Resampling Techniques and their Application

Frank Konietschke

Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Organization

- Instructor: Frank Konietschke (Frank.Konietschke@charite.de)
- Assistant: Kerstin Rubarth (Kerstin.Rubarth@charite.de)
- Materials: Available on blackboard (`https://lms.fu-berlin.de`)
- Syllabus: Course outline

Motivation and Examples

An experiment is conducted to study the side effects of a pain reliever on arthritis patients. Out of 480 patients, 60 patients suffered adverse symptom.

- Data: 60 times ’1’, 420 times ’0’

Motivation and Examples

An experiment is conducted to study the side effects of a pain reliever on arthritis patients. Out of 480 patients, 60 patients suffered adverse symptom.

- Data: 60 times ’1’, 420 times ’0’
- We can compute that 12.5% of patients suffered from side effects. Is that a good estimate?

Motivation and Examples

An experiment is conducted to study the side effects of a pain reliever on arthritis patients. Out of 480 patients, 60 patients suffered adverse symptom.

- Data: 60 times ’1’, 420 times ’0’
- We can compute that 12.5% of patients suffered from side effects. Is that a good estimate?
- Confidence interval is required.

Motivation and Examples

An experiment is conducted to study the side effects of a pain reliever on arthritis patients. Out of 480 patients, 60 patients suffered adverse symptom.

- Data: 60 times ’1’, 420 times ’0’
- We can compute that 12.5% of patients suffered from side effects. Is that a good estimate?
- Confidence interval is required.
- Quality of the confidence interval?

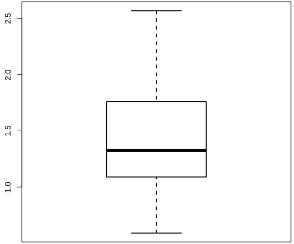
Motivation and Examples

An experiment is conducted to study the side effects of a pain reliever on arthritis patients. Out of 480 patients, 60 patients suffered adverse symptom.

- Data: 60 times '1', 420 times '0'
- We can compute that 12.5% of patients suffered from side effects. Is that a good estimate?
- Confidence interval is required.
- Quality of the confidence interval?
- We will learn quality criteria and how to investigate them ($\Rightarrow$ **Simulations**)

Motivation and Examples-II

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of **n=36** bottles and obtains:

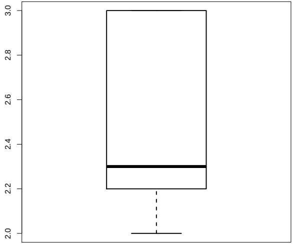



```
x=c(
0.59, 1.23, 1.00, 0.84, 0.88, 1.71,
1.81, 1.84, 2.03, 1.39, 1.30, 1.31,
1.96, 1.33, 2.57, 1.19, 1.01, 2.06,
1.32, 1.55, 1.28, 0.93, 1.63, 1.24,
1.83, 1.81, 0.94, 1.46, 1.25, 1.56,
0.61, 0.83, 1.17, 2.24, 1.68, 1.51)
```

- Study aim: **Estimation** and **Testing**. Quality of the method? Can we do better using resampling?

Motivation and Examples-III

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail.




```
x=c(
2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
```

- Study aim: **Estimation** and **Testing**. Quality of the method? Can we do better using resampling?

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Motivation and Examples-IV

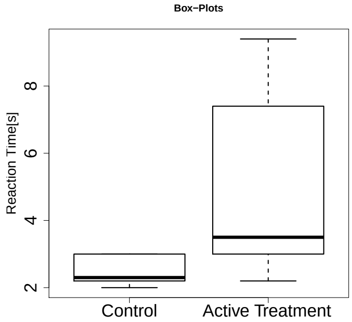
- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

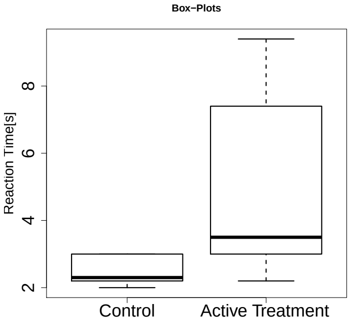
Boxplot



Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot

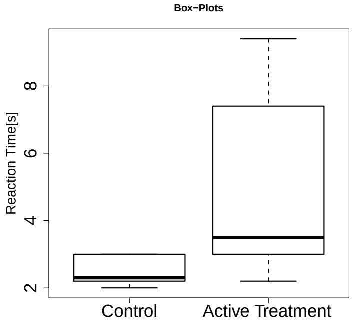


Study Aims?

