Lecture 7: Hypothesis testing part I

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 25, 2021

Today

1. Terminology
   - Experiment and parameter of interest
   - Null hypothesis and alternative hypothesis
   - Test statistic
   - Null distribution $f(s \mid H_0)$
   - Significance level α, power and p-value
2. Example
3. p-hacking
4. Summary

Learning outcome

- Be able to explain the following terminology
  - Null hypothesis H_0 and alternative hypothesis H_A
  - Test statistic s
  - Null distribution $f(s \mid H_0)$
  - Significance level α and power
  - p-value
- Be able to design and interpret the one-sample z-test
- Be able to explain the concept of p-hacking

Today

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

- If company A tested drug D on 30 ducks and the average weight loss after one month is 2.2 kg, would you buy drug D instead of regular diet control?

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

- If company A tested drug D on 30 ducks and the average weight loss after one month is 2.2 kg, would you buy drug D instead of regular diet control?
- What if company A tested drug D on 30 ducks and the average weight loss after one month is 2.3 kg? Would you buy drug D instead of regular diet control in this case?

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

- If company A tested drug D on 30 ducks and the average weight loss after one month is 2.2 kg, would you buy drug D instead of regular diet control?
- What if company A tested drug D on 30 ducks and the average weight loss after one month is 2.3 kg? Would you buy drug D instead of regular diet control in this case?
- What if company A tested drug D on 100 ducks and the average weight loss after one month is 2.3 kg?

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

- If company A tested drug D on 30 ducks and the average weight loss after one month is 2.2 kg, would you buy drug D instead of regular diet control?
- What if company A tested drug D on 30 ducks and the average weight loss after one month is 2.3 kg? Would you buy drug D instead of regular diet control in this case?
- What if company A tested drug D on 100 ducks and the average weight loss after one month is 2.3 kg?
- Now company B tested drug E on 30 ducks and the average weight loss after one month is 2.5 kg, while drug D results in 2.3 kg weight loss with the same setup, would you buy drug E instead of drug D?

Example

If you control the diet of your ducks, they lose 2.1 kg after one month on average

- Company A has developed a drug D to help ducks lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E and they claim that drug E is more effective than drug D **on average**

You need to help your chonker ducks lose weight. Which drug do you buy? Or should you just control their diet?

- If company A tested drug D on 30 ducks and the average weight loss after one month is 2.2 kg, would you buy drug D instead of regular diet control?
- What if company A tested drug D on 30 ducks and the average weight loss after one month is 2.3 kg? Would you buy drug D instead of regular diet control in this case?
- What if company A tested drug D on 100 ducks and the average weight loss after one month is 2.3 kg?
- Now company B tested drug E on 30 ducks and the average weight loss after one month is 2.5 kg, while drug D results in 2.3 kg weight loss with the same setup, would you buy drug E instead of drug D?

How would you make your decision?

Hypothesis

- **Hypothesis:** a hypothesis is a proposed explanation for a phenomenon (wikipedia)

Hypothesis

- **Hypothesis:** a hypothesis is a proposed explanation for a phenomenon (wikipedia)
- **Statistical hypothesis:** a proposed distribution that explains a set of random variables

Hypothesis

- **Hypothesis:** a hypothesis is a proposed explanation for a phenomenon (wikipedia)
- **Statistical hypothesis:** a proposed distribution that explains a set of random variables
- **Hypothesis testing in statistics:** we want to decide if it is likely that a random variable follows the distribution proposed by the statistical hypothesis

Hypothesis

- **Hypothesis:** a hypothesis is a proposed explanation for a phenomenon (wikipedia)
- **Statistical hypothesis:** a proposed distribution that explains a set of random variables
- **Hypothesis testing in statistics:** we want to decide if it is likely that a random variable follows the distribution proposed by the statistical hypothesis
    - The test is based on sample statistics, which are computed from data

Hypothesis

- **Hypothesis:** a hypothesis is a proposed explanation for a phenomenon (wikipedia)
- **Statistical hypothesis:** a proposed distribution that explains a set of random variables
- **Hypothesis testing in statistics:** we want to decide if it is likely that a random variable follows the distribution proposed by the statistical hypothesis
  - The test is based on sample statistics, which are computed from data
  - Hypothesis + data $\rightarrow$ decision on rejecting/not rejecting the hypothesis

Hypothesis testing: a list to go through

- A "boring" statement
- Experiment
- Data x, random variable X
- Parameter of interest θ
- Parameter estimate $\hat{\theta}$
- Null hypothesis H_0
- Alternative hypothesis H_A
- Test statistic s
- Null distribution $f(s \mid H_0)$
- Significance level α
- p-value

Experiment and parameter of interest

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement?

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec):

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec):

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.
  - **Parameter of interest** (5 sec):

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.
  - **Parameter of interest** (5 sec): the average weight loss μ_D

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
    - **A "boring" statement**: drug D is not more effective than regular diet on average
    - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
    - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.
    - **Parameter of interest** (5 sec): the average weight loss μ_D
    - **Parameter estimate** (5 sec):

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.
  - **Parameter of interest** (5 sec): the average weight loss μ_D
  - **Parameter estimate** (5 sec): $\hat{\mu_D} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Experiment design

- Before formulating the statistical hypothesis, we need **a "boring" statement**: a claim that we would like to test against, e.g. drug D is not more effective than regular diet on average; drug E works the same as drug D on average
- How do we test the "boring" statement? We design and run **experiments** to collect evidence (**data**)
- Example 1: recall if you control the diet of your ducks, they lose 2.1 kg after one month on average
  - **A "boring" statement**: drug D is not more effective than regular diet on average
  - **Experiment** (5 sec): test drug D on N chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss after one month for $i = 1, \cdots, N$; random variable - X_i i.i.d.
  - **Parameter of interest** (5 sec): the average weight loss μ_D
  - **Parameter estimate** (5 sec): $\hat{\mu_D} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

  Then we can use $\bar{x}$ to approximate μ_D and check if it is greater than diet control (2.1 kg)

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec):

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec):

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E
  - **Parameter of interest** (5 secs):

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E
  - **Parameter of interest** (5 secs): the average weight loss μ_D and μ_E for drug D and E, respectively

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E
  - **Parameter of interest** (5 secs): the average weight loss μ_D and μ_E for drug D and E, respectively
  - **Parameter estimate** (5 secs):

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E
  - **Parameter of interest** (5 secs): the average weight loss μ_D and μ_E for drug D and E, respectively
  - **Parameter estimate** (5 secs): $\hat{\mu}_D = \bar{x} = \frac{1}{N_D} \sum_{i=1}^{N_D} x_i$ and $\hat{\mu_E} = \bar{y} = \frac{1}{N_E} \sum_{j=1}^{N_E} y_j$

Experiment design (cont.)

