Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Seventh lecture, December 4, 2023

Hypothesis testing

Hypothesis testing

Statistical tests are applied extensively in various fields of science. We might want to test, for example:

- If one concrete type is stronger than another (competing) concrete type.
- If there a difference in the average salaries of men and women across the population.
- Whether or not a new medicine lowers systolic blood pressure.

⋮

Hypothesis testing

A statistical hypothesis is a hypothesis that is tested using probabilities. Statistical testing is based on setting general statistical assumptions, a null hypothesis and an alternative hypothesis, and on selecting a suitable test statistic. The value of the selected test statistic is calculated from a sample of observations.

Assumptions

- General statistical assumptions include assumptions about the population, sampling method, and about the distribution of the observations.
- Statistical assumptions hold throughout the testing process.
- Statistical assumptions may, and should, be tested separately.

Null hypothesis

- The statement about a population parameter that is being tested is called the null hypothesis H_0.
- The null hypothesis is assumed to be true, unless there is strong evidence that indicates otherwise.
- If strong evidence against the null hypothesis is found, then it is rejected.
- In simple statistical tests, the null hypothesis can often be stated as

$$H_0 : \theta = \theta_0,$$

  where θ is the parameter being tested and θ_0 is a fixed value of the parameter.
- The null hypothesis is often of the form "is the same" or "no difference".

Alternative hypothesis

- If the null hypothesis H_0 is rejected, then the alternative hypothesis H_1 is accepted.
- If the alternative hypothesis can be stated as $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$, then it is called a one tailed alternative hypothesis.
- If the alternative hypothesis can be stated as $H_1 : \theta \neq \theta_0$, then it is called a two tailed alternative hypothesis.
- The alternative hypothesis is often of the form "not the same" or "different".

It is not always easy to decide whether one tailed or two tailed alternative hypothesis should be used.

Do not fish for favorable results by using one tailed alternative hypothesis!

The use of one tailed alternative hypothesis must be justified by the context.

Test statistic

- A test statistic compares the observations and the null hypothesis H_0.
- A test statistic is a random variable and its value depends on the observations.
- A test statistic is used in evaluating the probability of getting the observed value of the statistic, under the assumption that the null hypothesis H_0 is true.
- The distribution of the test statistic under the null hypothesis H_0 must be known for comparing the observations and the null hypothesis H_0.

Critical value

- The expected value of a chosen test statistic is calculated under the null hypothesis H_0.
- If the observed value of the test statistic is close to the expected value, no strong argument against the null hypothesis H_0 is found.
- If the observed value of the test statistic is far away from the expected value, then evidence against the null hypothesis H_0 is found.
- The set of values of the test statistic for which the null hypothesis is rejected (i.e., the set of the values that are far away from the expected value) is called the critical region.
- The threshold values defining the critical region are called the critical values.

p-value

The p-value of a statistical test is the probability, assuming that the null hypothesis H_0 is true, of observing at least as extreme value as the observed value of the test statistic.

Rejecting or not rejecting the null hypothesis H_0 is based on the p-value. Statistical software can be used to calculate the p-value.

The significance level α of a test statistic is the smallest p-value that is accepted without rejecting the null hypothesis H_0. It is possible to use pre-selected significance levels and the corresponding critical regions. Commonly used significance levels α are 0.05, 0.1, 0.01, and 0.001.

If the significance level is $\alpha = 0.05$ and the p-value of the test statistic is < 0.05, then the null hypothesis H_0 is rejected.

p-value

The p-value of a test statistic is calculated as follows:

1. Calculate the value of the test statistic using the observations.
2. Assuming that the null hypothesis H_0 is true and based on the known distribution of the test statistic, calculate the probability of the value of the test statistic being as extreme, or more extreme, as it is.

The null hypothesis H_0 can be rejected, if the p-value is small enough.
The smaller the p-value, the stronger the evidence against H_0.

Errors

There are two types of errors related to the rejection of the null hypothesis H_0:

- Type 1 error: True null hypothesis is rejected.
- Type 2 error: False null hypothesis is not rejected.

The type 1 error rate is the probability of rejecting the null hypothesis given that it is true. Thus type 1 error rate is equal to the significance level α.

The type 2 error rate is the probability of not rejecting the null hypothesis given that it is false. Type 2 error rate is in general a function of the possible distributions, often determined by a parameter, under the alternative hypothesis. The power of a test statistic is equal to $1 -$ (type 2 error rate). Thus, the power of a test statistic is also a function of the possible distributions. As the power increases, the chance of a type 2 error decreases – one is more likely to detect significant differences when they truly exist.

