<td>101</td>
<td>98</td>
<td>106</td>
<td>98</td>
<td>100</td>
<td>101</td>
<td>95</td>
<td>104</td>
<td>97</td>
</tr>
<tr>
<td>Week 3 -1</td>
<td>-2</td>
<td>3</td>
<td>-2</td>
<td>9</td>
<td>-5</td>
<td>-4</td>
<td>-2</td>
<td>-8</td>
<td>4</td>
<td>-2</td>
</tr>
</tbody>
</table>
Non-normal distributions

How do we construct confidence intervals for non-normal distributions?
CIs for binomial proportions

**Example:** A basketball player has attempted a total of $n = 25$ three point shots and has succeeded in $x = 9$. Can you estimate the player’s average success rate?

There are several alternatives for constructing CIs for binomial proportions, where some of them work well only for large samples (Control Limit Theorem) or proportions in the region of $.5$.

Baguley recommends the exact Blaker CI under most circumstances.
R Code

install.packages("exactci")
library(exactci)

ci <- binom.exact(9, 25, tsmethod='blaker',
                   conf.level=.95)$conf.int

cat(ci[1]*100, ci[2]*100, "\n")

The success rate of the player is: 36%, 95% CI [19%, 56%]
CI for a difference in proportions

Example: A second basketball player has attempted a total of 20 three-point shots and has succeeded in 10. Estimate the difference between the success rate of the two players.

Baguley recommends the continuity corrected version of the Wilson CI.
R Code

```r
prop.test(c(25, 20), c(9, 8))$conf.int
```

The success rate of the 2nd player is higher by a 14%, 95% CI [-19%, 47%]
Lognormal distributions

1. Data values are first transformed to a logarithmic scale
2. CIs are computed over log-transformed values
3. CIs are then transformed back to the original scale
Simple math with logarithms

\[ \log_b(x) = a \iff b^a = x \]
\[ \log_b(1) = 0 \iff b^0 = 1 \]
\[ \log_b(b) = 1 \iff b^1 = b \]

\[ \log_b(x) + \log_b(y) = \log_b(xy) \]
\[ \log_b(x) - \log_b(y) = \log_b\left(\frac{x}{y}\right) \]

Differences between logs translate to ratios
Example

The following data show mean task completion times (in ms) of 10 participants for two selection techniques A and B. This is a repeated-measures (within-participants) design.

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tech A</td>
<td>530</td>
<td>600</td>
<td>556</td>
<td>480</td>
<td>578</td>
<td>532</td>
<td>740</td>
<td>590</td>
<td>612</td>
<td>679</td>
</tr>
<tr>
<td>Tech B</td>
<td>511</td>
<td>552</td>
<td>430</td>
<td>455</td>
<td>610</td>
<td>520</td>
<td>731</td>
<td>483</td>
<td>610</td>
<td>539</td>
</tr>
</tbody>
</table>

The research team wants to compare their performance but suspects that task-completion times follow a skewed log-normal distribution.
Example

1. We log-transform the data, using natural logarithms.

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
</tr>
</thead>
</table>

2. We compute 95% CIs by assuming that the sampling distribution of the means follow $t$ distributions:

**Tech A**: 95% CI = [6.29, 6.46]

**Tech B**: 95% CI = [6.18, 6.40]
3. We then transform the CIs back to their original scale (ms) by using the inverse transformation $f(x) = e^x$:

**Tech A:** 95% CI = [535 ms, 640 ms]

**Tech B:** 95% CI = [481 ms, 602 ms]

If we hadn’t log-transformed our data:

**Tech A:** 95% CI = [536 ms, 644 ms]

**Tech B:** 95% CI = [481 ms, 607 ms]
Example

What about the time difference between the two techniques?

In logarithmic scale: 95% CI = [0.008, 0.161]

But if we now know apply the inverse transform $f(x) = e^x$, we get the following:

$$95\% \text{ CI} = [1.008, 1.175]$$

Clearly, these values do not represent seconds!
Remember: Differences in logarithmic scales correspond to ratios in the original scale.

\[ \log_b(x) - \log_b(y) = \log_b \left( \frac{x}{y} \right) \]

\[ \Rightarrow b^{(\log_b(x) - \log_b(y))} = \frac{x}{y} \]

if \( b = e \): 
\[ e^{(\ln(x) - \ln(y))} = \frac{x}{y} \]
Example

We interpret the results as follows:

The mean selection time of Technique A is **109.4%, 95% CI [100.8%, 117.5%]** the mean selection time of Technique B.