- Example 2:
  - **A "boring" statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): test drug D on N_D chonker ducks and record the average weight loss after one month; test drug E on another N_E chonker ducks and record the average weight loss after one month
  - **Data** and **random variable** (5 sec): data - x_i weight loss using drug D after one month for $i = 1, \cdots, N_D$; random variable - X_i i.i.d.; likewise, we have data y_j and random variable Y_j for drug E
  - **Parameter of interest** (5 secs): the average weight loss μ_D and μ_E for drug D and E, respectively
  - **Parameter estimate** (5 secs): $\hat{\mu}_D = \bar{x} = \frac{1}{N_D}\sum_{i=1}^{N_D} x_i$ and $\hat{\mu_E} = \bar{y} = \frac{1}{N_E}\sum_{j=1}^{N_E} y_j$

  Then we use $\bar{x}$ and $\bar{y}$ to approximate μ_D and μ_E to see if they are the same

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the "boring" statement, we **reject the statement**; otherwise, we **do not reject the statement**

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the "boring" statement, we **reject the statement**; otherwise, we **do not reject the statement**
- But we can never prove or accept the statement - we can only **reject** a statement by showing counterexamples

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the "boring" statement, we **reject the statement**; otherwise, we **do not reject the statement**
- But we can never prove or accept the statement - we can only **reject** a statement by showing counterexamples
- The logic here is: if a statement is true, then the evidence should support the statement

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the "boring" statement, we **reject the statement**; otherwise, we **do not reject the statement**
- But we can never prove or accept the statement - we can only **reject** a statement by showing counterexamples
- The logic here is: if a statement is true, then the evidence should support the statement $\iff$ if the evidence does not support the statement, the statement is considered false

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the "boring" statement, we **reject the statement**; otherwise, we **do not reject the statement**
- But we can never prove or accept the statement - we can only **reject** a statement by showing counterexamples
- The logic here is: if a statement is true, then the evidence should support the statement $\iff$ if the evidence does not support the statement, the statement is considered false $\not\Longleftrightarrow$ if the evidence supports the statement, the statement must be true

Null hypothesis and alternative hypothesis

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

- **Alternative hypothesis H_A**: a complementary alternative explanation to the "boring" statement

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

- **Alternative hypothesis H_A**: a complementary alternative explanation to the "boring" statement
  - Example 1: drug D is more effective than regular diet on average (5 sec)

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

- **Alternative hypothesis H_A**: a complementary alternative explanation to the "boring" statement
  - Example 1: drug D is more effective than regular diet on average (5 sec)

$$H_A : \mu_D > 2.1$$

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis H_0**: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

- **Alternative hypothesis H_A**: a complementary alternative explanation to the "boring" statement
  - Example 1: drug D is more effective than regular diet on average (5 sec)

$$H_A : \mu_D > 2.1$$

  - Example 2: drug E and drug D do not work the same on average (5 sec)

Hypotheses H_0 and H_A

- **Statistical hypothesis**: a proposed distribution - a statement about the **parameter of interest**
- **Null hypothesis** H_0: the "boring" statement translated into a mathematical expression
  - Example 1: drug D is not more effective than regular diet on average

$$H_0 : \mu_D = 2.1$$

  - Example 2: drug E and drug D work the same on average (5 sec)

$$H_0 : \mu_D = \mu_E$$

- **Alternative hypothesis** H_A: a complementary alternative explanation to the "boring" statement
  - Example 1: drug D is more effective than regular diet on average (5 sec)

$$H_A : \mu_D > 2.1$$

  - Example 2: drug E and drug D do not work the same on average (5 sec)

$$H_A : \mu_D \neq \mu_E$$

Hypotheses H_0 and H_A (cont.)

Questions:

- Question 1: why are $H_A : \mu_D > 2.1$ and $H_0 : \mu_D = 2.1$ complementary to each other? What about $H_A : \mu_D < 2.1$?

Hypotheses H_0 and H_A (cont.)

Questions:

- Question 1: why are $H_A : \mu_D > 2.1$ and $H_0 : \mu_D = 2.1$ complementary to each other? What about $H_A : \mu_D < 2.1$?
  Answer: an implicit assumption here is that μ_D will not be smaller than 2.1

Hypotheses H_0 and H_A (cont.)

Questions:

- Question 1: why are $H_A : \mu_D > 2.1$ and $H_0 : \mu_D = 2.1$ complementary to each other? What about $H_A : \mu_D < 2.1$?
  Answer: an implicit assumption here is that μ_D will not be smaller than 2.1
- Question 2: can H_0 and H_A be ANYTHING I want? Like a magic mirror!?

Hypotheses H_0 and H_A (cont.)

Questions:

- Question 1: why are $H_A : \mu_D > 2.1$ and $H_0 : \mu_D = 2.1$ complementary to each other? What about $H_A : \mu_D < 2.1$?
  Answer: an implicit assumption here is that μ_D will not be smaller than 2.1
- Question 2: can H_0 and H_A be ANYTHING I want? Like a magic mirror!?
  Answer: no

Hypotheses H_0 and H_A (cont.)

Questions:

- Question 1: why are $H_A : \mu_D > 2.1$ and $H_0 : \mu_D = 2.1$ complementary to each other? What about $H_A : \mu_D < 2.1$?
  Answer: an implicit assumption here is that μ_D will not be smaller than 2.1
- Question 2: can H_0 and H_A be ANYTHING I want? Like a magic mirror!?
  Answer: no
- Follow up question: what are the choices for H_0 and H_A?

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis** H_0: two cases

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis H_0**: two cases
    - **One-sample test**: to test a data distribution against a theoretical probability distribution, i.e. for a given constant c

$$H_0 : \theta = c$$

For example, is a binary classifier more accurate than random?