In statistical testing, type 1 errors are generally considered worse than type 2 errors. That is why the significance level α is usually selected to be small.

p-value, one tailed and two tailed alternative hypothesis

Let z be the value of a test statistic Z calculated from the observations.

If the one tailed alternative hypothesis is given as $H_1 : \theta > \theta_0$, then the p-value of the test is

$$p = \mathbb{P}(Z \geq z \mid H_0).$$

If the one tailed alternative hypothesis is given as $H_1 : \theta < \theta_0$, then the p-value of the test is

$$p = \mathbb{P}(Z \leq z \mid H_0).$$

If the alternative hypothesis is two tailed, $H_1 : \theta \neq \theta_0$, then the p-value of the test is

$$p = 2 \min(\mathbb{P}(Z \leq z \mid H_0), \mathbb{P}(Z \geq z \mid H_0)).$$

Steps of statistical hypothesis testing

1. State the hypotheses and general assumptions.
2. Select a test statistic.
3. Pick a sample such that the general assumptions hold.
4. Calculate the value of the test statistic using the sample.
5. Calculate the p-value corresponding to the observed value of the test statistic.
6. Draw conclusions and reject/do not reject the null hypothesis.

t-tests

One sample t-test

The one sample t-test compares the expected value of a random variable to a given constant.

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x. Assume that the observed values come from the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The null hypothesis: $H_0 : \mu = \mu_0$.

The possible alternative hypotheses:

$$H_1 : \mu > \mu_0 \text{ (one tailed)},$$
$$H_1 : \mu < \mu_0 \text{ (one tailed)},$$
$$H_1 : \mu \neq \mu_0 \text{ (two tailed)}.$$

One sample t-test

- The t-test statistic is
$$t = \frac{\overline{x} - \mu_0}{s_n/\sqrt{n}}.$$
- If the null hypothesis H_0 is true, then the test statistic follows Student's t-distribution with $n-1$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0, i.e., $\mathbb{E}[t] = 0$.
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found. On the other hand, the null hypothesis H_0 is rejected if the p-value is small enough.
- Python:
`t_stat,p_value = scipy.stats.ttest_1samp(a=`x`,popmean=`μ_0`)`

One sample t-test, normality assumption

- When the one-sample t-test is used, it is assumed that the observations follow the normal distribution.
- If the sample size is large, then one sample t-test is not very sensitive to moderate deviations from normality.
- Even without normality, the one sample t-test is quite reliable if the sample size $n > 25$. That is, unless the distribution is very skewed.
- With sample size $n > 40$, the one sample t-test is quite reliable even for clearly skewed distributions.

One sample t-test – implementation in Python

```
import numpy as np
from scipy.stats import t as tdist

def tTest_1sample(x,mu0,alternative='two-sided'):
    n = len(x)
    xbar = np.mean(x)
    std = np.std(x,ddof=1) # Use Bessel's correction
    t_stat = (xbar-mu0)/(std/np.sqrt(n))
    q = tdist.cdf(t_stat,n-1)
    if alternative == 'less':
        return t_stat,q
    elif alternative == 'greater':
        return t_stat,1-q
    else:
        return t_stat,2*min(q,1-q)
```

Numerical example, one sample t-test

According to the package text, Brand X cookies have 12 chocolate chops in each cookie. The number of chocolate chips of ten randomly selected cookies were calculated and the following data was obtained:

$$\{12, 11, 10, 13, 14, 12, 11, 12, 12, 12\}.$$

We want to test, on significance level 5%, the hypothesis that the expected value of the number of chocolate chops in Brand X cookies is 12.

The sample mean of the chocolate chips is 11.9 and the sample standard deviation is 1.1005. One sample t-test is used and the value of the test statistic is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{11.9 - 12}{1.1005/\sqrt{10}} = -0.287.$$

Assuming normality and i.i.d. observations, under the null hypothesis ($\mu = \mu_0 = 12$), the test statistic follows Student's t-distribution with 9 degrees of freedom.

With significance level 5% and 9 degrees of freedom, the critical values of the test statistic are ± 2.262. Since the observed value of the test statistic $-0.287 > -2.262$ and $-0.287 < 2.262$, the null hypothesis is not rejeced.

