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

Hypothesis Testing
Dependent samples

Samples

So far we have studied hypothesis testing in different sampling scenarios with (approximately) known distributions (parametric tests):

- Type of samples
  - Large (≥30) samples: Z (normal) distribution
  - Small (<30) samples: T Student distribution

- Number of samples
  - **Single sample**: only one group of observations; test against a hypothetical **mean** or **proportion**; Z test for large samples and T test for small ones.
  - **Two samples**: two groups of observations; test the difference between **means** or **proportions**; Z test for large samples and T test for small ones.
  - **Three or more samples**: several groups; test the variance (**ANOVA**)

- Nature of the samples
  - **Independent samples**: groups are not related and observations are truly independent
  - **Dependent samples**: when one observation/measurement in a group is related to one observation in a another group. Also called matched pairs, matched samples, etc.

Examples of dependent samples

Example 1:

Sample 1: Downloads per day of the Android applications being marketed by your company

Sample 2: Downloads per day of the same group of Android applications after an advertisement campaign of your company at Google

Example 2:

Sample 1: Number of code security vulnerabilities found in code inspections by 10 engineers

Sample 2: Number of code security vulnerabilities found in code inspections by the same 10 engineers after a security programming training.

Examples of dependent samples

Example 1:

Sample 1: Downloads per day of the Android applications being marketed by your company

Sample 2: Downloads per day of the same group of Android applications after an advertisement campaign of your company at Google.

Example 2:

Sample 1: Number … eers

Sample 2: Number … 10 engineers after a security programming training.

Samples are dependent because the measurements can be paired with respect to each application (example 1) or each engineer (example 1).

But there are subtle differences between these two examples…

Test for the difference between means
Two dependent samples

The two-sample hypothesis test with dependent samples is based on the mean $\bar{d}$ of the differences between paired data entries in the dependent samples.

$$\bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

Difference between entries for a data pair

Number of pairs

The standard deviation S_d of the differences between the paired data entries in the dependent samples

$$s_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023

Test for the difference between means
Two dependent samples

We can use the mean $\bar{d}$ of the differences between paired data entries in the dependent samples and the standard deviation S_d if and only if the following **conditions** are met:

$$\bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

$$s_d = \sqrt{\frac{n\left(\sum d^2\right) - \left(\sum d\right)^2}{n(n-1)}}$$

Conditions:

1. The samples must be randomly selected.
2. The samples must be dependent (paired).
3. Both populations must be normally distributed.

If these conditions are met, then the sampling distribution for $\bar{d}$ is approximated by a normal distribution for $n \geq 30$ or by a ***T*** distribution with n-1 degrees of freedom if $n < 30$.

Test statistic for two dependent samples

$$z = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

Test statistics for **large sets** of paired samples ($n \geq 30$)

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

Test statistics for **small sets** of paired samples ($n < 30$). The degree of freedom is n-1.

Test statistic for two dependent samples

$$z = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

Mean of the difference between paired entries in the dependent samples

Assumed difference between means (usually zero because H_0 is conservative)

The standard deviation of the differences between the paired data entries in the dependent samples

Number of paired samples. In ***t statistics*** (formula below) the degree of freedom is n-1

Two dependent samples hypothesis test: **Steps**

Follows the same basic steps of the other hypothesis testing:

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

For large samples

$$z = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$\bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

$$s_d = \sqrt{\frac{n\left(\sum d^2\right) - \left(\sum d\right)^2}{n(n-1)}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023

Two dependent samples hypothesis test: **Steps**

Follows the same basic steps of the other hypothesis testing:

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

$$\bar{d} = \frac{\sum(x_{1i} - x_{2i})}{n}$$

For small samples

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$s_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023

Two dependent samples hypothesis test: **Steps**

Follows the same basic steps of the other hypothesis testing:

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

Using the standard Z table or online calculators.

Or T table/calculators with df = n-1 for small samples

Example: Hypothesis test for two dependent samples

Your company is unhappy with the quality of the code produced by the Web application developers, as the number of security vulnerabilities such as SQL injection and cross-site scripting (XSS) is quite high.

You gave the developers a written test that consists in asking them to write code snippets for typical situations where developers tend to make code with security vulnerabilities and you record their grades in the test.

The developers went through a specific training (quite expensive, by the way…) on how to write safe Web applications code and after the training you repeated the written test (not exactly the same test but a similar one).

The table below show the scores of the developers in both tests.

Developer	1	2	3	4	5	6	7	8	9	10
Score (before)	85	79	70	76	81	78	72	65	78	65
Score (after)	80	85	89	86	92	75	78	60	85	80

Can you report with 95% confidence that the training improved the skills of the Web application designers?

Example: hypothesis test for two dependent samples
Step 1 - State the hypothesis be tested

- **H_0: $\mu_d \leq 0$**
  The test scores after the training are not better than the test scores before the training (i.e., the training did not improve the skills of the Web developers)

- **H_1: $\mu_d > 0$**
  The training improved the skills of the Web developers and the test scores after the training are better than the ones before the training (**Claim)**.

Henrique Madeira, DEI-FCTUC, 2018-2023

Example: hypothesis test for two dependent samples
Step 2 - Compute the test statistic

Add intermediate calculations to the table

Developer	1	2	3	4	5	6	7	8	9	10
Score (before)	85	79	70	76	81	78	72	65	78	65
Score (after)	80	85	89	86	92	75	78	60	85	80
d	-5	6	19	10	11	-3	6	-5	7	15
d²	25	36	361	100	121	9	36	25	49	225

$\sum d = 61$

$\sum d^2 = 987$

$$\bar{d} = \sum \frac{\sum d}{n} = \frac{61}{10} = 6.1$$

$$S_d = \sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n(n-1)}} = \sqrt{\frac{10(987) - 3721}{10(10-1)}} = \sqrt{68.32} = 8.27$$

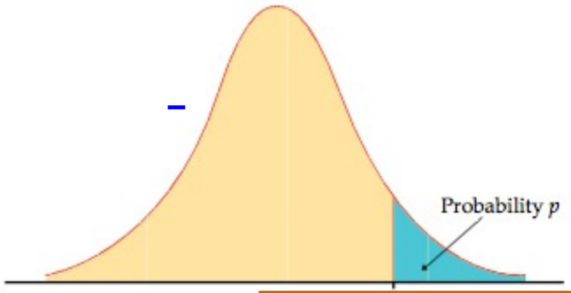
$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{6.1 - 0}{8.27/\sqrt{10}} = \frac{6.1}{2.61} = 2.33$$

Test statistic

Henrique Madeira, DEI-FCTUC, 2018-2023

Example: hypothesis test for two dependent samples
Step 3 – Obtain p value

Find the probability for
t = 2.33
n = 10 → df = 9



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

The probability is between 2% and 2.5%

TABLE D

t distribution critical values

	Upper-tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
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	1.156	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	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	[illegible]	[illegible]	2.896	3.355	3.833	4.501	5.041
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	[illegible]	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
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

Example: hypothesis test for two dependent samples
Step 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"**

As ***p*** is between 2% and 2.5% the effect is significant.
We can reject H0 with 95% of confidence.

<u>The training really improved the skills of the developers!</u>

Use of dependent-samples: summary

Use when you have:

- Repeated measures for the same individual/system/component/…
- Studies with matched pairs of family members.

Advantages:

- Known sources of potential bias are controlled
- The standard deviation of the test statistic is usually smaller, making the power of the test greater than in a Z or T test

Disadvantages:

- In some cases is hard to find the same objects/participants
- When the null hypothesis is rejected, often is difficult to argue that the difference is due to global events and not to the test-retest of the same individuals.