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 steps: **a more pragmatic approach**

Approach already study:

1. State the hypothesis or claim to be tested
2. Select the criteria for a decision (e.g., $\alpha = 0.05$)
3. Compute the test statistic
4. Make a decision

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing steps: **a more pragmatic approach**

Approach already study:

1. State the hypothesis or claim to be tested
2. ~~Select the criteria for a decision~~ ~~(e.g., $\alpha = 0.05$)~~
3. Compute the test statistic
4. Make a decision

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing steps: **a more pragmatic approach**

Pragmatic approach:

1. State the hypothesis or claim to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

Henrique Madeira, DEI-FCTUC, 2018-2023

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?**

That is, is the execution time in the new configuration smaller than in the previous one?

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

Example 2 - Step 1: State the hypothesis

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

- **H_0** – The new configuration has no effect on the execution time of the SQL packaged. → The average execution time is 83.54
- **H_1** – The execution time of the SQL packaged is **smaller** in the new configuration

We are testing whether the null hypothesis H_0 is true

Note that only the alternate hypothesis changed.

Directional or one-tailed tests are hypothesis tests where the alternative hypothesis is stated as greater than (>) or less than (<) the value stated in the null hypothesis

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 7

7

Example 2 - Step 2: Compute the test statistic

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

Test statistic:

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

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 8

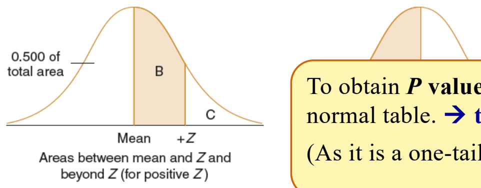
8

Example 2 - Step 3: Obtain the p value

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

Standard normal table example

Z = 1.86



A Z	B Area Between Mean and Z	C Area Beyond Z	A Z	B Area Between Mean and Z	C Area Beyond Z	A Z	B Area Between Mean and Z	C Area Beyond Z
0.00	0.0000	0.5000	0.11	[illegible]	0.4562	0.21	0.0832	0.4168
0.01	0.0040	0.4960	0.12	[illegible]	0.4522	0.22	0.0871	0.4129
0.02	0.0080	0.4920	0.13	0.0517	0.4483	0.23	0.0910	[illegible]
1.84	0.4671	0.0329	2.24	0.4875	0.0125	2.64	0.4959	0.0041
1.85	0.4678	0.0322	2.25	0.4878	0.0122	2.65	0.4960	0.0040
1.86	0.4686	0.0314	2.26	0.4881	0.0119	2.66	0.4961	0.0039
1.87	0.4693	0.0307	2.27	0.4884	0.0116	2.67	0.4962	0.0038

To obtain ***P* value** look for 1.86 in the standard normal table. → **the value is 0.0314 → p = 3.14%**

(As it is a one-tailed we do not double the p)

Henrique Madeira, DEI-FCTUC, 2018-2023

Example 2 - Step 4: Make a decision

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

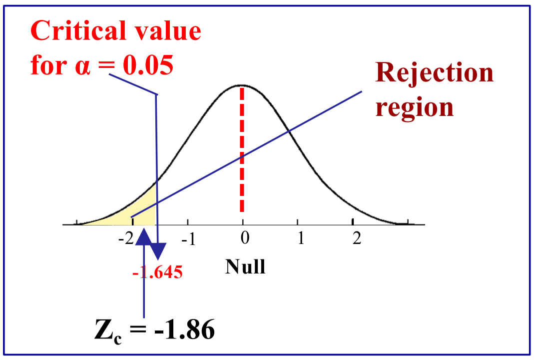
The probability of obtaining $\mathbf{Z_c = -1.86}$ is given by the ***p* value**.

Since **p = 0.0314** this means that the probability of getting an average of 78.15 <u>if H_0 is true</u> is 3.14%

Then I can **reject the H_0** with at least 95% of confidence

The execution time of the SQL packaged is smaller in the new configuration

Classic view



Henrique Madeira, DEI-FCTUC, 2018-2023

Types of errors

The conclusion in Step 4 could be wrong, as we are looking at a sample with a limited number n of elements

False positive

		Decision	
		Retain H_0	Reject H_0
Truth in the population	True	**Correct** $1 - \alpha$	**Type I error** α
	False	**Type II error** β	**Correct** $1 - \beta$ **(Power)**

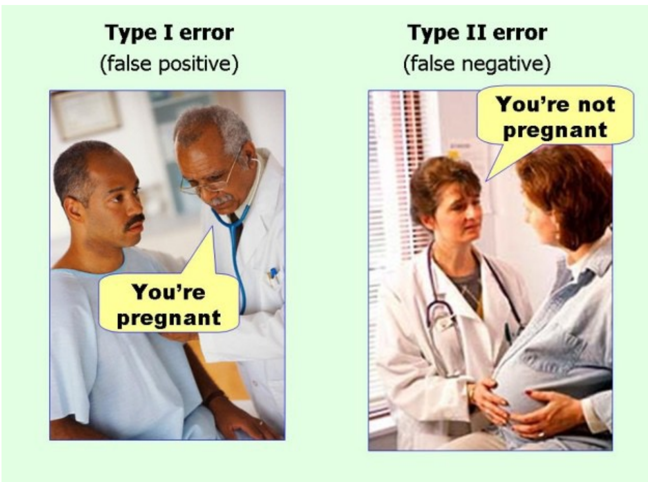
False negative

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 11

11

Types of errors



Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 12

12

Some errors are worse than others

The incorrect decision is to retain a false null hypothesis. This is equivalent of doing nothing. We can do more experiments and test again the null hypothesis. Not so bad…

