Experimental Methods in Computer Science

(*Metodologias Experimentais em Informática*)

Henrique Madeira

Master in Informatics Engineering
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia da Universidade de Coimbra
2023/2024

1

Hypothesis Testing

Hypothesis testing slides are mainly based on chapter 8 of the book "Essentials of Social Statistics for a Diverse Society" Second Edition by Anna Leon-Guerrero, Chava Frankfort-Nachmias , SAGE Publications, Inc, 2010.

2

Hypothesis testing scenario 1 (test for a mean)

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

Package exec. time
74
66
88
68
⋮
87
79
78
72
86
85
86

32 times

Avg = 78.15

Henrique Madeira, DEI-FCTUC, 2018-2023

3

Hypothesis testing scenario 2 (test for means)

Assume you are the database administrator of a big information system. The database has just been installed and you are trying two tuning configurations: Conf. **A** and Conf. **B**.

You use a given SQL package to test the execution time for each configuration.

After running several times the SQL package in both configurations you want to take a decision.

Question: **what is the best configuration?**

Conf. A exec. time	Conf. B exec. time
74	69
66	71
88	80
68	88
79	64
68	65
87	74
79	76
78	89
72	68
86	67
85	72
86	

Avg A = 78.15 **Avg B = 73.58**
n = 13 **n = 12**

Henrique Madeira, DEI-FCTUC, 2018-2023

4

Hypothesis

- **What is an hypothesis?**
  - A proposed explanation for a given phenomenon
  - An assumption about the efficiency of a given component/system
  - A statement about the parameters of a population (**statistical view**)
- Scope
  - **Abstract**: about the world *(lato senso)*
  - **Concrete**: about a given design or apparatus
- Types
  - **Explanatory**: explains the phenomenon, identifies relations and/or causality between variable/elements of the phenomenon
  - **Predictive**: predicts the observation of a phenomenon, anticipates the outcome of an experiment,...

An hypothesis is a tentative answer!

Henrique Madeira, DEI-FCTUC, 2018-2023

5

Hypothesis

- **What is an hypothesis?**
  - A proposed explanation for a given phenomenon
  - An assumption about the efficiency of a given component/system
  - A statement about the parameters of a population (**statistical view**)
- Scope
  - **Abstract**: abo
  - **Concrete**: abo
- Types
  - **Explanatory**:
    between varia
  - **Predictive**: pr
    an experiment

- An **hypothesis requires evaluation** to be considered true. It can be **rejected** or, in the absence of rejection, it is **confirmed.**
- True hypothesis means the probability of it being correct is 'high' and the probability of it being incorrect is 'low'.
- Statistics is necessary to quantify the meaning of "high" and "low" and to decide about the validity of the hypothesis.
- Hypotheses are rejected or accepted with some **degree of certainty**

Henrique Madeira, DEI-FCTUC, 2018-2023

6

Hypothesis: put it into perspective

Topic, problem and hypothesis

- **Topic:** Subject (focused area) of interest, where the gap or difficulty to be solved is included. Essential to provide context to the hypothesis.
- **Problem:** Object of the study. Presumes clear and explicit **questions** that formulate the problem to be solved.
- **Hypothesis:** Provisional answer to the question(s). If the hypotheses is confirmed, the answer is considered correct (to a given degree of certainty).

Henrique Madeira, DEI-FCTUC, 2018-2023

7

Hypothesis: put it into perspective

Topic, problem and hypothesis

- **Topic:** Subject (focused area) of interest, where the gap or difficulty to be solved is included. Essential to provide context to the hypothesis.
- **Problem:** Object of the s… …ns that formulate the problem to…
- **Hypothesis:** Provisional… …es is confirmed, the answer is considered correct (to a given degree of certainty).

Example:
Quality of the code (absence of bugs) produced by programmers.

Henrique Madeira, DEI-FCTUC, 2018-2023

8

Hypothesis: put it into perspective

Topic, problem and hypothesis

- **Topic:** Subject (focused area) of
solved is included. Esse
- **Problem:** Object of the study. Presumes clear and explici **questions** that formulate the problem to be solved.
- **Hypothesis:** Provisional answer to the question(s). If the hypotheses is confirmed, the answer is considered correct (to a given degree of certainty).

Example:
Is the software development methodology related to the number of bugs in deployed software?