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot



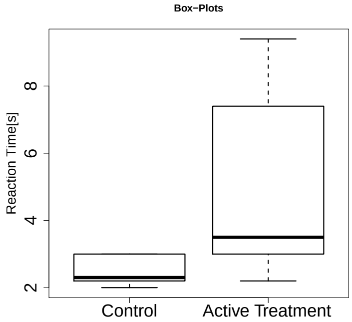
Study Aims?

- Compare the groups (**Testing** and **Estimation**). But

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot



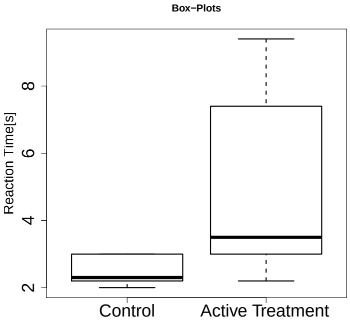
Study Aims?

- Compare the groups (**Testing** and **Estimation**). But
- Sample sizes are very small

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot



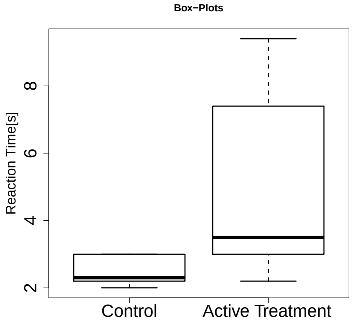
Study Aims?

- Compare the groups (**Testing** and **Estimation**). But
- Sample sizes are very small
- Data might come from skewed distributions

Motivation and Examples-IV

- $N = 20$ mice were randomized to $a = 2$ treatment groups
- $n_1 = 10$ control; $n_2 = 10$ active
- Response: reaction time [sec]

Boxplot



Study Aims?

- Compare the groups (**Testing** and **Estimation**). But
- Sample sizes are very small
- Data might come from skewed distributions
- Quality of available methods? Can we improve using resampling?

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

  - **Estimation**

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

  - **Estimation**
    - Parameters

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

  - **Estimation**
    - Parameters
    - Standard errors,...

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution
  - How accurate is the method?

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution
  - How accurate is the method?
- **Tests of significance**

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution
  - How accurate is the method?
- **Tests of significance**
  - Hypotheses

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution
  - How accurate is the method?
- **Tests of significance**
  - Hypotheses
  - Distribution of statistics?

Motivation and Review - I

- Statistical inference: Methods for drawing conclusions about a population from sample data.
- Three of the most common types of statistical inference:

- **Estimation**
  - Parameters
  - Standard errors,...
  - Goal: Estimate a population parameter
- **Confidence intervals**
  - Accuracy depends on the distribution
  - How accurate is the method?
- **Tests of significance**
  - Hypotheses
  - Distribution of statistics?
  - Accuracy of the methods?

Motivation and Review - II

- **Estimation:** Some parameters are very hard to estimate. Sometimes formulas do not even exist. Think about variances of empirical correlation coefficients, third moments, etc.

Motivation and Review - II

- **Estimation:** Some parameters are very hard to estimate. Sometimes formulas do not even exist. Think about variances of empirical correlation coefficients, third moments, etc.
- **Confidence intervals**: Accuracy depends on the distribution. How **accurate** is the method?