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis** H_0: two cases
  - **One-sample test**: to test a data distribution against a theoretical probability distribution, i.e. for a given constant c

$$H_0 : \theta = c$$

For example, is a binary classifier more accurate than random? $H_0 : p = 50\%$

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis** H_0: two cases
  - **One-sample test**: to test a data distribution against a theoretical probability distribution, i.e. for a given constant c

    $$H_0 : \theta = c$$

    For example, is a binary classifier more accurate than random? $H_0 : p = 50\%$
  - **Two-sample test**: to test a data distribution against another data distribution, i.e.

    $$H_0 : \theta_1 = \theta_2$$

    For example, is classifier A better than classifier B?

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis H_0**: two cases
  - **One-sample test**: to test a data distribution against a theoretical probability distribution, i.e. for a given constant c

    $$H_0 : \theta = c$$

    For example, is a binary classifier more accurate than random? $H_0 : p = 50\%$
  - **Two-sample test**: to test a data distribution against another data distribution, i.e.

    $$H_0 : \theta_1 = \theta_2$$

    For example, is classifier A better than classifier B? $H_0 : p_A = p_B$

Choices for H_0

- In this course, we only deal with null hypotheses **with an equal sign** in them - only one fixed choice for the distribution proposed by H_0
- **Null hypothesis H_0**: two cases
  - **One-sample test**: to test a data distribution against a theoretical probability distribution, i.e. for a given constant c

$$H_0 : \theta = c$$

    For example, is a binary classifier more accurate than random? $H_0 : p = 50\%$
  - **Two-sample test**: to test a data distribution against another data distribution, i.e.

$$H_0 : \theta_1 = \theta_2$$

    For example, is classifier A better than classifier B? $H_0 : p_A = p_B$
- We have seen one-sample test and two-sample test in the Q-Q plot lecture
- In practice, you can narrow down your choice of hypotheses by making a Q-Q plot

Choices for H_A

Given

$$H_0 : \theta = \beta$$

where β can be either a constant (one-sample test) or a parameter from another data distribution (two-sample test)

- **Alternative hypothesis H_A**: H_A can be **one-tailed** or **two-tailed**
  - **One-tailed**:

$$H_A : \quad \theta > \beta$$

or

$$H_A : \quad \theta < \beta$$

  - **Two-tailed**:

$$H_A : \quad \theta \neq \beta \iff \theta < \beta \text{ or } \theta > \beta$$

Summary: choices for H_0 and H_A

Putting everything together,

	One-sample test	Two-sample test
Two-tailed	$H_0: \theta = c,\ H_A: \theta \neq c$	$H_0: \theta_1 = \theta_2,\ H_A: \theta_1 \neq \theta_2$
One-tailed	$H_0: \theta = c,\ H_A: \theta > c$	$H_0: \theta_1 = \theta_2,\ H_A: \theta_1 > \theta_2$
	$H_0: \theta = c,\ H_A: \theta < c$	$H_0: \theta_1 = \theta_2,\ H_A: \theta_1 < \theta_2$

where θ, θ_1, θ_2 are the parameters of interest and c is a constant

Test statistic

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis
  - s is the **observation**

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis
  - s is the **observation**
  - Given a set of parameters of interest and a set of estimates, s is typically a **standardized statistic** computed from the estimates

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis
  - s is the **observation**
  - Given a set of parameters of interest and a set of estimates, s is typically a **standardized statistic** computed from the estimates
  - **Purpose:** to compare s with **a standard distribution**, e.g. the standard Gaussian distribution $\mathcal{N}(0, 1)$, to see if it is likely that the standard distribution is the underlying distribution of S, i.e. if the null hypothesis is plausible

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis
  - s is the **observation**
  - Given a set of parameters of interest and a set of estimates, s is typically a **standardized statistic** computed from the estimates
  - **Purpose:** to compare s with **a standard distribution**, e.g. the standard Gaussian distribution $\mathcal{N}(0, 1)$, to see if it is likely that the standard distribution is the underlying distribution of S, i.e. if the null hypothesis is plausible
- What is needed for computing the test statistic?
  - Assumptions on random variables X_i
  - We only need the null hypothesis H_0 (not H_A) to choose the test statistic

Test statistic

- **Test statistic s, random variable S**: the statistic used for testing the hypothesis
  - s is the **observation**
  - Given a set of parameters of interest and a set of estimates, s is typically a **standardized statistic** computed from the estimates
  - **Purpose:** to compare s with **a standard distribution**, e.g. the standard Gaussian distribution $\mathcal{N}(0, 1)$, to see if it is likely that the standard distribution is the underlying distribution of S, i.e. if the null hypothesis is plausible
- What is needed for computing the test statistic?
  - Assumptions on random variables X_i
  - We only need the null hypothesis H_0 (not H_A) to choose the test statistic

Note: in this course, we only deal with null hypothesis where we are able to express the PDF/PMF $f(s \mid H_0)$, i.e. H_0 with an equal sign in them

Test statistic (cont.)

Example 1. one-sample test

- Data: $x_1, \cdots, x_N$
- Random variable: $X_1, \cdots, X_N$ i.i.d. Gaussian with known σ
- Parameter of interest: μ_D
- Parameter estimate: $\bar{x}$
- Null hypothesis: $H_0 : \mu_D = 2.1$
- Test statistic: standardized $\bar{x}$ assuming the null hypothesis

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$
  - Recall: what are we trying to do?

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X-\mu_X}{\sigma_X}$
    - Data x: $y = \frac{x-\mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?
    - Gaussian distribution with standard deviation σ and mean $\mu_D = 2.1$

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?
    - Gaussian distribution with standard deviation σ and mean $\mu_D = 2.1$
  - **Assuming the null hypothesis**: data are assumed to be generated from the distribution described by the null hypothesis

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known** σ
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X-\mu_X}{\sigma_X}$
    - Data x: $y = \frac{x-\mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?
    - Gaussian distribution with standard deviation σ and mean $\mu_D = 2.1$
  - **Assuming the null hypothesis**: data are assumed to be generated from the distribution described by the null hypothesis - $X_i \sim \mathcal{N}(\mu_D, \sigma^2)$

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X-\mu_X}{\sigma_X}$
    - Data x: $y = \frac{x-\mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?
    - Gaussian distribution with standard deviation σ and mean $\mu_D = 2.1$
  - **Assuming the null hypothesis**: data are assumed to be generated from the distribution described by the null hypothesis - $X_i \sim \mathcal{N}(\mu_D, \sigma^2)$

Standardize $\bar{x}$ (15 sec)

Test statistic (cont.)