Or:

The mean selection time of Technique A is **9.4%, 95% CI [0.8%, 17.5%]** higher than the mean selection time of Technique B.
Example

Graphing the results: choose your axes carefully!

![Graph showing mean selection time for Technique A and Technique B, with error bars for 95% CIs and mean ratio graph with a point at 1.05 and 95% CIs shown as bars.](image)
Working with ratios

Interpreting results in terms of ratios (rather than differences) has several advantages: there are no units, and comparisons are based on relative values.

However, interpretation depends on the specific application context. Unfortunately, many people are not familiar with this approach. You may need to further justify it. Provide a clear interpretation of your results.
Other considerations

The approach discussed above is often characterized as a “naive method,” as it produces biased estimates.

If a log-normal distribution is heavily skewed, it may even result in CIs that do not include the sample mean.

There are several improved methods (Cox, modified Cox, etc.): http://ww2.amstat.org/publications/jse/v13n1/olsson.html

But they are out of the scope of this course.
Bootstrap Confidence Intervals
Monte Carlo methods

What if there are good reasons to doubt the assumption of normality?

What if an analytic solution to calculate CIs is unknown or not available?

One can switch to Monte Carlo methods.
Random sampling without replacement

Randomly drawing samples from a population or a sample by not taking any member more than once.

```r
> data <- c(12, 10, 20, 11, 8, 34, 23, 19)
> sample(data, size = 6, replace = FALSE)
[1] 19 23  8 12 10 20
```
Random sampling with replacement

Randomly drawing samples from a population or a sample, where the same member can be chosen multiple times.

```r
> data <- c(12, 10, 20, 11, 8, 34, 23, 19)
> sample(data, size = 6, replace = TRUE)
[1] 23 20 11 20 12 10
> sample(data, size = 6, replace = TRUE)
> sample(data, size = 10, replace = TRUE)
[1] 34 19 10 19 23  8 12  8 10  8
```

The new sample can be larger than the initial one.
The Bootstrap method

It is a Monte Carlo method for constructing confidence intervals without any assumption about the underlying population distribution.

**Approach:** Sample with replacement from the initial sample to approximate the sampling distribution of a statistic.
Bootstrapping algorithm

Input:

\textit{data}: our initial sample of \( n \) values

\( R \): the number of bootstrap iterations

Output:

\textit{bootstrap}: the bootstrap samples of a statistic
# Input
R <- 3000
data <- c(10, 8, 9, 12, 23, 14, 11, 13, 12, 16, 13, 15, 13, 7)
statistic <- mean # We are interested in the mean

bootstrap <- rep(NA, R) # Init the bootstrap samples

for(i in 1:R){
  s <- sample(data, replace = TRUE) # Sample with replacement
  bootstrap[i] <- statistic(s) # The mean of this last sample
}

# Visualize the approximated sampling distribution
hist(bootstrap)
Bootstrapping algorithm

Step 1 : s = { 7 12 14 8 7 9 7 11 8 12 13 16 23 11 }, bootstrap[i] = 11.28571
Step 2 : s = { 14 11 14 7 14 11 15 16 12 15 13 16 12 12 }, bootstrap[i] = 13
Step 5 : s = { 13 23 13 12 12 7 12 13 11 13 11 13 15 7 }, bootstrap[i] = 12.5
Step 6 : s = { 11 7 13 10 13 14 14 7 12 13 12 14 23 7 }, bootstrap[i] = 12.14286

...
Bootstrapping algorithm

Result:
Bootstrap confidence interval

We can now construct a 95% CI by taking the 2.5th and the 97.5th percentile of the bootstrap distribution.
# The bootstrap algorithm in one line code

```r
bootstrap <- function(data, statistic = mean, R = 3000){
  replicate(R, statistic(sample(data, replace = TRUE)))
}
```

# Constructs the CI of the bootstrap samples

```r
ci <- function(boot.sample, conf.level = .95){
  alpha <- 1 - conf.level

  quantile(boot.sample, c(alpha/2, 1 - alpha/2))
}
```
# Bootstrap of the difference between data1 and data2

```r
bootstrap <- function(data1, data2, statistic = mean, R = 3000){
  replicate(R,
    statistic(sample(data1, replace = TRUE))
    - statistic(sample(data2, replace = TRUE))
  )
}
```

Difference between independent samples
How many bootstrap samples?

The more, the better! Computers makes this process fast.

It is recommended to use more than 1000 bootstrap samples. Some authors recommend more than 3000 samples for 95% CIs or even more.

However, increasing the number of bootstrap samples does not increase the amount of information from the original dataset. A very large number of samples may not necessarily improve results.
The percentile bootstrap

The simple but very generic bootstrap method that we described is known as the **percentile bootstrap**.