Motivation and Review - II

- **Estimation:** Some parameters are very hard to estimate. Sometimes formulas do not even exist. Think about variances of empirical correlation coefficients, third moments, etc.
- **Confidence intervals**: Accuracy depends on the distribution. How **accurate** is the method?
- **Tests of significance**: Control of the **Type-1 error rate** of the method. Most methods rely on large sample sizes. What if sample sizes are small?

Motivation and Review - II

- **Estimation:** Some parameters are very hard to estimate. Sometimes formulas do not even exist. Think about variances of empirical correlation coefficients, third moments, etc.
- **Confidence intervals**: Accuracy depends on the distribution. How **accurate** is the method?
- **Tests of significance**: Control of the **Type-1 error rate** of the method. Most methods rely on large sample sizes. What if sample sizes are small?
- We will explore ***Resampling methods*** as modern ways to counter these issues.

Motivation and Review - III

- **Resampling methods**

Motivation and Review - III

- **Resampling methods**
  - Cannot be computed without a computer (we use R)

Motivation and Review - III

- **Resampling methods**
  - Cannot be computed without a computer (we use R)
  - The quality of statistical methods can only be judged by **simulations**

Motivation and Review - III

- **Resampling methods**
  - Cannot be computed without a computer (we use R)
  - The quality of statistical methods can only be judged by **simulations**
  - Learning simulations will therefore play a major role in this class

Motivation and Review - III

- **Resampling methods**
  - Cannot be computed without a computer (we use R)
  - The quality of statistical methods can only be judged by **simulations**
  - Learning simulations will therefore play a major role in this class
  - Understanding of statistical testing theory, p-value etc is fundamental and we therefore refresh today

Review: Basic Idea of Tests of Significance

Example: Every week, two teams ("In-laws" and the "Outlaws") get together and play a football game. They decide who receives the kickoff by a flip of a **fair coin** provided by the In-laws (heads = outlaws receive, tails = in-laws receive).

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the "fair" coin really is fair.

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the “fair” coin really is fair.
- They then flip the coin **100 times** on the sidelines, and get **35** heads.

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the “fair” coin really is fair.

- They then flip the coin **100 times** on the sidelines, and get **35** heads.

  - How many heads did they **expect**,
    if the coin was fair?

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the “fair” coin really is fair.
- They then flip the coin **100 times** on the sidelines, and get **35** heads.
  - How many heads did they **expect**, if the coin was fair?
    - They expected 50 heads but got 35 heads.

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the "fair" coin really is fair.

- They then flip the coin **100 times** on the sidelines, and get **35** heads.

  - How many heads did they **expect**, if the coin was fair?
    - They expected 50 heads but got 35 heads.

  - Two possible explanations:

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the "fair" coin really is fair.

- They then flip the coin **100 times** on the sidelines, and get **35** heads.

  - How many heads did they **expect**, if the coin was fair?
    - They expected 50 heads but got 35 heads.

- Two possible explanations:
  1. The coin is fair and this outcome just happened by chance (bad sample!)

Review: Hypotheses

- Well, after losing the coin toss for a long time, the Outlaws have become suspicious! They decide to see if the "fair" coin really is fair.

- They then flip the coin **100 times** on the sidelines, and get **35** heads.

  - How many heads did they **expect**, if the coin was fair?
    - They expected 50 heads but got 35 heads.

  - Two possible explanations:
    1. The coin is fair and this outcome just happened by chance (bad sample!)
    2. For a fair coin, this outcome is so unlikely (extreme) that we can conclude that the coin is not fair.

Hypotheses - II

- So, we have two competing hypotheses:

Hypotheses - II

- So, we have two competing hypotheses:
  - **Null Hypothesis** (H_0): The chance of heads = 50% (This is the statement of “status quo”: It says the observed outcome is different from the expected just by chance variation).

$$H_0 : p = p_0 \geq 0.5$$

Hypotheses - II

- So, we have two competing hypotheses:
  - **Null Hypothesis** (H_0): The chance of heads = 50% (This is the statement of “status quo”: It says the observed outcome is different from the expected just by chance variation).

$$H_0 : p = p_0 \geq 0.5$$

  - **Alternative Hypothesis** (H_1):

$$H_1 : p < 50\%$$

    The chance of heads $<$ 50%. (This is the statement we’d like to prove. It says chance variation is not enough to explain the outcome).