Example 1. one-sample test

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ i.i.d. **Gaussian with known σ**
- **Parameter of interest**: μ_D
- **Parameter estimate**: $\bar{x}$
- **Null hypothesis**: $H_0 : \mu_D = 2.1$
- **Test statistic**: **standardized** $\bar{x}$ **assuming the null hypothesis**
  - Recall: what is **standardization**?
    - Random variable X: $Y = \frac{X - \mu_X}{\sigma_X}$
    - Data x: $y = \frac{x - \mu_X}{\sigma_X}$
  - Recall: what are we trying to do? - Decide how likely data follows **the distribution described by the null hypothesis**?
  - What is the **distribution described by the null hypothesis**?
    - Gaussian distribution with standard deviation σ and mean $\mu_D = 2.1$
  - **Assuming the null hypothesis**: data are assumed to be generated from the distribution described by the null hypothesis - $X_i \sim \mathcal{N}(\mu_D, \sigma^2)$

Standardize $\bar{x}$ (15 sec)

$$z = \frac{\bar{x} - 2.1}{\sigma / \sqrt{N}}$$

Test statistic (cont.)

Example 2. two-sample test

- **Data**: $x_1, \cdots, x_{N_D}$ and $y_1, \cdots, y_{N_E}$
- **Random variable**: $X_1, \cdots, X_{N_D}$ i.i.d. **Gaussian with known σ_D**; $Y_1, \cdots, Y_{N_E}$ i.i.d. **Gaussian with known σ_E**; X_i and Y_j independent
- **Parameter of interest**: μ_D, μ_E
- **Parameter estimate**: $\bar{x}$, $\bar{y}$
- **Null hypothesis**: $H_0 : \mu_D = \mu_E \iff H_0 : \mu_D - \mu_E = 0$
- **Test statistic**: standardized $\bar{x} - \bar{y}$ assuming the null hypothesis

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_D^2/N_D + \sigma_E^2/N_E}} \text{ (explained later)}$$

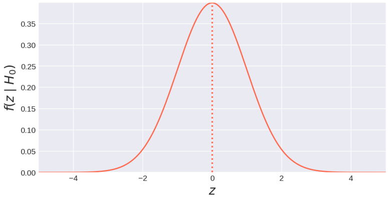
Null distribution $f(s \mid H_0)$

Null distribution

- **Null distribution** $f(s \mid H_0)$: the distribution of the test statistic given the null hypothesis
- Example:
  - **Data**: $x_1, \cdots, x_N$
  - **Random variable**: $X_1, \cdots, X_N$ i.i.d. Gaussian with known σ
  - **Parameter of interest**: μ
  - **Parameter estimate**: $\bar{x}$
  - **Null hypothesis**: $H_0 : \mu = \mu_0$
  - **Test statistic**:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$$

  - **Null distribution**: standard Gaussian distribution



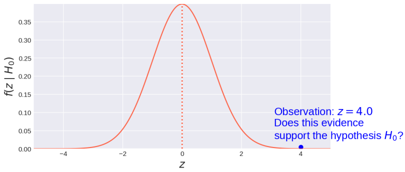
Significance level α, power and p-value

Significance level

Given a null hypothesis $H_0 : \mu = 2.1$ and the null distribution $f(s \mid H_0)$, we decide if we reject the hypothesis or not by observing data

- Run some experiments and collect data $x_1, \cdots, x_N$
- Estimate the parameter of interest $\hat{\theta}$, e.g. $\hat{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Standardize $\hat{\theta}$ assuming H_0 to compute the test statistic, e.g.

$$z = \frac{\bar{x} - 2.1}{\sigma / \sqrt{N}} = 4.0$$



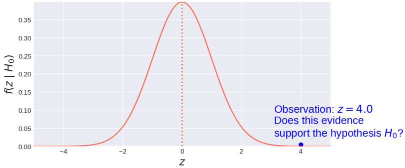
- Does this evidence support the hypothesis H_0?

Significance level

Given a null hypothesis $H_0 : \mu = 2.1$ and the null distribution $f(s \mid H_0)$, we decide if we reject the hypothesis or not by observing data

- Run some experiments and collect data $x_1, \cdots, x_N$
- Estimate the parameter of interest $\hat{\theta}$, e.g. $\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Standardize $\hat{\theta}$ assuming H_0 to compute the test statistic, e.g.

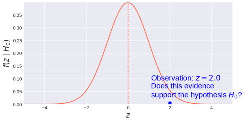
$$z = \frac{\bar{x} - 2.1}{\sigma/\sqrt{N}} = 4.0$$



- Does this evidence support the hypothesis H_0? Probably not since it's so far away from the center?

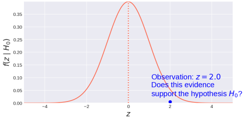
Significance level (cont.)

- What about this observation?



Significance level (cont.)

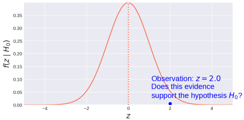
- What about this observation?



- To be able to answer the question, you need to decide where you draw the line

Significance level (cont.)

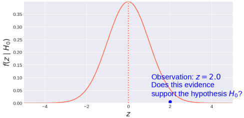
- What about this observation?



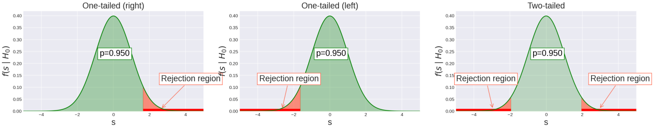
- To be able to answer the question, you need to decide where you draw the line - define a **rejection region** by choosing a significance level

Significance level (cont.)

- What about this observation?



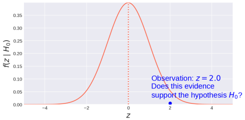
- To be able to answer the question, you need to decide where you draw the line - define a **rejection region** by choosing a significance level
- **Significance level** α: red area under the curve



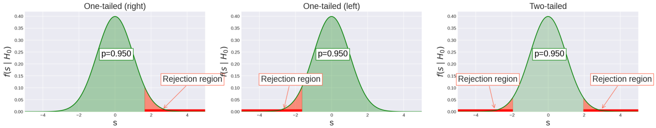
In these three images, $\alpha = 0.05$

Significance level (cont.)

- What about this observation?