The p-value is often observed directly without setting any pre-selected significance level. Probabilities $\mathbb{P}(T \leq t|H_0)$ and $\mathbb{P}(T \geq t|H_0)$ are 0.6098 and 0.3902, respectively. Then the p-value is

$$p = 2\min(\mathbb{P}(Z \leq z \mid H_0), \mathbb{P}(Z \geq z \mid H_0)) = 2 \cdot 0.3902 = 0.7804.$$

The p-value is large and no evidence against the null hypothesis is found.

What went wrong in the previous example? What are the general statistical assumptions when one sample t-test is used?

Two sample t-test

The two sample t-test compares the expected values of two independent variables. We first consider the case when the variances are *not assumed to be equal*.

Two sample t-test, assumptions

Let $x_1, \ldots, x_n$ be the observed values of a random variable x and let $y_1, \ldots, y_m$ be the observed values of a random variable y. Assume that the observed values $x_1, \ldots, x_n$ are i.i.d. and come from the normal distribution $\mathcal{N}(\mu_x, \sigma_x^2)$ and assume that the observed values $y_1, \ldots, y_m$ are i.i.d. and come from the normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, assume that x_i and y_j are independent for all i, j.

The null hypothesis: $H_0 : \mu_x = \mu_y$.

The possible alternative hypotheses:

$$H_1 : \mu_x > \mu_y \text{ one tailed,}$$
$$H_1 : \mu_x < \mu_y \text{ one tailed,}$$
$$H_1 : \mu_x \neq \mu_y \text{ two tailed.}$$

Two sample t-test

- The t-test statistic
$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2/n + s_y^2/m}}.$$
- If the null hypothesis H_0 is true, then the test statistic follows Student's t-distribution with v degrees of freedom, where
$$v = \frac{(s_x^2/n + s_y^2/m)^2}{((s_x^2/n)^2/(n-1)) + ((s_y^2/m)^2/(m-1))}.$$
- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- Python:

```
t_stat,p_value =
scipy.stats.ttest_ind(a=x,b=y,equal_var=False)
```

Two sample t-test – implementation Python

```
import numpy as np
from scipy.stats import t as tdist

def tTest_2sample(x,y,alternative='two-sided'):
    n = len(x); m = len(y)
    xbar = np.mean(x); ybar = np.mean(y)
    stdx = np.std(x,ddof=1); stdy = np.std(y,ddof=1)
    t_stat = (xbar-ybar)/np.sqrt(stdx**2/n+stdy**2/m)
    v = (stdx**2/n+stdy**2/m)**2 \
        /((stdx**2/n)**2/(n-1)+((stdy**2/m)**2/(m-1)))
    q = tdist.cdf(t_stat,v)
    if alternative == 'less':
        return t_stat,q
    elif alternative == 'greater':
        return t_stat,1-q
    else:
        return t_stat,2*min(q,1-q)
```

Two sample t-test, normality assumption

- When the two sample t-test is used, it is assumed that the observations follow the normal distribution.
- If the sample sizes are large, then the two sample t-test is not very sensitive to moderate deviations from normality.
- Even without normality, the two sample t-test is quite reliable, if the sample sizes $n > 25$ and $m > 25$. That is, unless the distributions are very skewed.
- If $n > 40$ and $m > 40$, then the test can be quite safely used even with clearly skewed distributions.

Two sample t-test, equal variances

The two sample t-test has a bit simpler form if the variances are assumed to be equal.

Assumptions and hypotheses are the same as in the general two sample t-test, but the variances of the distributions are assumed to be equal – that is, it is assumed that $\sigma_x^2 = \sigma_y^2$.

Two sample t-test, equal variances

- The t-test statistic
$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{1/n + 1/m}},$$
where
$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$
- If the null hypothesis H_0 is true, then the test statistic follows Student's t-distribution with $n + m - 2$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- Normality assumption can be relaxed as in the general two sample t-test.
- Python:
`t_stat,p_value = scipy.stats.ttest_ind(a=x,b=y,equal_var=True)`

Paired t-test

General two sample t-tests can be applied when the two samples are *independent*.

The paired t-test can be used to compare two measuring equipments by using both equipments to measure the same subject in the same circumstances. (Do two pedometers give the same result?) A paired t-test can be used for example to study if a treatment works by measuring the same subjects before and after the treatment. (Does drinking have an effect on reaction time? Does malnutrition have an effect on memory?) The aim can also be to compare two populations by measuring the same variables of fitted pairs. (Do the voting preferences of couples living together differ from each other?)