Test Statistic

- Philosophy of Hypothesis Test: “Innocent until proven guilty”. Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.

Test Statistic

- Philosophy of Hypothesis Test: “Innocent until proven guilty”. Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.

- To decide between the null and alternative hypothesis, we find two things:

Test Statistic

- Philosophy of Hypothesis Test: “Innocent until proven guilty”. Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.
- To decide between the null and alternative hypothesis, we find two things:
  - how **much different** is the **observed outcome** from the **expected outcome** if the null hypothesis is true (test statistic)

Test Statistic

- Philosophy of Hypothesis Test: "Innocent until proven guilty". Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.
- To decide between the null and alternative hypothesis, we find two things:
  - how **much different** is the **observed outcome** from the **expected outcome** if the null hypothesis is true (test statistic)
  - what is the chance of that type of difference occuring if null hypothesis is true (p-value).

Test Statistic

- Philosophy of Hypothesis Test: “Innocent until proven guilty”. Assume null hypothesis to be true and see if there is convincing evidence in the data to prove otherwise.
- To decide between the null and alternative hypothesis, we find two things:
  - how **much different** is the **observed outcome** from the **expected outcome** if the null hypothesis is true (test statistic)
  - what is the chance of that type of difference occuring if null hypothesis is true (p-value).
- If such a difference is very unlikely to occur if null hypothesis is true, then we reject the null hypothesis.

Test Statistic - II

- Test Statistic

$$Z = \frac{\text{Observed} - \text{Expected under } H_0}{\text{Standard Error}}$$

Test Statistic - II

- Test Statistic

$$Z = \frac{\text{Observed} - \text{Expected under } H_0}{\text{Standard Error}}$$

- In-laws vs. Outlaws: As we are dealing with %, our test statistic under H_0:

$$Z = \frac{\text{oberved } \% - EV(\%)}{SE(\%)} = (0.35 - 0.5)/\sqrt{0.25/100} = -3$$

Test Statistic - II

- Test Statistic

$$Z = \frac{\text{Observed} - \text{Expected under } H_0}{\text{Standard Error}}$$

- In-laws vs. Outlaws: As we are dealing with %, our test statistic under H_0:

$$Z = \frac{\text{oberved } \% - EV(\%)}{SE(\%)} = (0.35 - 0.5)/\sqrt{0.25/100} = -3$$

- So, the observed outcome is 3 SD below the expected outcome.

Test Statistic - II

- Test Statistic

$$Z = \frac{\text{Observed} - \text{Expected under } H_0}{\text{Standard Error}}$$

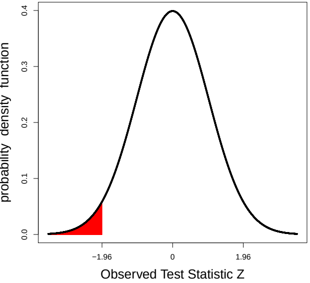
- In-laws vs. Outlaws: As we are dealing with %, our test statistic under H_0:

$$Z = \frac{\text{oberved }\% - EV(\%)}{SE(\%)} = (0.35 - 0.5)/\sqrt{0.25/100} = -3$$

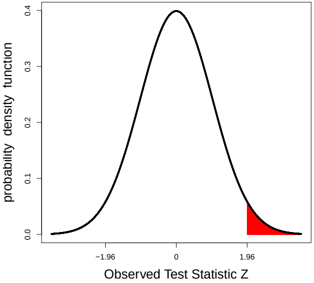
- So, the observed outcome is 3 SD below the expected outcome.
- How likely is the outcome? ($\Rightarrow$ Distribution of Z)

Computation of P-Value

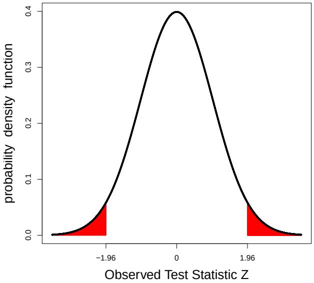
$H_0 : p = p_0$ versus
$H_1 : p < p_0$



$H_1 : p > p_0$



$H_1 : p \neq p_0$



P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.