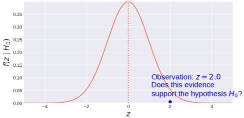
- To be able to answer the question, you need to decide where you draw the line - define a **rejection region** by choosing a significance level
- **Significance level** α: red area under the curve



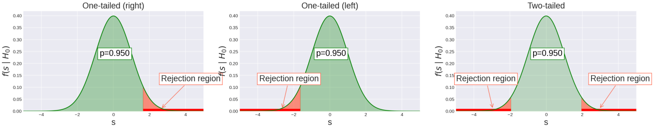
In these three images, $\alpha = 0.05$
More conservative $\Rightarrow$ less probable to reject H_0, which indicates a smaller rejection region

Significance level (cont.)

- What about this observation?



- To be able to answer the question, you need to decide where you draw the line - define a **rejection region** by choosing a significance level
- **Significance level α**: red area under the curve



In these three images, $\alpha = 0.05$
More conservative $\Rightarrow$ less probable to reject H_0, which indicates a smaller rejection region
Two-tailed H_A is more conservative

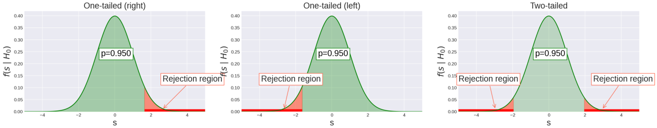
Significance level (cont.)

What is needed for choosing a meaningful α?

- Null distribution
- H_A one-tailed or two-tailed

Interpretation of α

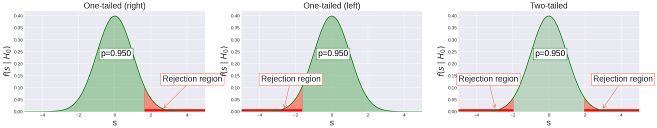
- $\alpha = P(\textit{reject } H_0 \mid H_0 \textit{ is true})$ - the probability of making such a mistake



  - The rejection region indicates that H_0 is **unlikely**, but the probability is not zero
  - It is possible that H_0 is true, but our observation happens to fall in the rejection region
  - If H_0 is true and our observation falls in the rejection region, we will **mistakenly** reject H_0
  - The probability of making this type of mistakes is α
- Similar to the confidence interval, $1 - \alpha$ is called the **confidence level**

Interpretation of α

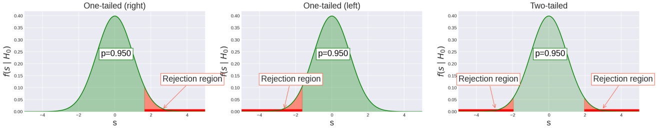
- $\alpha = P(\textit{reject } H_0 \mid H_0 \textit{ is true})$ - the probability of making such a mistake



  - The rejection region indicates that H_0 is **unlikely**, but the probability is not zero
  - It is possible that H_0 is true, but our observation happens to fall in the rejection region
  - If H_0 is true and our observation falls in the rejection region, we will **mistakenly** reject H_0
  - The probability of making this type of mistakes is α
- Similar to the confidence interval, $1 - \alpha$ is called the **confidence level** - "with 95% confidence, rejecting H_0 is the right thing to do"

Interpretation of α

- $\alpha = P(\textit{reject } H_0 \mid H_0 \textit{ is true})$ - the probability of making such a mistake



  - The rejection region indicates that H_0 is **unlikely**, but the probability is not zero
  - It is possible that H_0 is true, but our observation happens to fall in the rejection region
  - If H_0 is true and our observation falls in the rejection region, we will **mistakenly** reject H_0
  - The probability of making this type of mistakes is α
- Similar to the confidence interval, $1 - \alpha$ is called the **confidence level** - "with 95% confidence, rejecting H_0 is the right thing to do"
- Define the significance level **before you run the experiments** so that you can't cheat!

Significance level and power

- Contingency table:

	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

Significance level and power

- Contingency table:

	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

- **Significance level** α: incorrectly rejecting H_0

$$\alpha = P(\text{type I error})$$

Significance level and power

- Contingency table:

	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

- **Significance level** α: incorrectly rejecting H_0

$$\alpha = P(\text{type I error})$$

- **Power**: correctly rejecting H_0

$$\text{power} = P(\text{reject } H_0 \mid H_A) = 1 - P(\text{type II error})$$

Significance level and power

- Contingency table:

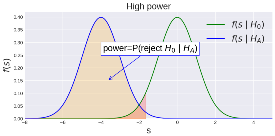
	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

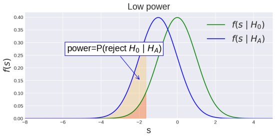
- **Significance level** α: incorrectly rejecting H_0

$$\alpha = P(\text{type I error})$$

- **Power**: correctly rejecting H_0

$$\text{power} = P(\text{reject } H_0 \mid H_A) = 1 - P(\text{type II error})$$





Significance level and power

- Contingency table:

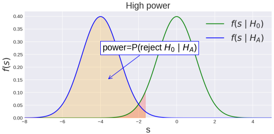
	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

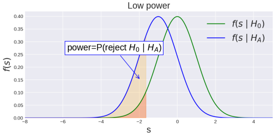
- **Significance level** α: incorrectly rejecting H_0

$$\alpha = P(\text{type I error})$$

- **Power**: correctly rejecting H_0

$$\text{power} = P(\text{reject } H_0 \mid H_A) = 1 - P(\text{type II error})$$





- What is needed for computing the power?

Significance level and power

- Contingency table:

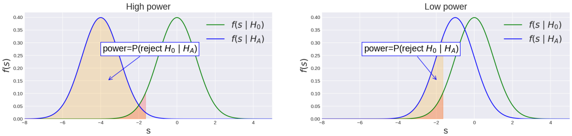
	$y = H_A$	$y = H_0$
$\hat{y}$ = reject H_0	**TP**	**FP** (Type I error)
$\hat{y}$ = do not reject H_0	**FN** (Type II error)	**TN**