Paired t-test

Paired t-test:

- Observations $(x_{i,1}, x_{i,2})$, $i = 1, \ldots, n$, consist of measured pairs of a random variable x.
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- General two sample t-tests should not be used for paired observations.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \ldots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.
- Measurements $x_{i,1}$ and $x_{i,2}$ have on average about the same value if the differences are on average about 0.
- It is now possible to apply the standard one sample t-test to the differences d_i.
- Python:
  `t_stat,p_value = scipy.stats.ttest_rel(a=x,b=y)`
  Remark: `scipy.stats.ttest_rel` assumes $\mu_0 = 0$ by default.

Paired t-test

- General statistical assumptions: differences d_i are i.i.d. and come from the normal distribution.
- The null hypothesis: $H_0 : \mu_d = 0$.
- Possible alternative hypotheses: $H_1 : \mu_d > 0$ (one tailed), $H_1 : \mu_d < 0$ (one tailed) or $H_1 : \mu_d \neq 0$ (two tailed).
- The t-test statistic
$$t = \frac{\overline{d}}{s_d/\sqrt{n}}.$$
- If the null hypothesis H_0 is true, then the test statistic follows Student's t-distribution with $n - 1$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- The normality assumption can be relaxed as in the general one sample t-test.

Paired t-test – implementation in Python

```
def tTest_paired(x,y,alternative='two-sided'):
    return tTest_1sample(x-y,0,alternative)
```

Variance tests

Variance test, assumptions

Let $x_1, \ldots, x_n$ be observed values of a random variable x. Assume that the observed values are i.i.d. and come from the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$.

The possible alternative hypotheses:

$$H_1 : \sigma^2 > \sigma_0^2 \text{ (one tailed)},$$
$$H_1 : \sigma^2 < \sigma_0^2 \text{ (one tailed)},$$
$$H_1 : \sigma^2 \neq \sigma_0^2 \text{ (two tailed)}.$$

Variance test

- The χ^2 test statistic
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}.$$
- If the null hypothesis is true, then the test statistic follows the χ^2 distribution with $n-1$ degrees of freedom.
- The expected value of the test statistic is $n-1$.
- Large and small values of the test statistic (compared to the expected value $n-1$) suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p-value is small enough.
- This test is sensitive to deviations from normality! Variance test does not work, not even with large sample sizes, if the distribution of the observations is skewed.

Variance test – implementation in Python

```
import numpy as np
from scipy.stats import chi2

def varTest(x,sigma_squared,alternative='two-sided'):
    n = len(x)
    Q_stat = (n-1) * np.var(x,ddof=1)/sigma_squared
    q = chi2.cdf(Q_stat,n-1)
    if alternative == 'less':
        return Q_stat,q # one-sided variance test
    elif alternative == 'greater':
        return Q_stat,1-q # one-sided variance test
    else:
        return Q_stat,2*min(q,1-q) # two-sided variance test
```

Variance comparison test, assumptions

Let $x_1, \ldots, x_n$ be observed values of a random variable x and let $y_1, \ldots, y_m$ be observed values of a random variable y. Assume that the observations $x_1, \ldots, x_n$ are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_x, \sigma_x^2)$ and assume that $y_1, \ldots, y_m$ are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, assume also that x_i and y_j are independent for all i, j.

The null hypothesis: $H_0 : \sigma_x^2 = \sigma_y^2$.

The possible alternative hypotheses:

$$H_1 : \sigma_x^2 > \sigma_y^2 \text{ (one tailed)},$$
$$H_1 : \sigma_x^2 < \sigma_y^2 \text{ (one tailed)},$$
$$H_1 : \sigma_x^2 \neq \sigma_y^2 \text{ (two tailed)}.$$

Variance comparison test

- The F-test statistic
$$F = \frac{s_x^2}{s_y^2}.$$
- If the null hypothesis is true, then the test statistic follows F-distribution with $n-1$ and $m-1$ degrees of freedom.
- The expected value of the test statistic is ≈ 1.
- Large and small values of the test statistic (compared to the expected value ≈ 1) suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- This test is also sensitive to deviations from normality and does not work, not even with large sample sizes, if the distribution of the observations is skewed.

Variance comparison test – implementation in Python

```
import numpy as np
from scipy.stats import f as Fdist

def Ftest(x,y,alternative='two-sided'):
    dfx = len(x)-1
    dfy = len(y)-1
    F_stat = np.var(x,ddof=1)/np.var(y,ddof=1)
    q = Fdist.cdf(F_stat,dfx,dfy)
    if alternative == 'less':
        return F_stat,q
    elif alternative == 'greater':
        return F_stat,1-q
    else:
        return F_stat,2*min(q,1-q)
```

Nonparametric (distribution free) statistical tests

Sign tests and rank tests

Parametric tests are usually preferred over non-parametric tests since they usually have more statistical power (= lower type 2 error rate) than non-parametric tests, given that the statistical assumptions are satisfied.