P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
  - Note: “Extreme” is in the direction of H_1.

P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
  - Note: "Extreme" is in the direction of H_1.
  - This p-values tell us: If H_0 is true (i.e., the coin is fair), then the chance of getting 35% or less heads in 100 tosses is ......

P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
  - Note: "Extreme" is in the direction of H_1.
  - This p-values tell us: If H_0 is true (i.e., the coin is fair), then the chance of getting 35% or less heads in 100 tosses is ......
  - What does this tell us about H_0?

P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
  - Note: “Extreme” is in the direction of H_1.
  - This p-values tell us: If H_0 is true (i.e., the coin is fair), then the chance of getting 35% or less heads in 100 tosses is ......
  - What does this tell us about H_0?
- As our p-value is so small, we conclude that we have enough evidence to reject the null.

P-Value

- **P-value**: The chance of getting a test statistic as extreme or more extreme than what we observed if H_0 is true.
  - Note: "Extreme" is in the direction of H_1.
  - This p-values tell us: If H_0 is true (i.e., the coin is fair), then the chance of getting 35% or less heads in 100 tosses is ......
  - What does this tell us about H_0?
- As our p-value is so small, we conclude that we have enough evidence to reject the null.
- Does this prove that H_0 is false?

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called “significance level” and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:
  - State H_0 and H_1 (H_1 is what we are interested in proving. One-sided or two-sided).

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:
  - State H_0 and H_1 (H_1 is what we are interested in proving. One-sided or two-sided).
  - Calculate the appropriate test statistic (assuming H_0 to be true).

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:
  - State H_0 and H_1 (H_1 is what we are interested in proving. One-sided or two-sided).
  - Calculate the appropriate test statistic (assuming H_0 to be true).
  - Calculate the p-value.

Steps for Test of Significance

- For p-value, how small is small enough to reject H_0? This cutoff is called "significance level" and is denoted by α.
- Typical choices of α are 0.05 or 0.01.
- Steps in Hypothesis Test:
  - State H_0 and H_1 (H_1 is what we are interested in proving. One-sided or two-sided).
  - Calculate the appropriate test statistic (assuming H_0 to be true).
  - Calculate the p-value.
  - Use the p-value and the given significance level to draw the conclusion:
    If p-value $\leq \alpha$, reject H_0. If p-value $> \alpha$, do not reject H_0.

Example

An experiment is conducted to study the side effects of pain reliever arthritis patients. Out of 480 patients, 60 suffered from adverse events. Is there evidence at 5% level that the proportion of all Bioxx users suffering from adverse symptoms is more than 10%? What if the level was 1%?

- Hypothesis: $H_0 : p \leq p_0 = 10\%$ vs. $H_1 : p > 10\%$

Example

An experiment is conducted to study the side effects of pain reliever arthritis patients. Out of 480 patients, 60 suffered from adverse events. Is there evidence at 5% level that the proportion of all Bioxx users suffering from adverse symptoms is more than 10%? What if the level was 1%?

- Hypothesis: $H_0 : p \leq p_0 = 10\%$ vs. $H_1 : p > 10\%$
- Estimator: $\widehat{p} = \frac{60}{480} = 12.5\%$

Example

An experiment is conducted to study the side effects of pain reliever arthritis patients. Out of 480 patients, 60 suffered from adverse events. Is there evidence at 5% level that the proportion of all Bioxx users suffering from adverse symptoms is more than 10%? What if the level was 1%?