- **Significance level** α: incorrectly rejecting H_0

$$\alpha = P(\text{type I error})$$

- **Power**: correctly rejecting H_0

$$\text{power} = P(\text{reject } H_0 \mid H_A) = 1 - P(\text{type II error})$$



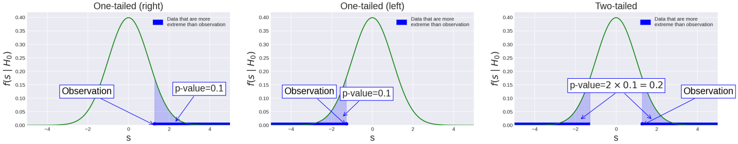
- What is needed for computing the power? $f(s \mid H_0)$, $f(s \mid H_A)$

p-value

- ***p*-value**:
  - One-tailed:
    - Right tail: $p = P(S \geq s \mid H_0)$, e.g. 1-stats.norm.cdf(s, 0, 1)
    - Left tail: $p = P(S \leq s \mid H_0)$, e.g. stats.norm.cdf(s, 0, 1)
  - Two-tailed:
    - $p = 2\min\left(P(S \leq s \mid H_0), P(S \geq s \mid H_0)\right)$, e.g. 2*min(stats.norm.cdf(s, 0, 1), 1-stats.norm.cdf(s, 0, 1))

    Note: for example, if $f(s \mid H_0)$ is symmetric around zero and $s < 0$,

$$p = 2P(S \leq s \mid H_0)$$



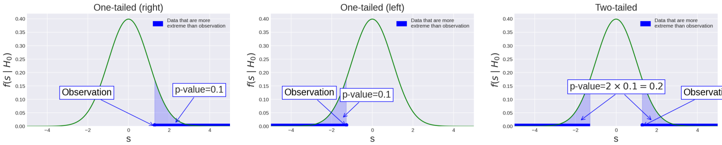
- What is needed for computing the p-value? (10 sec)

p-value

- ***p*-value**:
  - One-tailed:
    - Right tail: $p = P(S \geq s \mid H_0)$, e.g. 1-stats.norm.cdf(s, 0, 1)
    - Left tail: $p = P(S \leq s \mid H_0)$, e.g. stats.norm.cdf(s, 0, 1)
  - Two-tailed:
    - $p = 2\min\left(P(S \leq s \mid H_0), P(S \geq s \mid H_0)\right)$, e.g. 2*min(stats.norm.cdf(s, 0, 1), 1-stats.norm.cdf(s, 0, 1))

  Note: for example, if $f(s \mid H_0)$ is symmetric around zero and $s < 0$,

$$p = 2P(S \leq s \mid H_0)$$



- What is needed for computing the p-value? (10 sec)
  - Null distribution
  - Alternative hypothesis H_A to know one-tailed or two-tailed
  - Observation - test statistic computed from data

Summary: steps for hypothesis testing

- Step 1 Make a "boring" statement
- Step 2 Design an **experiment**
- Step 3 Describe the **data** generated from the experiment and the corresponding random variables
- Step 4 Describe the parameter of interest and their estimates
- Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis** H_0
- Step 6 Find the expression for the **test statistic** s
- Step 7 Find the expression for the **null distribution**
- Step 8 Define **an alternative hypothesis** H_A: one-tailed or two-tailed
- Step 9 Choose a **significance level** α (the tail), which defines the **rejection region**
- Step 10 Collect **data**
- Step 11 Compute the test statistic from data
- Step 12 Compute the p-value
- Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0; otherwise, we fail to reject H_0.

Today

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables**

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a “boring” statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss** $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ **i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

- Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**
- Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**
- Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known
- Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**
- Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis** H_0 (10 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss** $x_1, \cdots, x_{30}$**;** $X_1, \cdots, X_{30}$ **i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value** μ_D **and** $\hat{\mu}_D = \bar{x}$

Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis** H_0 (10 secs): $H_0 : \mu_D = 2.1$

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**

Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis H_0** (10 secs): $H_0 : \mu_D = 2.1$

Step 6 Find the expression for the **test statistic s** (60 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**

Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis H_0** (10 secs): $H_0 : \mu_D = 2.1$

Step 6 Find the expression for the **test statistic s** (60 secs):

$$s = z = \frac{\bar{x} - 2.1}{\sigma/\sqrt{30}}$$

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**

Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis** H_0 (10 secs): $H_0 : \mu_D = 2.1$

Step 6 Find the expression for the **test statistic** s (60 secs):

$$s = z = \frac{\bar{x} - 2.1}{\sigma / \sqrt{30}}$$

Step 7 Find the expression for the **null distribution** $f(s \mid H_0)$ (10 secs):

Example

Recall example: if you control the diet of your ducks, they lose 2.1 kg after one month on average. Company A has developed a drug D to help ducks lose weight. They claim that on average the drug works better than diet control. Here is the set up for the experiment.

Step 1 Make a "boring" statement (5 secs): **drug D works the same as diet**

Step 2 Design an **experiment** (choose $N = 30$) (10 secs): **let 30 chonker ducks take drug D and measure their weight loss after one month**

Step 3 Describe the **data** and **random variables** with assumptions about their distributions (5 secs): **weight loss $x_1, \cdots, x_{30}$; $X_1, \cdots, X_{30}$ i.i.d. Gaussian random variables** - let's make an additional assumption to simplify the problem - the standard deviation of X_i $\sigma = 0.6$ is known

Step 4 Describe the parameter of interest and their estimates (10 secs): **the mean value μ_D and $\hat{\mu}_D = \bar{x}$**

Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis H_0** (10 secs): $H_0 : \mu_D = 2.1$

Step 6 Find the expression for the **test statistic s** (60 secs):

$$s = z = \frac{\bar{x} - 2.1}{\sigma/\sqrt{30}}$$

Step 7 Find the expression for the **null distribution $f(s \mid H_0)$** (10 secs):

$$f(z \mid H_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Example (cont.)

Step 8 Define **an alternative hypothesis H_A** (10 secs):

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$H_A : \mu_D \neq 2.1$ **or** $H_A : \mu_D > 2.1$

One-tailed or two-tailed

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):**

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \text{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):**

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs):

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs): e.g. $\alpha = 0.05$

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs): e.g. $\alpha = 0.05$

Step 10 Collect 30 ducks in 20 secs and feed them drugs

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs): e.g. $\alpha = 0.05$

Step 10 Collect 30 ducks in 20 secs and feed them drugs - great job! Weights measured after one month $x_1, \cdots, x_{30}$

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs): e.g. $\alpha = 0.05$

Step 10 Collect 30 ducks in 20 secs and feed them drugs - great job! Weights measured after one month $x_1, \cdots, x_{30}$

Say $\frac{1}{30}\sum_{i=1}^{30} x_i = 2.2$

Step 11 Compute the test statistic from data (5 secs):

Example (cont.)