The advantage of sign tests and rank tests is that they do not require strong distributional assumptions. Sign tests and rank tests are suitable for continuous quantitative variables, but can also be used for any ordinal data.

Sign test

One sample sign test

The one sample sign test is applied in similar testing problems as the one sample t-test. However, the sign test requires milder distributional assumptions.

Let $x_1, \ldots, x_n$ be observed values of a continuous random variable x with population median m. Assume that the observed values are i.i.d.

The null hypothesis: $H_0 : m = m_0$.

Possible alternative hypotheses:

$$H_1 : m > m_0 \text{ (one tailed)},$$
$$H_1 : m < m_0 \text{ (one tailed)},$$
$$H_1 : m \neq m_0 \text{ (two tailed)}.$$

One sample sign test

- Calculate the differences $d_i = x_i - m_0$, $i = 1, \ldots, n$.
- The test statistic S is the number of cases where $d_i > 0$. (Alternatively, the number of cases where $d_i < 0$.)
- If the null hypothesis H_0 is true, then the test statistic follows the binomial distribution with parameters n and $1/2$.
- Under H_0, the expected value of the test statistic is $\frac{1}{2}n$ and the variance is $\frac{1}{4}n$.
- Large and small values of the test statistic (compared to the expected value $\frac{1}{2}n$) suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p-value is small enough.

One sample sign test, p-value

The distribution of the test statistic S is tabulated and many softwares give exact p-values of the test.

Let s denote the observed value of the test statistic S. Then the p-value of the test is given as follows:

- If the alternative hypothesis is $H_1 : m > m_0$, then the p-value is $p = \mathbb{P}(S \geq s)$.
- If the alternative hypothesis is $H_1 : m < m_0$, then the p-value is $p = \mathbb{P}(S \leq s)$.
- If the alternative hypothesis is $H_1 : m \neq m_0$, then the p-value is $p = 2\min(\mathbb{P}(S \geq s), \mathbb{P}(S \leq s))$.

Naturally, the probabilities $\mathbb{P}(S \leq s)$ and $\mathbb{P}(S \geq s)$ are calculated under H_0.

Remark. The sign test can also be used for discrete variables as well. Then it is possible that for some of the observations $d_i = x_i - m_0 = 0$. If the number of zeros is small compared to the sample size, these observations can be deleted and the sample size can be modified accordingly. If the number of zeros is large, then the zeros should be dealt with such that they are against rejecting the null hypothesis. For example: consider the two-tailed null hypothesis, 3 negative signs, 15 positive signs and 6 zeros. Now the test should be conducted as if there were 9 negative signs and 15 positive ones.

Python:

```
S_stat = sum(x-m_0>0)
n = sum(i!=0 for i in x - m_0)
#n = len(x) # if the number of zeros in x - m_0 is large
p_value = scipy.stats.binom_test(S_stat,n,p=0.5)
```

If using Scipy 1.12.0 or newer, use `binomtest` instead of `binom_test`

One sample sign test – implementation in Python

```
import numpy as np
from scipy.stats import binom

def signTest_1sample(x,m0,alternative='two-sided'):
    diff = x-m0
    S_stat = sum(diff>0)
    n = sum(i!=0 for i in diff)
    #n = len(x) # if the number of zeros in x − m0 is large
    q = binom.cdf(S_stat,n,0.5)
    q2 = binom.pmf(S_stat,n,0.5)+1-q # (*)
    if alternative == 'less':
        return S_stat,q
    elif alternative == 'greater':
        return S_stat,q2
    else:
        return S_stat,2*min(q,q2)
# Note that in (*), we used P(S≥s) = P(S=s)+1-P(S≤s)
```

Asymptotic one sample sign test

If the sample size is large, then under the null hypothesis H_0, the standardized test statistic $Z = \frac{S - n/2}{\sqrt{n/4}}$ approximately follows the standard normal distribution.

The approximation is usually good enough if $n > 20$. For smaller samples, the test relies on the exact distribution of the test statistic S.

Paired sign test

The paired sign test is applied in similar testing problems as the paired t-test.

