

# Experimental Analysis of Algorithms.

## Lab 12: Statistical Analysis

May, 15 2026

### Hypothesis testing

- **parametric tests:** assume the data arise from a normal distribution (with mean  $\mu$  and variance  $\sigma^2$ )  
Unpaired/paired *t-test*
- **non-parametric tests:** do not make parametric assumptions (most often based on ranks, as opposed to raw values)  
*Sign test, Wilcoxon signed-ranks test, Mann Whitney U-test*

## 1 Parametric tests

1. **Z-test** for mean (the population standard deviation  $\sigma$  is known)

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Example 9.25 from Baron's book (section 9.4.6)

2. **t-test** (the population standard deviation  $\sigma$  is unknown)  
Assumption: the data are normally distributed.

**Two sample t-test** is used to determine if two population means are equal. The test also specify how significant the differences are.

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- Estimate the means and variances:  $\bar{x}, s_x^2, \bar{y}, s_y^2$ .
- The *t* statistic is computed as follows, assuming equal variance:

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{1/n_x + 1/n_y}}$$

where  $s = \sqrt{\frac{s_x^2(n_x-1)+s_y^2(n_y-1)}{n_x+n_y-2}}$  is an estimator of the common standard deviation of the two samples. The degrees of freedom:  $n_x + n_y - 2$ .

- Compute the  $p$ -value using the Student's  $t$ -distribution table. The level of statistical significance is expressed as a  $p$ -value (between 0 and 1). The smaller the  $p$ -value, the stronger the evidence that you should reject  $H_0$ .  
If the calculated  $p$ -value is less than the threshold chosen for statistical significance (usually 0.05, or 0.01), then the null hypothesis is rejected in favor of the alternative hypothesis.

**Example:** compare the blood pressure of male consultant doctors with the junior resident female doctors.

*The null hypothesis:* there is no statistically significant difference in the mean of male consulting doctors and junior resident female doctors.

Python: `ttest_ind()` function: [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest\\_ind.html](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html).

```
from scipy import stats
female_doctor_bps = [128, 127, 118, 115, 144, 142, 133, 140, 132, 131,
111, 132, 149, 122, 139, 119, 136, 129, 126, 128]
male_consultant_bps = [118, 115, 112, 120, 124, 130, 123, 110, 120, 121,
123, 125, 129, 130, 112, 117, 119, 120, 123, 128]

stats.ttest_ind(female_doctor_bps, male_consultant_bps)
```

Interpretation: the  $t$ -statistic value: 3.514; calculate a  $p$ -value using the  $t$ -statistic value and the degrees of freedom (38). The  $p$ -value is 0.0012, less than the standard thresholds of 0.05 or 0.01, so reject the null hypothesis: there is a statistically significant difference between the resting systolic blood pressure of the resident female doctors and the male consultant doctors.

### Paired $t$ -test

Previously, the observations in the two samples have been completely independent of one another. To compare two related samples, e.g. before and after test, use a paired  $t$ -test.

The  $t$  statistic is computed as follows:

$$t = \frac{\bar{x}_D - \mu_0}{s_D/\sqrt{n}}$$

where  $\bar{x}_D$  and  $s_D$  are average and standard deviation of differences of all pairs,  $\mu_0$  is a constant (is non-zero if we want to test whether the average of the difference is significantly different from  $\mu_0$ ). The degrees of freedom:

$n - 1$ ,  $n$  the number of pairs.

$$H_0 : \mu_{diff} = 0$$

$$H_a : \mu_{diff} \neq 0$$

**Example:** measure the amount of sleep got by patients before and after taking soporific drugs to help them sleep.

*The null hypothesis:* the soporific drug has no effect on the sleep duration of the patients.

Python: use `ttest_rel()` function from *scipy*:

```
control = [8.0, 7.1, 6.5, 6.7, 7.2, 5.4, 4.7, 8.1, 6.3, 4.8]
treatment = [9.9, 7.9, 7.6, 6.8, 7.1, 9.9, 10.5, 9.7, 10.9, 8.2]
```

```
stats.ttest_rel(control, treatment)
```

Interpretation: the  $t$ -statistic value: -3.624; compute the  $p$ -value by using the degrees of freedom (9). The  $p$ -value: 0.0055, which is below than the standard thresholds of 0.05 or 0.01, so reject the null hypothesis (there is a statistically significant difference in sleep duration caused by the soporific drug).

