Lecture 9: Hypothesis testing part I

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

December 01 and 05, 2022

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

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

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

Important 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 (aka. Duckyphanomin) to help duckies lose weight. They claim that **on average** the drug works better than diet control
- Company B has developed a drug E (aka. Everyduckyslim) 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 should you buy? Or should you just control their diet without drugs?

- 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?

What would you do?

Hypothesis

- **Hypothesis:**
  - A proposed explanation for a phenomenon (Wikipedia)
  - An idea or explanation of something that is based on a few known facts but that has not yet been proved to be true or correct (Oxford dictionary)
- **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 proposed distribution
  - The test is based on sample statistics, which are computed from data
  - Hypothesis + data $\rightarrow$ decision on rejecting or not rejecting the hypothesis

Hypothesis testing: a list to go through

- A default 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 to propose a **default statement**: a "boring" and unsurprising claim that we would like to **test**, e.g.,
  - Drug D is **not more effective** than regular diet on average
  - Drug E works the same as drug D on average

  In science, we are hoping for new discoveries and excitement, but we need to earn it by showing that the trivial explanation does not hold
- How do we test the default statement? We need to 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 default statement**: drug D is not more effective than regular diet on average What experiments can we run to test if this statement is true?
  - **Experiment** (5 sec): give drug D to 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 mean of the weight loss μ_D
  - **Parameter estimate** (5 sec): the sample mean $\hat{\mu_D} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

  Then we test if $\bar{x}$ is greater than diet control (2.1 kg)

Experiment design (cont.)

- Example 2:
  - **A default statement**: drug E and drug D work the same on average
  - **Experiment** (5 sec): give drug D to 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 mean μ_D and μ_E for drug D and E, respectively
  - **Parameter estimate** (5 secs): the sample mean $\hat{\mu}_D = \bar{x} = \frac{1}{N_D}\sum_{i=1}^{N_D} x_i$ and $\hat{\mu_E} = \bar{y} = \frac{1}{N_E}\sum_{j=1}^{N_E} y_j$

  Then we test if $\bar{x}$ and $\bar{y}$ are the same

Experiment design (cont.)

- We make our decision by observing data; if the evidence does not support the default 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
- Intuition: "If the statement is true, then the evidence should support the statement", which is the same as ($\iff$) "if the evidence does not support the statement, the statement is considered false" , which is not the same as ($\not\Leftrightarrow$) "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** - a statement about the **parameter of interest**
- **Null hypothesis** H_0: the default 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: an alternative hypothesis that is **complementary** (the opposite) to the null hypothesis
  - Example 2 (5 sec): drug E and drug D do not work the same on average (5 sec)

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

  - Example 1 (5 sec): drug D is more effective than regular diet on average (5 sec)

$$H_A : \mu_D > 2.1$$

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: One implicit assumption here is that μ_D will not be smaller than 2.1
  Question 1.1: Do I need to make this assumption?
  Answer: No
  Question 1.2: Could you elaborate on that?
  Answer: Yes
  Question 1.3: When?
  Answer: In a few slides
  Okay
- Question 2: Can H_0 and H_A be ANYTHING I want? Like a magic mirror!?
  Answer: No
  Question 2.2: What are the choices for H_0 and H_A then?

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 this (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 looking at Q-Q plots

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

Note: this is the answer to question 1.1 (cf. page 14): if you choose the one-tailed test, then you are making the assumption $H_A : \mu_D > 2.1$; if you choose the two-tailed test, then you are not making this assumption

Test statistic

Test statistic

- **Test statistic** s (random variable S): a (typically standardized) statistic computed from data
- **Purpose:**
  - Assume the null hypothesis is true, we can calculate the sampling distribution of S
  - Then we observe s; s will indicate how plausible this sampling distribution
- 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

  Disclaimer: 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 (is drug D more effective than diet control)

- **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: $y = \frac{x - \mu_X}{\sigma_X}$
  - What are we trying to do here? - To test if we can reject the null hypothesis by asking does data follow the **distribution described by the null hypothesis**?
  - Why are we standardizing the statistic $\bar{x}$? We want to use standard tools for our analysis
  - 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 - $X_i \sim \mathcal{N}(2.1, \sigma^2)$

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

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

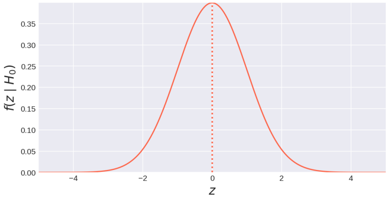
Null distribution $f(s \mid H_0)$

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