…ld be wrong, as we are looking
…mber n of elements

False positive

		Decision	
		Retain H_0	Reject H_0
Truth in the population	True	**Correct** $1 - \alpha$	**Type I error** α
	False	**Type II error** β	**Correct** $1 - \beta$ **(Power)**

False negative

Henrique Madeira, DEI-FCTUC, 2018-2023

13

Some errors are worse than others

The incorrect decision is to reject a true null hypothesis. This means rejecting a previous notions of truth that are in fact true (this is equivalent to finding an innocent person guilty).

The … …s we are looking
at a … …nents

False positive

		…ion	
		Retain H_0	Reject H_0
Truth in the population	True	**Correct** $1 - \alpha$	**Type I error** α
	False	**Type II error** β	**Correct** $1 - \beta$ **(Power)**

False negative

Note that you can directly control the probability of a Type I error by stating an alpha level

Henrique Madeira, DEI-FCTUC, 2018-2023

14

The decision we are looking for

Strong conclusion
This is the decision we are looking for when we test the hypothesis. If we test it, it means we have doubts about such hypothesis.

ng, as we are looking
elements

		Decision	
		Retain	Reject H_0
Truth in the population	True	**Correct** $1 - \alpha$	**Type I error** α
	False	**Type II error** β	**Correct** $1 - \beta$ **(Power)**

False positive

False negative

Decision we want

Henrique Madeira, DEI-FCTUC, 2018-2023

The example again: **some questions**

Assume you a
information s
execution tim

What should we do if we cannot have a relatively large number of samples?

From historical data (thousands of previous packag
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 ... of the database and run the package several ti

What should we do if we don't know the standard deviation of the population?

Questions:

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

In this cases we should use the t Test

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

32 times

Avg = 78.15

Henrique Madeira, DEI-FCTUC, 2018-2023

T-test

- The **t test** follows a Student's T-distribution (if the null hypothesis is true)
- T-test should be applied when:
  - The **sample size is small** (n < 30)
  - The populations' **standard deviation is not known**

(when the number of samples is large, t test and z test give similar results)

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing using T-test (one sample)

- Follows the same steps as for the Z test
- The critical value comes from the **T table** (considering $n - 1$ degrees of freedom)
- The **test statistics** is now the t-test (similar formula)

Mean of the sample

Mean of the population (assumed, often stated as a target)

$$t_c = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Standard deviation of the sample

Number of elements of the sample

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing steps: pragmatic approach

Steps:

1. State the hypothesis or claim to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

$$t_c = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 19

19

Example 3 - Hypothesis testing using T-test (one sample)

- A professor wants to know if their students are proficient in C programming. The professor wants the class to be able to score above 70 (0-100 scale) on the test (but doesn't want to examine all the students).
- The professor selects 6 students at random from the class and give them a C programming test.
- The six students get scores of 62, 92, 75, 68, 83, and 95.
- **Can the professor have 90% confidence that the mean score for the class on the test would be above 70?**

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 20

20

Example 3: **t test** (one sample)
Step 1- State the hypothesis

- **H_0: $\mu = 70$**
  In words: the class knows how to program in C with a proficiency equivalent to 70 in the C programming test
- **H_1: $\mu_1 > 70$**
  The class is better on C programming than the score of 70

Example 3: **T test** (one sample)
Step 2 - Compute the test statistic

- Average of the sample: 79.17
- Standard deviation of the sample: 13.17

Test statistic:

$$t_c = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = 1.71$$

Example 3: **T test** (one sample)
Step 3 – Obtain p

- Co
- Lo
  tai
- As
- Lo
  –
  –

In a non-directional two-tailed test, we would divide the **α** value in half

The probability of obtaining **t = 1.71** is given by the ***p* value**. To obtain the ***P* value** look for 1.71 in the **t table**, for **df = 5**

df = 5

Table entry for *p* and *C* is the critical value *t** with probability *p* lying to its right and probability *C* lying between −*t** and *t**.

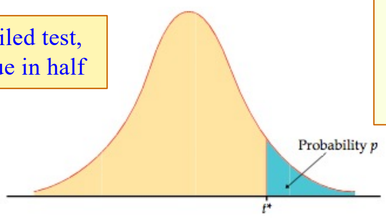


TABLE D

t distribution critical values

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
	Upper-tail probability *p*											
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	[illegible]	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140

→ t score = 1.71

Example 3: **T test** (one sample)
Step 4 - Make a decision

The probability of obtaining **t = 1.71** is given by the ***p* value**. To obtain the ***P* value** look for 1.71 in the **t table**, for **df = 5**

→ the P value is between 5% and 10% (P = 7.4%)

As *p* < 10% → **Reject the null hypothesis(reach significance)**

Means that the probability of getting an average score of 79.17 if **H_0 is true** is 7.4%

Conclusion: The class is better on C programming than the score of 70