## 2 Non-parametric tests

1. **Sign test** is the simplest non-parametric test that can be used instead of  $t$ -test if the data violates the assumptions of normality. The approach analyzes only the signs of the difference scores.

**Example:** The marks of a surprise test for students were as follows: 8, 6, 4, 2, 5, 6. After a week of self-practice, the marks of a new test are: 6, 8, 8, 9, 4, 10 (the student marks in the same order). Check if the marks of the students improved.

- **Null hypothesis  $H_0$ :** The median difference is 0.
- **Alternative hypothesis  $H_A$ :** The median difference is positive.

If we had a similar number of positive and negative differences, then  $H_0$  would be true. We have 4 positive and 2 negative signs; the statistic  $S = 4$ , the  $p$ -value  $\text{Binomial}(6, 0.5)$  is .34. Not reject  $H_0$ .

2. **Wilcoxon signed-rank test** is a non-parametric univariate test; it is an alternative to the *dependent t-test*. The approach takes into account the values of the difference scores.

**Example:** analyze the blood pressure before and after the intervention (test if the intervention had a significant effect on the blood pressure). Dataset.

- $H_0$ : The **difference between the pairs** follows a symmetric distribution around zero.
- $H_A$ : The difference between the pairs does not follow a symmetric distribution around zero.

If the  $p$ -value is less than the chosen significance level (0.05), reject the null hypothesis.

Use from python <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.wilcoxon.html>

```
import pandas as pd
df = pd.read_csv("data/blood_pressure.csv")
print(df[['bp_before', 'bp_after']].describe())

#compute the difference between the two conditions
df['bp_difference'] = df['bp_before'] - df['bp_after']
df['bp_difference'][df['bp_difference']==0]

print(stats.wilcoxon(df['bp_difference']))
print(stats.wilcoxon(df['bp_before'], df['bp_after']))
```

The blood pressure before the intervention was higher ( $M = 156.45 \pm 11.39$  units), compared to the blood pressure post intervention ( $M = 151.36 \pm 14.18$  units);  $t = 2234.5$ ,  $p = 0.0014$ ; the  $p$ -value is below the significance level, so reject  $H_0$ . The samples were likely drawn from populations with differing distributions - there was a statistically significant decrease in blood pressure.

3. **The Mann Whitney U-test** allows comparison of two groups of data where the data is not normally distributed. It is the non-parametric version of the *Student t-test for independent data* samples.

The test is used to interpret whether **there are differences in the "distributions" of two groups** or differences in the "medians" of two groups. The hypothesis is:

- $H_0$ : the distribution of scores for the two groups are equal

- $H_A$ : the distribution of scores for the two groups are not equal

**Example:**

Python: use <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html>

```
# Create two groups of data
group1 = [2, 5, 7, 3, 5, 8, 34, 1, 3, 5, 150, 3]
group2 = [10, 19, 11, 12, 15, 19, 9, 17, 1, 22, 9, 8]

from scipy import stats
# Calculate u and probability of a difference
u_statistic, pVal = stats.mannwhitneyu(group1, group2)
print (pVal)
```

Because the  $p$ -value (0.0149) is less than the chosen significance level (0.05), reject the null hypothesis.

**Homework**

1. Exercises 9.7b, 9.9b, 9.13, 9.14
2. HCI text background color  
Determine which of two background colors for computer text (yellow or cyan) is easier to read (consider the speed with which a task described by the text is performed).  
The study randomly assigns 35 students to one of two versions of a computer program that presents text describing which of several icons the user should click on. The two versions differ in the background color for the text. The program measures how long it takes until the correct icon is clicked (“reaction time” measured in milliseconds). The program reports the average time for 20 trials per subject. The data file.
3. Exercises 10.10, 10.16, 10.23
4. Optional: exercises 10.15, 10.22

Exercises are from *M. Baron. Probability and Statistics for Computer Scientists*.

For python examples, see *Statistics and Machine Learning in Python, Ch. 4.1 Univariate statistics*.