In many cases, percentile bootstrap CIs are too narrow, especially when working with small sample sizes.

Also, the method is not recommended in case of:

- **Bias**: the bootstrap distribution is not centered around the parameter of interest - systematically disagrees with the sample

- **Asymmetry**: the bootstrap distribution is skewed.
Alternative methods

**Bias-corrected bootstrap:** Corrects for bias in the bootstrap distribution

**Bias-corrected and accelerated bootstrap (BCa):** Corrects for both bias and skewness in the bootstrap distribution (Efron, 1987).

**Note:** Out of our scope and not simple to implement, but there are specialized R libraries (boot package) to help you.
Bootstrapping: when?

When the underlying distribution is unknown or there are not known (or available) analytical methods for dealing with such a distribution.
From confidence intervals to significance tests
Uses of confidence intervals

1. To provide an estimate of plausible values that a population parameter may take.

2. To support formal inference about a parameter
Example

Let’s review these results:

- **Mean Selection Time (ms):**
  - Technique A: 0.95, 1.00, 1.05, 1.10, 1.15, 1.20
  - Technique B: 1.00, 1.05, 1.10, 1.15, 1.20

- **Mean Ratio (Techn.A / Techn.B):**
  - Error bars: 95% CIs

It allows to conclude that “Technique A is slower than Technique B” (with a certain level of confidence).
Significance testing

Formal inference with a confidence interval is a form of significance testing.

A significance test involves explicitly or implicitly setting up a hypothesis about the value of a population parameter:

Hypothesis example:

“Mean selection time of Technique A is equal to the mean selection time of Technique B.”

or

“The mean ratio of selection times of Techniques A and B is 1”
Null hypothesis

Such hypotheses that make a statement about a hypothetical value of a population parameter (a mean, a mean difference, a mean ratio, etc.) are known as null hypotheses.

The goal of an experiment is commonly to provide evidence against a null hypothesis.
Examples of null hypotheses

"The mean height of men is equal to the mean height of women."

“The mean IQ score of adults lacking enough sleep is 100.”

“Mean selection time with a mouse is equally fast to mean selection time with a trackpad.”
Rejecting a null hypothesis

If the C% confidence interval excludes the hypothesized population value, then the hypothesis is rejected with C% confidence.

(remember that the wording “with C% confidence” reflects a rather informal interpretation and should be used with care)
Example

Here, we reject the null hypothesis with 95% confidence because the 95% CI does not include the value 1.
Example

What about here? Can we reject the null hypothesis concerning the mean difference (in IQ scores) between the two groups?

We cannot reject the null hypothesis.  
**But this does not necessarily mean that the difference is zero!**  
We just do not have enough evidence to say anything more.
Example

And if we reduce the confidence level?

For this level, we can reject the null hypothesis! But is a 80% confidence level high enough?
95% confidence level

It is commonly used as the threshold for rejecting a null hypothesis.

It is not a magic number and there is no reason why a different level (e.g., 92% or 97%) is not used.

It is fixed for consistency reasons and its use reflects a long tradition in science.
Null vs. Alternative hypothesis

Our goal is usually to find enough statistical evidence to reject the null hypothesis ($H_0$) in order to establish an alternative hypothesis ($H_1$).

The alternative hypothesis is the hypothesis of interest, i.e., what the researcher actually seeks to show by rejecting the null hypothesis.
Example

An HCI researcher studies whether visual grouping in menus help users locate menu items faster.

<table>
<thead>
<tr>
<th>Control</th>
<th>Visual Grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>earthquake</td>
<td>earthquake</td>
</tr>
<tr>
<td>thunder</td>
<td>thunder</td>
</tr>
<tr>
<td>skirt</td>
<td>skirt</td>
</tr>
<tr>
<td>eyeliner</td>
<td>eyeliner</td>
</tr>
<tr>
<td>pants</td>
<td>pants</td>
</tr>
<tr>
<td>jacket</td>
<td>jacket</td>
</tr>
<tr>
<td>lipstick</td>
<td>lipstick</td>
</tr>
<tr>
<td>powder</td>
<td>powder</td>
</tr>
<tr>
<td>toffee</td>
<td>toffee</td>
</tr>
<tr>
<td>jellybeans</td>
<td>jellybeans</td>
</tr>
<tr>
<td>caramel</td>
<td>caramel</td>
</tr>
<tr>
<td>hurricane</td>
<td>hurricane</td>
</tr>
</tbody>
</table>

(Brumby & Zhuang, 2015)
Example

Null Hypothesis:

\[ H_0: \text{“Mean selection time is the same for Control and Visual Grouping.”} \]

Alternative Hypothesis:

\[ H_1: \text{“Mean selection is faster with Visual Grouping than with Control.”} \]
Significance tests

Null Hypothesis Significance Testing (NHST) can be supported by confidence internals and significance tests.