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 9

9

Hypothesis: put it into perspective

Topic, problem and hypothesis

- **Topic:** Subject (focused ar
solved is included. Essentia
- **Problem:** Object of the
formulate the prob
- **Hypothesis:** Provisional answer to the question(s). If the hypotheses is confirmed, the answer is considered correct (to a given degree of certainty).

Example:
H_0 - Software developed and deployed using CMMi 5 has the same bug density of software developed using Scrum.

H_1 - Software developed and deployed using CMMi 5 has **not** the same bug density of software developed using Scrum.

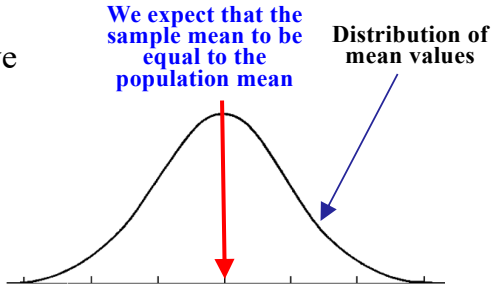
Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 10

10

Inferential statistics and hypothesis testing

- Allows us to evaluate the behavior in samples to learn more about the behavior in the entire population
- Quite often, the entire population is too large (or even infinite) or is not accessible
- From the **central limit theorem**, we know that the probability of selecting any other sample mean value from this population is normally distributed.



Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 23

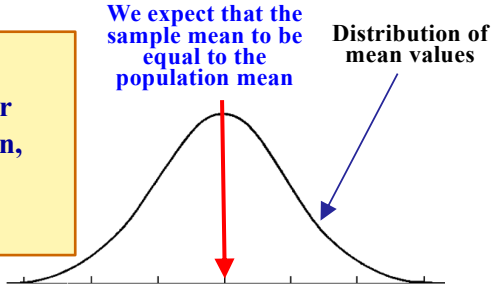
23

Inferential statistics and hypothesis testing

- Allows us to evaluate the behavior in samples to learn more about the behavior in the entire population
- Quite often, the entire population is too large (or even infinite) or is not accessible
- Fror
  kno
  sele
  valu
  norr

Hypothesis testing

A systematic way to test claims or ideas about a group or population, based on selected samples of such population.



Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 24

24

Hypothesis testing steps

1. State the hypothesis or claim to be tested
2. Select the criteria for a decision
3. Compute the test statistic
4. Make a decision

Laboratory experiments, controlled experiments

1. Problem statement (or research question)
2. Identify variables
3. Generate hypothesis
4. Define the experimental setup/scenario
5. Develop tools and procedures for the experiment

Design of the experiment

6. Run experiments and collect the data/measurements — Measurements
7. Perform data analysis — Analysis
8. Draw conclusions (often go back to the beginning and reformulate the problem statement or test a different hypothesis) — Conclusions

Henrique Madeira, DEI-FCTUC, 2018-2023

25

Hypothesis testing scenario 1 (test for a mean)

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

We consider the data distribution normal because:

- Each execution is independent from previous executions;
- The variability in the measurements results from random changes in the execution conditions.

If we are not sure that the data follows a normal distribution, we must test it for normality.

Package exec. time
74
66
88
68
⋮
87
79
78
72
86
85
86

32 times

Avg = 78.15

Henrique Madeira, DEI-FCTUC, 2018-2023

26

Step 1 - State the hypothesis

- **Null hypothesis** (H_0) is a statement about the population parameter (e.g., the population mean) that is assumed to be true.
  This is a provisional answer to the research question or problem under study. For example:
  H_0 - The new configuration has no effect on the execution time of the SQL packaged

- **Alternative hypothesis** (H_1) is a statement that directly contradicts the null hypothesis by stating that the actual value of the population is not equal to the value stated in the null hypothesis.
  This is what we think is wrong about the null hypothesis. For example:
  H_1 – The execution time of the SQL packaged is different in the new configuration (could be smaller or bigger)

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 27

27

Step 1 - State the hypothesis

- **Null hypothesis** (H_0) is a statement about the population parameter (e.g., the population mean) that is assumed to be true.
  This is a provisional answer to the research question or problem under study. For example:
  ed

- alue

The decision made in hypothesis testing centers on the null hypothesis H_0

- The idea is to show evidences that H_0 is unlikely, in order to reject the null hypothesis. If failing to do so, the null hypothesis is retained.
- The bias is do nothing. In other words, the burden is put on the researcher to demonstrate that H_0 is not likely to be true. → **The experiments must be defined to collect data to show that H_0 is not true**