Step 8 Define **an alternative hypothesis** H_A (10 secs):

$$H_A : \mu_D \neq 2.1 \textbf{ or } H_A : \mu_D > 2.1$$

One-tailed or two-tailed

- **Two-tailed (5 secs):** $H_A : \mu_D \neq 2.1$
- **One-tailed (5 secs):** $H_A : \mu_D > 2.1$

Step 9 Choose a **significance level** α (the tail), which defines the rejection region (5 secs): e.g. $\alpha = 0.05$

Step 10 Collect 30 ducks in 20 secs and feed them drugs - great job! Weights measured after one month $x_1, \cdots, x_{30}$

Say $\frac{1}{30}\sum_{i=1}^{30} x_i = 2.2$

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.2 - 2.1}{0.6/\sqrt{30}} = 0.91$$

Example (cont.)

Step 12 Compute the p-value (20 secs):

Example (cont.)

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.1807 > \alpha$

Example (cont.)

Step 12 Compute the *p*-value (20 secs):

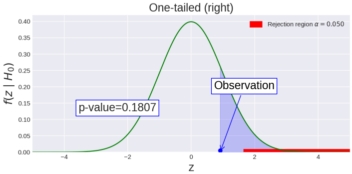
- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.1807 > \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.3613 > \alpha$

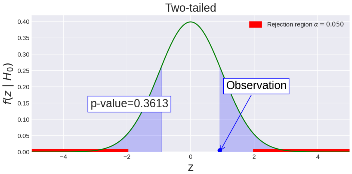
Example (cont.)

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.1807 > \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.3613 > \alpha$

Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0





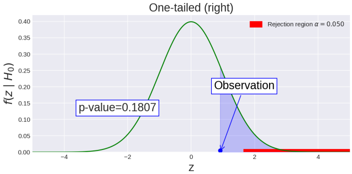
Do not reject H_0 for both one-tailed and two-tailed H_A

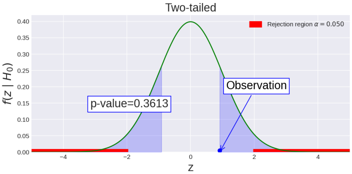
Example (cont.)

Step 12 Compute the p-value (20 secs):

- **For $H_A : \mu_D > 2.1$ (one-tailed): $p = P(Z \geq z_0 \mid H_0) = 0.1807 > \alpha$**
- **For $H_A : \mu_D \neq 2.1$ (two-tailed): $p = 2P(Z \geq z_0 \mid H_0) = 0.3613 > \alpha$**

Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0





Do not reject H_0 for both one-tailed and two-tailed H_A

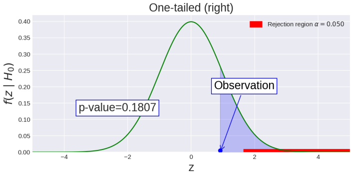
What does it mean?

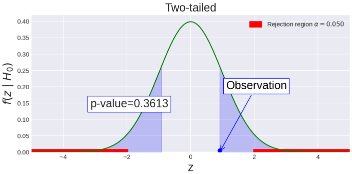
Example (cont.)

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.1807 > \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.3613 > \alpha$

Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0





Do not reject H_0 for both one-tailed and two-tailed H_A

What does it mean? - Based on this test, you will stick to diet control instead of buying drug D.

Example (cont.)

What if $\bar{x} = 2.3$?

Example (cont.)

What if $\bar{x} = 2.3$?

Step 11 Compute the test statistic from data (5 secs):

Example (cont.)

What if $\bar{x} = 2.3$?

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{30}} = 1.826$$

Example (cont.)

What if $\bar{x} = 2.3$?

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{30}} = 1.826$$

Step 12 Compute the p-value (20 secs):

Example (cont.)

What if $\bar{x} = 2.3$?

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{30}} = 1.826$$

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.0339 < \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.0679 > \alpha$

Example (cont.)

What if $\bar{x} = 2.3$?

Step 11 Compute the test statistic from data (5 secs):

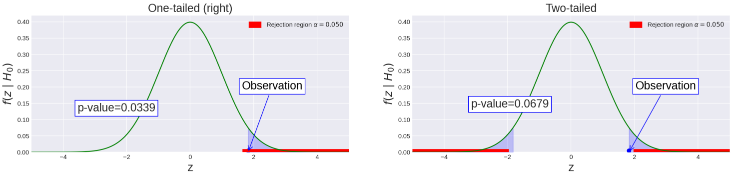
$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{30}} = 1.826$$

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.0339 < \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.0679 > \alpha$

Example (cont.)

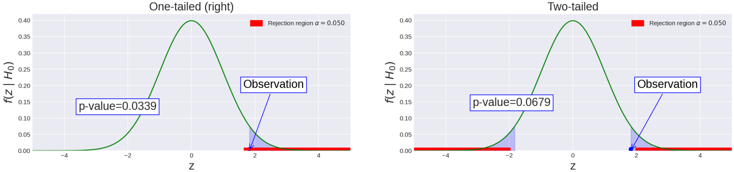
Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0



Reject H_0 for one-tailed H_A; do not reject H_0 for two-tailed H_A for the same confidence level $1 - \alpha = 95\%$

Example (cont.)

Step 13 If *p*-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0



Reject H_0 for one-tailed H_A; do not reject H_0 for two-tailed H_A for the same confidence level $1 - \alpha = 95\%$

Note: the two-tailed test is more conservative - if the data passes a two-tailed test, it is more conclusive than one-tailed test for the same confidence level

Example (cont.)

What if $\bar{x} = 2.3$ with $N = 100$?

Example (cont.)

What if $\bar{x} = 2.3$ with $N = 100$?

Step 11 Compute the test statistic from data (5 secs):

Example (cont.)

What if $\bar{x} = 2.3$ with $N = 100$?

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{100}} = 3.33$$

Example (cont.)

What if $\bar{x} = 2.3$ with $N = 100$?

Step 11 Compute the test statistic from data (5 secs):

$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{100}} = 3.33$$

Step 12 Compute the p-value (20 secs):

Example (cont.)

What if $\bar{x} = 2.3$ with $N = 100$?