- The observations $(x_{i,1}, x_{i,2})$, $i = 1, \ldots, n$, consist of measured pairs of a random variable x.
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \ldots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.

Paired sign test

- General statistical assumptions: the differences d_i are i.i.d. and follow a distribution with median m.
- The null hypothesis $H_0 : m = 0$.
- Possible alternative hypotheses: $H_1 : m > 0$ (one tailed), $H_1 : m < 0$ (one tailed) or $H_1 : m \neq 0$ (two tailed).
- Now it is possible to apply the one sample sign test for the differences d_i.

Paired sign test – implementation in Python

```
def signTest_paired(x,y,alternative='two-sided'):
    return signTest_1sample(x-y,0,alternative)
```

Numerical example

An imaginary medical study was conducted to examine the effect of medicine *a* in lowering beerium levels in plasma. High beerium levels in plasma are related to several diseases. Beerium levels were measured at the beginning of the study and again 8 weeks after the treatment. We wish to study, whether the medicine had the desired effect on 5% significance level.

Data

Patient	Level		Difference
	Before	After	
1	1384	1332	-52
2	1640	1564	-76
3	1122	1100	-22
4	1272	1260	-12
5	1380	1360	-20
6	624	1624	1000
7	360	1821	1461
8	456	450	-6
9	1726	1712	-14
10	332	821	489
11	1342	1338	-4
12	1630	1626	-4
13	1170	1160	-10

Table: Beerium levels ($\mu g/1000ml$) before and after treatment.

t-test

One sample t-test
Data: differences
$t = 1.5646$, $df = 12$, p-value=0.9282
Alternative hypothesis: true mean is less than 0
Sample estimates: mean of $x = 210$.

Sign test

One sample sign test
Data: differences
$s = 3$, p-value=0.04614
Alternative hypothesis: true median is less than 0
Sample estimates: median of $x = -10$.

Compare the results given by the two tests. Neither one of the tests alone gives a clear view on how the medicine *a* affects. Why? Based on this sample, how does the medicine seem to affect? Is there anything suspicious in the testing set up? Is it OK to use one sided alternative hypothesis here?

Wilcoxon signed rank test

The one sample Wilcoxon signed rank test is applied in similar testing problems as the one sample t-test. However, the one sample Wilcoxon signed rank test requires milder distributional assumptions.

Let $x_1, \ldots, x_n$ be observed values of a continuous symmetric random variable x with population median m. Assume that the observed values are i.i.d.

The null hypothesis $H_0 : m = m_0$.

Possible alternative hypotheses:

$$H_1 : m > m_0 \text{ (one tailed)},$$
$$H_1 : m < m_0 \text{ (one tailed)},$$
$$H_1 : m \neq m_0 \text{ (two tailed)}.$$

One sample Wilcoxon signed rank test

- Calculate the absolute values of the differences $|d_i| = |x_i - m_0|$ for $i = 1, \ldots, n$. Order the absolute values from the smallest to the largest. Define the signed ranks $R_\star(x_i)$ such that $R_\star(x_i)$ is the rank of the absolute value $|d_i| = |x_i - m_0|$ multiplied with the sign of the difference $x_i - m_0$.
- The test statistic $W_\star = \sum_{R_\star(x_i)>0} R_\star(x_i)$ is the sum of the positive ranks. (Alternatively, the sum of the negative ranks.)
- Under H_0, the expected value of the test statistic is $\frac{n(n+1)}{4}$ and the variance is $\frac{n(n+1)(2n+1)}{24}$.
- Large and small values (compared to the expected value $\frac{n(n+1)}{4}$) if the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p-value is small enough.
- Python:

  `_, p_value = scipy.stats.wilcoxon(x-m0)`

One sample Wilcoxon signed rank test, p-value

The distribution of the test statistic $W_\star$ is tabulated and many softwares give exact p-values of the test.

The p-value of the Wilcoxon signed rank test, where $w_\star$ is the observed value of the test statistic $W_\star$, is given as follows:

- If the alternative hypothesis is $H_1 : m > m_0$, then the p-value is $p = \mathbb{P}(W_\star \geq w_\star)$.
- If the alternative hypothesis is $H_1 : m < m_0$, then the p-value is $p = \mathbb{P}(W_\star \leq w_\star)$.
- If the alternative hypothesis is $H_1 : m \neq m_0$, then the p-value is $p = 2\min(\mathbb{P}(W_\star \geq w_\star), \mathbb{P}(W_\star \leq w_\star))$.