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 28

28

Step 2 - Select the criteria for a decision

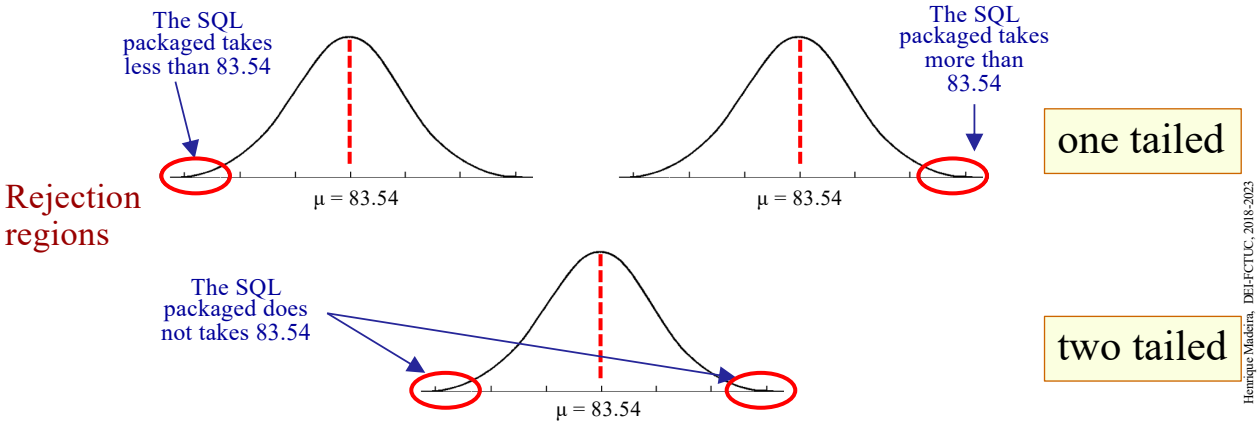
- To set a criteria means **to state the significance level for the test.**
- **Significance level** refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis.
- A typical significance level is 5%. This means that when the probability of obtaining a given sample mean is less than 5%, supposing that the null hypothesis is true, then we conclude that the sample used to calculate the mean is too unlikely, and so we **reject the null hypothesis**.

Henrique Madeira, DEI-FCTUC, 2018-2023

29

Step 2 - Select the criteria for a decision

The alternate hypothesis H_1 establish where to place the level of confidence
For example:

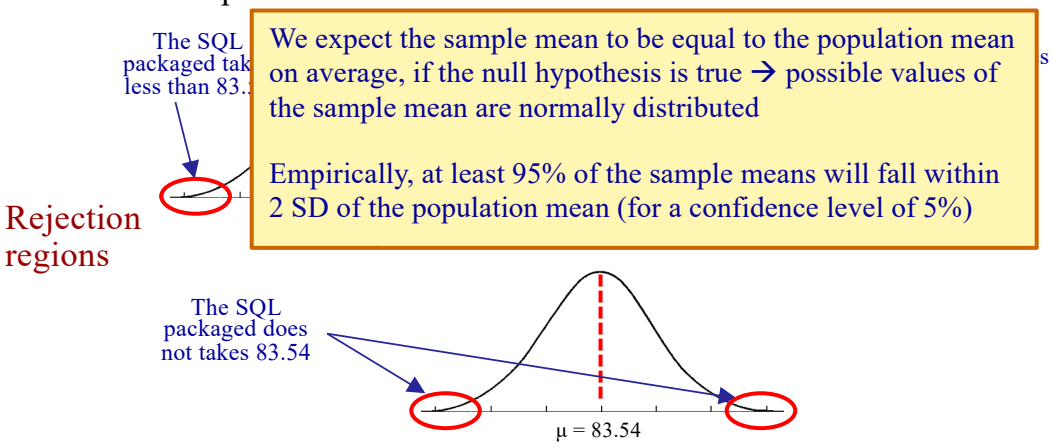


Henrique Madeira, DEI-FCTUC, 2018-2023

30

Step 2 - Select the criteria for a decision

The alternate hypothesis H_1 establish where to place the level of confidence
For example:



We expect the sample mean to be equal to the population mean on average, if the null hypothesis is true → possible values of the sample mean are normally distributed