Step 11 Compute the test statistic from data (5 secs):

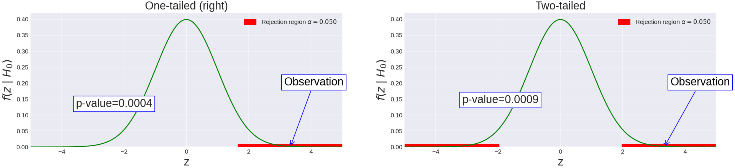
$$z_0 = \frac{2.3 - 2.1}{0.6/\sqrt{100}} = 3.33$$

Step 12 Compute the p-value (20 secs):

- **For** $H_A : \mu_D > 2.1$ **(one-tailed):** $p = P(Z \geq z_0 \mid H_0) = 0.0004 < \alpha$
- **For** $H_A : \mu_D \neq 2.1$ **(two-tailed):** $p = 2P(Z \geq z_0 \mid H_0) = 0.0009 < \alpha$

Example (cont.)

Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0



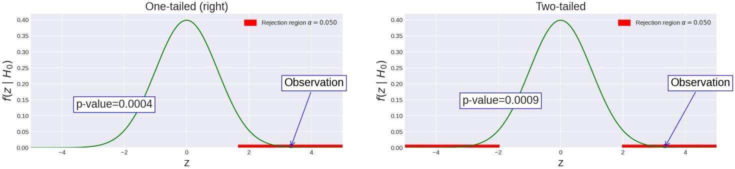
Reject H_0 for both one-tailed and two-tailed H_A

Note:

- With more data, it becomes more certain that we should reject H_0 in favor of H_A given the observation $\bar{x} = 2.3$

Example (cont.)

Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0



Reject H_0 for both one-tailed and two-tailed H_A

Note:

- With more data, it becomes more certain that we should reject H_0 in favor of H_A given the observation $\bar{x} = 2.3$

This test is called one-sample z-test

Today

Recall: one-tailed vs two-tailed tests

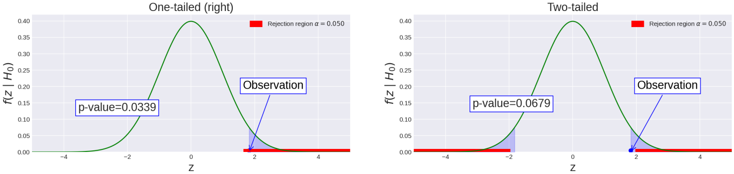
- p-value indicates how "surprising" the observation is

Recall: one-tailed vs two-tailed tests

- p-value indicates how "surprising" the observation is
- "Surprising" observation usually means potential novelty

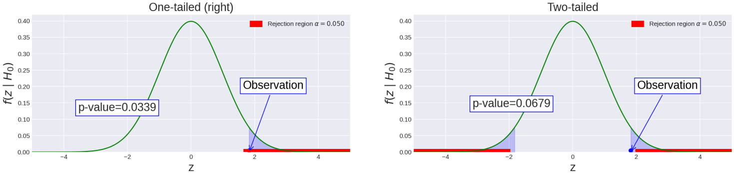
Recall: one-tailed vs two-tailed tests

- p-value indicates how "surprising" the observation is
- "Surprising" observation usually means potential novelty
- In one of the examples, we have shown that we reject the null hypothesis for the one-tailed test and we fail to reject the null hypothesis for the two-tailed test given the same significance level



Recall: one-tailed vs two-tailed tests

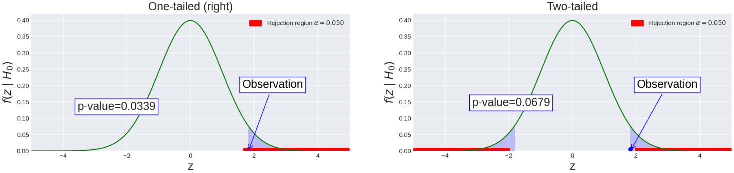
- *p*-value indicates how "surprising" the observation is
- "Surprising" observation usually means potential novelty
- In one of the examples, we have shown that we reject the null hypothesis for the one-tailed test and we fail to reject the null hypothesis for the two-tailed test given the same significance level



- In this example, if we use the two-tailed test, we will not claim that we have observed potential novelty with the experiment, whereas if we use the one-tailed test, we claim that we do observe potential novelty

Recall: one-tailed vs two-tailed tests

- *p*-value indicates how "surprising" the observation is
- "Surprising" observation usually means potential novelty
- In one of the examples, we have shown that we reject the null hypothesis for the one-tailed test and we fail to reject the null hypothesis for the two-tailed test given the same significance level



- In this example, if we use the two-tailed test, we will not claim that we have observed potential novelty with the experiment, whereas if we use the one-tailed test, we claim that we do observe potential novelty
- The conclusion we draw depends on which test we conduct

Variation of the p-value

- p-value is computed from data

Variation of the p-value

- p-value is computed from data
- Data is random - **p-value is random**

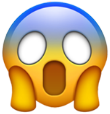
Variation of the p-value

- p-value is computed from data
- Data is random - **p-value is random**
- With the same experiment set up, if we switch to a different sample, p-value will be different

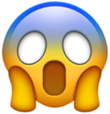
p-hacking

- Many factors can result in a different p-value
- p-hacking refers to situations where researchers are **trying multiple things until they get the desired result**
- This action can be a conscious decision, a subconscious decision or even an unconscious action
- p-hacking can be tricky to identify
- Suggestions to avoid p-hacking, e.g. one should always **report effect sizes and confidence intervals**
- Reference:
  - `https://www.nature.com/news/scientific-method-statistical-errors-1.14700`
  - Why Most Published Research Findings Are False?

p-hacking (cont.)

What should I do!? 

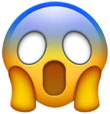
p-hacking (cont.)



What should I do!?

- Be honest and explicit about your assumptions

p-hacking (cont.)



What should I do!?

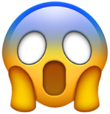
- Be honest and explicit about your assumptions
- Be "conservative"

p-hacking (cont.)

What should I do!? 

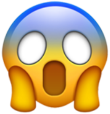
- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!

p-hacking (cont.)

What should I do!? 

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true**

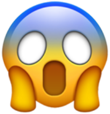
p-hacking (cont.)



What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true** - **try to prove yourself wrong**

p-hacking (cont.)

What should I do!? 

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true** - **try to prove yourself wrong** - be a proper scientist

Today

1. Terminology
2. Example
3. p-hacking
4. Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis test

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis test

Next:

- More examples, test statistics; comparison of two classifiers

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis test

Next:

- More examples, test statistics; comparison of two classifiers

Before next lecture:

- Steps for hypothesis testing