The probabilities $\mathbb{P}(W_\star \geq w_\star)$ and $\mathbb{P}(W_\star \leq w_\star)$ are calculated under the null H_0.

Asymptotic one sample Wilcoxon signed rank test

Under H_0, when the sample size is large, the standardized test statistic $Z = \frac{W_\star - \mathbb{E}[W_\star]}{\sqrt{\mathrm{Var}(W_\star)}}$, where $\mathbb{E}[W_\star] = \frac{n(n+1)}{4}$ and $\mathrm{Var}(W_\star) = \frac{n(n+1)(2n+1)}{24}$, approximately follows the standard normal distribution.

The approximation is usually good enough if $n > 20$. For smaller samples, the exact distribution of $W_\star$ is needed.

One sample Wilcoxon signed rank test

We assumed above that the observations come from a continuous distribution. The Wilcoxon signed rank test can be applied for discrete observations as well. However, it is then possible that some points share the same rank of absolute values $|x_i - m_0|$. In that case, all these points are assigned to have the median of the corresponding ranks. For example, if two sample points have the same rank, corresponding to ranks 7 and 8, then both points are assigned to have rank 7.5. If three sample points have the same rank corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Paired Wilcoxon signed rank test

The paired Wilcoxon signed rank test is applied in similar testing problems as the paired t-test.

- The observations $(x_{i,1}, x_{i,2})$, $i = 1, \ldots, n$, consist of measured pairs of a random variable x.
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \ldots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.

Paired Wilcoxon signed rank test

- General statistical assumptions: the differences d_i are i.i.d. and follow a symmetric distribution with median m.
- The null hypothesis is $H_0 : m = 0$.
- Possible alternative hypotheses: $H_1 : m > 0$ (one tailed), $H_1 : m < 0$ (one tailed) or $H_1 : m \neq 0$ (two tailed).
- Now it is possible to apply the one sample Wilcoxon signed rank test for the differences d_i.
- Python:
  `_, p_value = scipy.stats.wilcoxon(x,y)`

Numerical example

We want to compare the prices of Brand X and Brand Y cookies in different stores. The distribution of the prices is not known, but it can be assumed to be symmetrical. 10 different stores were selected randomly for this study. The cookie prices have been tabulated below.

Brand X	4.56	4.67	4.28	4.57	4.78	4.54	4.56	4.48	4.47	4.50
Brand Y	4.52	4.48	4.51	4.30	4.59	4.67	4.53	4.54	4.71	4.49
Difference	**0.04**	**0.19**	**-0.23**	**0.27**	**0.19**	**-0.13**	**0.03**	**-0.06**	**-0.24**	**0.01**

Table: Prices of Brand X and Brand Y cookie packages in different stores.

Numerical example

The price differences are assumed to be symmetrically distributed. The null hypothesis is that the theoretical medians of the prices of Brand X and Brand Y cookies do not differ, i.e., the difference of the population medians is zero. The ordered absolute values of the differences and the corresponding signed ranks are as follows.

Difference	**0.01**	**0.03**	**0.04**	**0.06**	**0.13**	**0.19**	**0.19**	**0.23**	**0.24**	**0.27**
Signed rank	1	2	3	-4	-5	6.5	6.5	-8	-9	10

Table: The ordered absolute values of the differences and the corresponding signed ranks.

The test statistic

$$W_\star = \sum_{R_\star(d_i)>0} R_\star(x_i) = 1 + 2 + 3 + 6.5 + 6.5 + 10 = 29.$$

The p-value (obtained using statistical software) is 0.9219. We do not reject the null hypothesis.

Signed test vs. Wilcoxon signed rank test

- Both tests are suitable for similar problems: one sample – comparison of the median to a constant, paired samples – comparison of the medians.
- The tests are non-parametric counterparts of the one sample t-test.
- The values of the test statistic do not depend on the numerical values of the observations – only the order of the observations matters.
- No assumption of the type of the population distribution is needed for the sign test. Symmetry assumption is required for the Wilcoxon signed rank test.
- The Wilcoxon signed rank test uses more information of the order of the observations.
- If the distribution can be assumed to be symmetric, use the Wilcoxon signed rank test. Otherwise, apply the sign test.

Two sample Wilcoxon rank test

The two sample Wilcoxon rank test is used in similar settings as the two sample t-test, but Wilcoxon rank test requires milder assumptions.

In practice, the two sample Wilcoxon rank test is exactly the same test statistic as the Mann-Whitney test – both names are used in the literature.