- Hypothesis: $H_0 : p \leq p_0 = 10\%$ vs. $H_1 : p > 10\%$
- Estimator: $\widehat{p} = \frac{60}{480} = 12.5\%$
- Standard error: $SE(\widehat{p}) = \sqrt{\frac{10*90}{480}}$ (under H_0 in %)

Example

An experiment is conducted to study the side effects of pain reliever arthritis patients. Out of 480 patients, 60 suffered from adverse events. Is there evidence at 5% level that the proportion of all Bioxx users suffering from adverse symptoms is more than 10%? What if the level was 1%?

- Hypothesis: $H_0 : p \leq p_0 = 10\%$ vs. $H_1 : p > 10\%$
- Estimator: $\widehat{p} = \frac{60}{480} = 12.5\%$
- Standard error: $SE(\widehat{p}) = \sqrt{\frac{10*90}{480}}$ (under H_0 in %)
- Test statistic: $Z = \frac{12.5-10}{\sqrt{\frac{10*90}{480}}} = 1.83$

Example

An experiment is conducted to study the side effects of pain reliever arthritis patients. Out of 480 patients, 60 suffered from adverse events. Is there evidence at 5% level that the proportion of all Bioxx users suffering from adverse symptoms is more than 10%? What if the level was 1%?

- Hypothesis: $H_0 : p \leq p_0 = 10\%$ vs. $H_1 : p > 10\%$
- Estimator: $\widehat{p} = \frac{60}{480} = 12.5\%$
- Standard error: $SE(\widehat{p}) = \sqrt{\frac{10*90}{480}}$ (under H_0 in %)
- Test statistic: $Z = \frac{12.5-10}{\sqrt{\frac{10*90}{480}}} = 1.83$
- p-value: : Area to the right of Z under the standard normal curve, Here, $p = 0.034$

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) ↓	Reject H_0	Do not reject H_0
H_0		
H_1		

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) $\downarrow$	Reject H_0	Do not reject H_0
H_0		
H_1		

- **Type I error:** Reject H_0 when H_0 is actually true.

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) $\downarrow$	Reject H_0	Do not reject H_0
H_0		
H_1		

- **Type I error:** Reject H_0 when H_0 is actually true.
- **Type II error:** Not reject H_0 when H_1 is actually true.

Types of Error in Hypothesis Tests

- When we carry out a test, what types of errors we can make?

	Decision	
Truth (unknown) $\downarrow$	Reject H_0	Do not reject H_0
H_0		
H_1		

- **Type I error:** Reject H_0 when H_0 is actually true.
- **Type II error:** Not reject H_0 when H_1 is actually true.
- These errors are defined conditional on the true status (H_0 or H_1).

Types of Error in Hypothesis Tests - II

- Bioxx Ex: What are the type I and type II errors?

Types of Error in Hypothesis Tests - II

- Bioxx Ex: What are the type I and type II errors?
- Ex: In a court trial, H_0: Person is Innocent vs. H_1: Person is Guilty

  Type I error:

  Type II error:

Types of Error in Hypothesis Tests - II

- Bioxx Ex: What are the type I and type II errors?
- Ex: In a court trial, H_0: Person is Innocent vs. H_1: Person is Guilty

  Type I error:

  Type II error:

- We cannot eliminate the possibility of errors because our decision is based on a sample, and not the whole population.

Types of Error in Hypothesis Tests - II

- Bioxx Ex: What are the type I and type II errors?
- Ex: In a court trial, H_0: Person is Innocent vs. H_1: Person is Guilty

  Type I error:

  Type II error:

- We cannot eliminate the possibility of errors because our decision is based on a sample, and not the whole population.
- But we can have some control over the chances of making errors.

Probabilities of Type I and Type II Errors & P-value

- Fact: P(Type I error) = α (the significance level used).

Probabilities of Type I and Type II Errors & P-value

- Fact: P(Type I error) = α (the significance level used).
- Consider the court trial: If we try to decrease P(type I error) (by concluding persons under trial innocent more often), what happens to P(type II error)?