Empirically, at least 95% of the sample means will fall within 2 SD of the population mean (for a confidence level of 5%)

one tail

two tail

Henrique Madeira, DEI-FCTUC, 2018-2023

31

Step 3 – Compute the test statistics

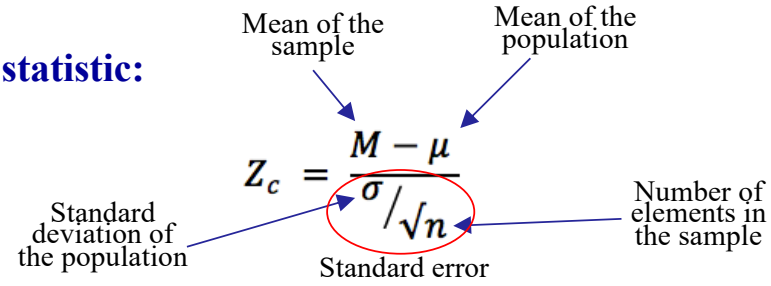
- Select a random sample from the population and measure the sample mean. For example: execute the SQL package n times and measure a mean = 78.15
- To make a decision we need to evaluate how likely this sample outcome is, if the population mean stated by the null hypothesis (83.54) is true.
- **Test statistic** is a formula to determine the likelihood of obtaining sample outcomes if the null hypothesis is true. The value of the test statistic is used to make a decision regarding the null hypothesis.

Henrique Madeira, DEI-FCTUC, 2018-2023

32

Step 3 – Compute the test statistics
(test for means, normal distribution)

Test statistic:



Measures how far the sample mean is from the population mean under H_0. The larger the value of $|Z_c|$ the more it will indicate that H_0 is not true.

Henrique Madeira, DEI-FCTUC, 2018-2023

33

Step 4 – Make a decision

- The value of the test statistic (Z_c) is the key to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true.
- ***P* value** is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true.
- Example:
  - $P < 5\%$ → reject the null hypothesis (reach significance)
  - $P > 5\%$ → retain the null hypothesis (fail reaching significance)

Henrique Madeira, DEI-FCTUC, 2018-2023

34

Hypothesis testing scenario 1 (test for a mean)

Assume you are the database administrator of a big information system and you are unhappy with the execution time of a given SQL package.

From historical data (thousands of previous package executions), you know that the average execution time of the package is **83.54** seconds with a standard deviation of **16.36**.

You change the tuning of the database and run the package several times to check the effect.

Questions:

- **Has the new tuning any effect?**
- **Is the new configuration better?**
- **Is the new configuration worse?**

Package exec. time
74
66
88
68
⋮
87
79
78
72
86
85
86

32 times

Avg = 78.15

Henrique Madeira, DEI-FCTUC, 2018-2023

35

Example 1: non-directional (two tailed)
Step 1- State the hypothesis

- $\mathbf{H_0}$ – The new configuration has no effect on the execution time of the SQL packaged.
- $\mathbf{H_1}$ – The execution time of the SQL packaged is different in the new configuration (could be smaller or bigger)

We are testing whether the null hypothesis H_0 is true

Henrique Madeira, DEI-FCTUC, 2018-2023

36

Example 1: non-directional (two tailed)
Step 2 - Set the criteria for a decision

- Consider the level of significance of 5% → $\alpha = 0.05$. → $1 - \alpha = 0.95$
- Locate the Z score (in the table for the standard normal distribution) that represents the **critical values**
- A **critical value** is a cutoff value that sets the boundaries beyond which less than 5% of sample means can be obtained if the null hypothesis is true.

37

Example 1: non-directional (two tailed)
Step 2 - Set the criteria for a decision

- Consider the level of significance of 5% → $\alpha = 0.05$. → $1 - \alpha = 0.95$
- Locate the Z score (in the table for the standard n the **critical values**
- A **critical value** is a cutoff value that sets the bou 5% of sample means can be obtained if the null h

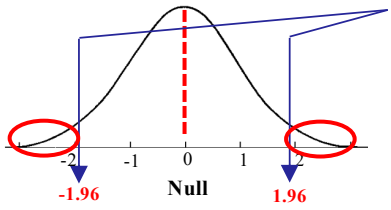
Significance Level	Z score (two tailed)	Z score (one tailed)
0.70	1.04	-0.525 or 0.525
0.75	1.15	-0.675 or 0.675
0.80	1.28	-0.84 or 0.84
0.85	1.44	-1.036 or 1.036
0.90	1.645	-1.28 or 1.28
0.91	1.70	-1.34 or 1.34
0.92	1.75	-1.41 or 1.41
0.93	1.81	-1.476 or 1.476
0.94	1.88	-1.556 or 1.556
0.95	1.96	-1.645 or 1.645
0.96	2.05	-1.751 or 1.751
0.97	2.17	-1.881 or 1.881
0.98	2.33	-2.054 or 2.054
0.99	2.575	-2.326 or 2.326

