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

2

Measuring the size of an effect

- A decision to reject the null hypothesis means that **an effect is significant**. Hypothesis testing does not inform on how big the effect is.
- **Effect size** is a statistical measure of the size of an effect in a population. It particularly makes sense when the null hypothesis is rejected.
- **Cohen's *d*** measures the number of standard deviations an effect shifted above or below the population mean stated by the null hypothesis

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 3

3

Cohen's d measure formula

Mean of the sample

Mean of the population

Standard deviation of the population

$$\text{Cohen's } d = \frac{M - \mu}{\sigma}$$

Cohen's effect size conventions are often used to interpreter the effect size

If values of *d* are negative, the effect shifted below the population mean

Description of Effect	Effect Size (*d*)
Small	$\lvert d \rvert < 0.2$
Medium	$0.2 < \lvert d \rvert < 0.8$
Large	$\lvert d \rvert > 0.8$

In practice, the value of *p* also gives an idea of the effect size.

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in Informatics Engineering , DEI-FCTUC, 2023/2024 4

4

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 co**

$$\textbf{Cohen's } \boldsymbol{d} = \frac{M-\mu}{\sigma} = \frac{78.15-83.54}{16.36} = -0.33$$

The observed effect shifted 0.33 standard deviations below the mean

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

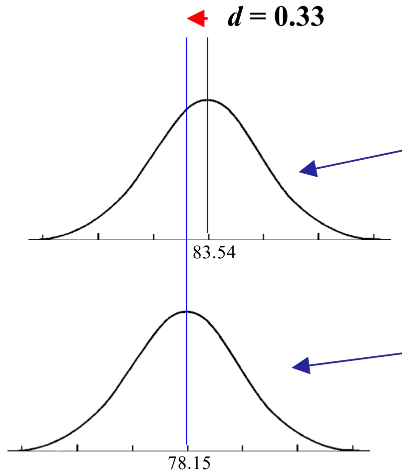
32 times

Avg = 78.15

Henrique Madeira, DEI-FCTUC, 2018-2023

5

The example again: *Cohen's d*



Population distribution assuming the null hypothesis is true

Population distribution assuming the null hypothesis is false with a 5.39 points effect

Henrique Madeira, DEI-FCTUC, 2018-2023

6

T-test

- The **T test** follows a Student's T-distribution (if the null hypothesis is true)
- Two types:
  - **One-sample T-tests** → used to compare a sample mean with the known population mean
  - **Two-sample T-tests** → used to compare two samples.
- T-test should be applied when:
  - The **sample size is small** (n < 30)
  - The populations' **standard deviation is not known**

Independent samples:
unrelated separate groups

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

Henrique Madeira, DEI-FCTUC, 2018-2023

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

7

Hypothesis testing using T-test (two samples)

- Follows the same steps as for the Z test
- The critical value comes from the **T table** (the degrees of freedom is the smaller n_1-1 and n_2-1)
- The **test statistics** is now the **two sample T-test**

Means of the two samples

Hypothesized difference between the population means (0 if testing for equal means)

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Standard deviation of the two samples

Number of elements of the two samples

Henrique Madeira, DEI-FCTUC, 2018-2023

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

8

Example 4 - Hypothesis testing using T-test
(two independent samples)

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?**

Important: we consider that the measurement samples obtained with each configuration are independent.

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	

$\mu_1 = 78.15$ $\mu_2 = 73.58$
$s_1 = 7.94$ $s_2 = 8.33$
$n = 13$ $n = 12$

Henrique Madeira, DEI-FCTUC, 2018-2023

Example 4: t test (two independent samples)
Step 1- State the hypothesis

- **H_0: $\mu_1 = \mu_2$**
  In words: configuration A and B are equivalent concerning the execution time of the SQL package
- **H_1: $\mu_1 > \mu_2$**
  Configuration B is faster than configuration A (i.e., the execution time of the SQL package is higher in configuration A)

Henrique Madeira, DEI-FCTUC, 2018-2023

Example 4: **t test** (two independent samples)

Step 2 - Compute the test statistic

Sample	Configuration	n	x	s
1	A	13	78.15	7.94
2	B	12	73.53	8.33

Test statistic:

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{78.15 - 73.58 - 0}{\sqrt{\frac{7.94^2}{13} + \frac{8.33^2}{12}}} = 1.402$$

Henrique Madeira, DEI-FCTUC, 2018-2023

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

11

Example 4: **t test** (two independent samples)

Step 3 – Calculate *p*

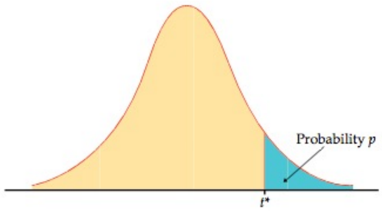


Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

***p* = 1.402**

As the sizes of the samples are n = 13 and n = 12, the degree of freedom is the smaller n -1 → 11

α = 0.05
df = 11

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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	0.765	0.978	1.250	1.638	2.353	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	0.741	0.941	1.190	1.533	2.132	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	0.727	0.920	1.156	1.476	2.015	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	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	[illegible]	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
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

→ p between 0.05 and 0.1

Henrique Madeira, DEI-FCTUC, 2018-2023

Experimental Methods in Computer Science, Master in I 12

12

Example 4: **t test** (two independent samples)
Step 3 – Make a decision

The ***p* value** for **t = 1.402** and **df = 11** is between 5% and 10% (from the T table)
→ the accurate ***p*** value is 0.0942 (***p*** = 9.42%) (from an online calculator)

Means that the probability of getting an average score of 73.58 if H_0 is true is 9.42%

→ **Retain the null hypothesis (fail reaching significance)**

We could not prove that configuration B is faster than A with 95% confidence

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing steps

Pragmatic approach:

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

When the sample size is large (n > 30) and the σ of the population is known

$$Z_c = \frac{M - \mu}{\sigma / \sqrt{n}}$$

Z test – to compare a sample mean with the population mean

Henrique Madeira, DEI-FCTUC, 2018-2023

Hypothesis testing steps

Pragmatic approach:

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

When the size of the samples is large (n ≥ 30)

$$Z_c = \frac{\bar{x}_1 - \bar{x}_1 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two samples Z test – to compare the means of two independent large samples

Henrique Madeira, DEI-FCTUC, 2018-2023

15

Hypothesis testing steps

Pragmatic approach:

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

When the sample size is small (n < 30) and the μ of the population is not known (it is a target).

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

One sample T test – to compare a sample mean with the population mean

Henrique Madeira, DEI-FCTUC, 2018-2023

16

Hypothesis testing steps

Pragmatic approach:

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

When the size of the samples is small (n < 30)

$$t_c = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two samples T test – to compare the means of two independent samples

Hypothesis testing steps

Pragmatic approach:

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

The test statistic is converted into a conditional probability, the **p value**. It can be obtained using the t tables or using p value calculation sites/programs.

The p value answers the question "**If the null hypothesis is true, what is the probability of observing the measured data?**"

Hypothesis testing steps

P

1.

2.

3. Obtain

4. Make a decision

Small ***p*** values provide evidence against the null hypothesis, as it means that the observed data are unlikely when the null hypothesis is true.

Conventions:
- $p \geq 0.10$ → **the observed difference is "not significant"**
- $0.05 \leq p < 0.10$ → **the observed difference is "marginally significant"**
- $0.01 \leq p < 0.05$ → **the observed difference is "significant"**
- $p < 0.01$ → **the observed difference is "highly significant"**

Henrique Madeira, DEI-FCTUC, 2018-2023

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

19