Probabilities of Type I and Type II Errors & P-value

- Fact: P(Type I error) = α (the significance level used).
- Consider the court trial: If we try to decrease P(type I error) (by concluding persons under trial innocent more often), what happens to P(type II error)?
- Type I error is typically considered more severe and so it's chance is controlled by setting α to be a small $\#$.

Probabilities of Type I and Type II Errors & P-value

- Fact: P(Type I error) = α (the significance level used).
- Consider the court trial: If we try to decrease P(type I error) (by concluding persons under trial innocent more often), what happens to P(type II error)?
- Type I error is typically considered more severe and so it's chance is controlled by setting α to be a small $\#$.
- After setting α to be small, the studies are designed in such a way that type II error is minimized.

Probabilities of Type I and Type II Errors & P-value

- Fact: P(Type I error) = α (the significance level used).
- Consider the court trial: If we try to decrease P(type I error) (by concluding persons under trial innocent more often), what happens to P(type II error)?
- Type I error is typically considered more severe and so it's chance is controlled by setting α to be a small $\#$.
- After setting α to be small, the studies are designed in such a way that type II error is minimized.
- What can we do to have both type I and type II error rates acceptably small?

Probabilities of Type I and Type II Errors & P-value - II

- Interpretation of P-value: P-value is **NOT** the probability that H_0 is true. H_0 is either true or not true. It does not vary from sample to sample. P-value tells how likely it is to get the observed sample (or something more extreme) if H_0 is true (Note: H_0 is held fixed). Smaller the P-value, stronger the evidence against H_0.

Test of Significance for Population Average: t-Test

- Setting: Random sample from a **normal** population with mean μ and SD σ

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

Test of Significance for Population Average: *t*-Test

- Setting: Random sample from a **normal** population with mean μ and SD σ

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu >$ or $<$ or $\neq \mu_0$

Test of Significance for Population Average: t-Test

- Setting: Random sample from a **normal** population with mean μ and SD σ

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu >$ or $<$ or $\neq \mu_0$
- Estimators:

$$\overline{X}_{\cdot} = \frac{1}{n} \sum_{k=1}^{n} X_k \qquad \widehat{\sigma}^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X}_{\cdot})^2$$

Test of Significance for Population Average: t-Test

- Setting: Random sample from a **normal** population with mean μ and SD σ

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu >$ or $<$ or $\neq \mu_0$
- Estimators:

$$\overline{X}_{\cdot} = \frac{1}{n} \sum_{k=1}^{n} X_k \qquad \widehat{\sigma}^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X}_{\cdot})^2$$

- Test Statistic and Confidence Interval:

$$T = \sqrt{n} \frac{\overline{X}_{\cdot} - \mu_0}{\widehat{\sigma}} \sim t_{n-1}, \quad CI = \left[\overline{X}_{\cdot} \mp \frac{t_{n-1}(1-\alpha/2)}{\sqrt{n}} \widehat{\sigma} \right]$$

Test of Significance for Population Average: t-Test

- Setting: Random sample from a **normal** population with mean μ and SD σ

$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu >$ or $<$ or $\neq \mu_0$
- Estimators:

$$\overline{X}_{\cdot} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_{\cdot})^2$$

- Test Statistic and Confidence Interval:

$$T = \sqrt{n}\frac{\overline{X}_{\cdot} - \mu_0}{\widehat{\sigma}} \sim t_{n-1}, \quad CI = \left[\overline{X}_{\cdot} \mp \frac{t_{n-1}(1-\alpha/2)}{\sqrt{n}}\widehat{\sigma}\right]$$

- Reject H_0, if $|T| \geq t_{n-1}(1-\alpha/2)$. Or compute the p-value

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$
- p-value: 2 times the area to the **left** of T under the t curve with 35 df. (Here, p=0.27)

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$
- p-value: 2 times the area to the **left** of T under the t curve with 35 df. (Here, p=0.27)
- Is this method of high quality?

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$
- Is this CI of high quality?

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$
- Is this CI of high quality?
- Next class: **Simulation of type-1 error**