38

Example 1: non-directional (two tailed)

Step 2 - Set the criteria for a decision

- Consider the level of significance of 5% → α = 0.05. → 1 - α = 0.95
- Locate the Z score (in the table for the standard normal distribution) that represents the **critical values**
- A **critical value** is a cutoff value that sets the boundaries beyond which less than 5% of sample means can be obtained if the null hypothesis is true.



Critical values for nondirectional (two tailed) test with α = 5%

→ **Z score = 1.96**

Rejection regions

39

Example 1: non-directional (two-tailed)

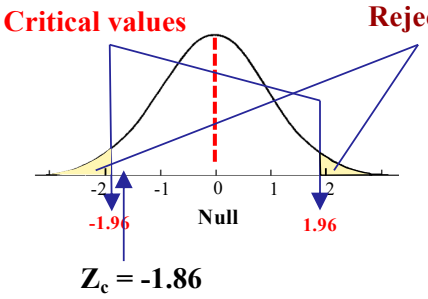
Step 3 - Compute the test statistic

Test statistic:

$$Z_c = \frac{M - \mu}{\sigma/\sqrt{n}} = \frac{78.15 - 83.54}{16.36/\sqrt{32}} = -1.86$$

40

Example 1: non-directional (two-tailed)
Step 4 - Make a decision



The probability of obtaining $\mathbf{Z_c = -1.86}$ is given by the ***P* value**. To obtain ***P* value** for look for 1.86 in the standard normal table.

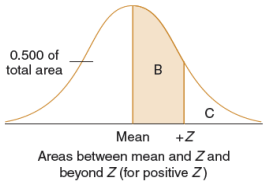
→ **the p value for Z_c = -1.86 is 0.0314**

41

Example 1 — Step 3: Obtain the p value

(test for a **mean, non-directional**, known population; normal Z distribution)

Standard normal table example



To obtain ***P* value** for look for 1.86 in the standard normal table. → **the value is 0.0314**

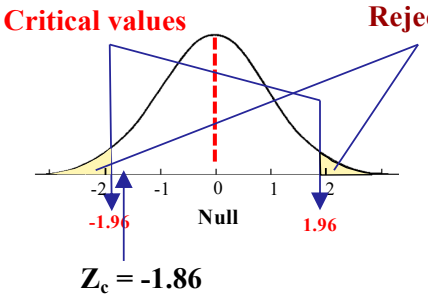
As it is a two-tailed

p = 0.0314 x 2 = 0.0628 → **p = 6.28%**

A Z	B Area Between Mean and Z	C Area Beyond Z	A Z	B Area Between Mean and		A Z	B Area Between Mean and Z	C Area Beyond Z
0.00	0.0000	0.5000	0.11		0.4562	0.21	0.0832	0.4168
0.01	0.0040	0.4960	0.12		0.4522	0.22	0.0871	0.4129
0.02	0.0080	0.4920			0.4483	0.23	0.0910	
1.84	0.4671	0.0329			0.4875	0.0125	2.64	0.4959	0.0041
1.86	0.4686	0.0314	2.25	0.4878	0.0122	2.65	0.4960	0.0040
1.87	0.4693	0.0307	2.26	0.4881	0.0119	2.66	0.4961	0.0039
			2.27	0.4884	0.0116	2.67	0.4962	0.0038

42

Example 1: non-directional (two-tailed)
Step 4 - Make a decision



The probability of obtaining Z_c = **-1.86** is given by the ***P* value**. To obtain ***P* value** for look for 1.86 in the standard normal table.

→ **the p value for Z_c = -1.86 is 0.0314**

As it is a two-tailed

P = 0.0314 x 2 = 0.0628 → **P = 6.28%**

Means that the probability of getting an average of 78.15 if H_0 is true is 6.28%

As *P* > 5%

Retain the null hypothesis (fail reach significance)

Henrique Madeira, DEI-FCTUC, 2018-2023