Significance tests is a very common research tool. However, they have many limitations. They are also very frequently overused or misused.

Note that significance tests rely on similar assumptions as the ones that we have already discussed for constructing confidence intervals.
$p$ value

The result of a significance test is a probability value $p$, which is commonly known as the $p$ value.

Given an observed value of a statistic (e.g., the mean) of a sample, the $p$ value gives the probability that:

if the null hypothesis $H_0$ was true, then a random sample of the same size would result in a value for the statistic that is equal or more extreme than the observed value.
Example

An experiment studies the IQ scores of people lacking enough sleep.

$H_1$: The mean IQ score of people lacking sleep is lower than 100.

$H_0$: The mean IQ score of people lacking sleep is equal to 100.
Example

Results from a sample of 15 participants are as follows:

94, 91, 96, 100, 103, 88, 98, 103, 87, 93, 97, 105, 99, 91, 88

The mean IQ score of the above sample is $M = 95.53$.

The researchers conduct a significance test and find that $p = .005$

**Interpretation:** If the null hypothesis was true ($M = 100$), then the probability to draw a random sample of 15 people and find that the mean IQ score is equal to or lower than 95.53 is 0.5%.
Interpreting the p value

**Interpretation:** If the null hypothesis was true (M = 100), then the probability to draw a random sample of 15 people and find that the mean IQ score is equal to or lower than 95.53 is 0.5%.

Notice that this is NOT the probability that the researcher’s hypothesis \( H_1 \) is false.
Threshold for rejecting $H_0$

By tradition, the null hypothesis ($H_0$) is rejected when the $p$ value is lower than $\alpha = .05$.

This alpha ($\alpha$) is the same as the one we discussed for $C\%$ confidence intervals, where the confidence level is:

$$C = 100(1 - \alpha)$$

(Clearly, there is a correspondence between significance tests and confidence intervals.)
Back to our example

Results from a sample of 15 participants are as follows:

94, 91, 96, 100, 103, 88, 98, 103, 87, 93, 97, 105, 99, 91, 88

The mean IQ score of the above sample is $M = 95.53$.

The researchers conduct a significance test and find that $p = .005$

Since $p < \alpha (\alpha = .05)$, the null hypothesis is rejected.

The researchers conclude that the lack of sleep results in statistically significantly lower IQ scores.
Statistical significance

Since $p < \alpha (\alpha = .05)$, the null hypothesis is rejected.

The researchers conclude that the lack of sleep results in statistically significantly lower IQ scores.

It is common to say that such an experiment has resulted in a “statistically significant” result.
Interpreting statistical significance

The term “significant” can be very misleading. Statistical significance does not refer to the actual significance of the result!

A significance test may often characterize a very tiny (or practically insignificant) difference as statistically significant!

Make sure that you refer to “statistical significance” when you characterize the results of a study. Don’t use the term “significant” alone.
Back to our example

Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:


The mean IQ score of the above sample is $M = 103.25$.

This mean is higher than 100, so clearly, the data do not support $H_1$.

Can we still test the null hypothesis ($H_0$) to check whether the lack of sleep leads to statistically significantly higher IQ scores?
Answer: It depends on what type of significance tests the researcher uses.

One-sided (or one-directional) tests do not allow for that. If the hypothesized direction is not supported, the hypothesis cannot be rejected.

However, the common practice is to use two-sided (or two-directional) tests. In this case, the hypothesis can be rejected in both directions.
Back to our example

Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:


The mean IQ score of the above sample is $M = 103.25$.

A two-sided significance test results in $p = 0.047$

**Interpretation:** If the null hypothesis was true ($M = 100$), then the probability to draw a random sample of 12 people and find that the absolute difference between the observed mean and 100 is equal to or higher than $\Delta M = 3.25$ is 4.7%.

$p$ values are now interpreted differently
Imagine that another researcher conducts the same experiment with 12 participants and finds the following IQ scores:


The mean IQ score of the above sample is $M = 103.25$.

A two-sided significance test results in $p = .047$

**Conclusion:** The null hypothesis is rejected ($\alpha = .05$). The researchers conclude that the lack of sleep results in statistically significantly higher IQ scores.
Exercise

Write a Monte Carlo simulation in R that computationally calculates the p value for the previous example.

Consider that the distribution of IQ scores is normal.