Let $x_1, \ldots, x_n$ be the observed values of a continuous random variable x and let $y_1, \ldots, y_m$ be the observed values of a continuous random variable y. Assume that the observations $x_1, \ldots, x_n$ are i.i.d. and assume that $y_1, \ldots, y_m$ are i.i.d. as well. Assume also that x_i and y_j are independent for all i, j. Assume that x is distributed as y up to a location shift (i.e., x and y follow otherwise the same distribution, but possibly with different medians) and assume that the variables have population medians m_x and m_y, respectively.
The null hypothesis $H_0 : m_x = m_y$.
Possible alternative hypotheses: $H_1 : m_x > m_y$ (one tailed), $H_1 : m_x < m_y$ (one tailed) or $H_1 : m_x \neq m_y$.

Two sample Wilcoxon rank test

Consider the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$. Assume (without loss of generality) that $n \leq m$.

The two sample Wilcoxon rank test is based in analyzing the order of all the observations. Combine the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ to one sample $z_1, \ldots, z_{n+m}$. Order the observations z_i from the smallest to the largest. Let $R(z_i)$ be the rank of z_i in the combined sample $z_1, \ldots, z_{n+m}$.

- The test statistic $W = \sum_{i=1}^{n} R(x_i)$ is the sum of the ranks of the smaller sample.
- Under H_0, the expected value of the test statistic is $n(n+m+1)/2$ and the variance is $nm(n+m+1)/12$.
- Large and small values of the test statistic (compared to the expected value $n(n+m+1)/2$) suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- Python:

```
_, p_value = scipy.stats.mannwhitneyu(x,y)
```

Two sample Wilcoxon rank test, p-value

The distribution of the test statistic W is tabulated and many softwares give the exact p-values.

The p-value of the two sample Wilcoxon rank test, where w is the observed value of the test statistic W, is defined as follows:

- If the alternative hypothesis is $H_1 : m_x > m_y$, then the p-value is $p = \mathbb{P}(W \geq w)$.
- If the alternative hypothesis is $H_1 : m_x < m_y$, then the p-value is $p = \mathbb{P}(W \leq w)$.
- If the alternative hypothesis is $H_1 : m_x \neq m_y$, then the p-value is $p = 2\min(\mathbb{P}(W \geq w), \mathbb{P}(W \leq w))$.

Naturally, $\mathbb{P}(W \geq w)$ and $\mathbb{P}(W \leq w)$ are calculated under H_0.

Asymptotic two sample Wilcoxon rank test

Assuming that the null hypothesis is true, if the sample size is large, the standardized test statistic $z = \frac{W - \mathbb{E}[W]}{\sqrt{\mathrm{Var}(W)}}$, where $\mathbb{E}[W] = n(n+m+1)/2$ and $\mathrm{Var}(W) = nm(n+m+1)/12$, approximately follows the standard normal distribution.

The approximation is usually good enough if $n, m > 10$. For smaller samples, the exact distribution of the test statistic W is needed.

Two sample Wilcoxon rank test

The Wilcoxon rank test can be used also when the observations are discrete. Then it is possible that some of the sample points have the same rank. In that case, all those points are assigned to have the median of the corresponding ranks. For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3,5, and 5, then each is assigned to have rank 4.

Note that ranks can be used even when the variables cannot be measured numerically, but they can be ordered. (For example, one could order/rank singers, or qualities of apartments, without measuring them numerically.)

Two sample Wilcoxon rank test

- The two sample Wilcoxon rank test is the non-parametric counterpart of the two sample t-test.
- The value of the test statistic depends on the order/rank of the observed value, not on the exact numerical values of the observations.
- The test is an excellent alternative to two sample t-test, when the populations are not normally distributed.

Numerical example

The height of 10 randomly chosen students was measured in the corridor of the Department of Mathematics. The students were put to stand in line from the shortest to the tallest. There were both, male and female students, in the sample. We wish to know if there is a difference in the distribution of male and female students. The null hypothesis is that the population median of the heights of the female students is equal to the population median of the heights of the male students.

The following table displays the gender and rank of the height of the students.

Student	F	F	M	M	M	F	M	M	F	M
Rank	1	2	3	4	5	6	7	8	9	10

Table: Female and male students ordered according to the rank of their height.

The test statistic

$$W = \sum_{i=1}^{4} R(x_i) = 1 + 2 + 6 + 9 = 18$$

is the sum of the ranks of the smaller, female, sample. We decide to use the two-tailed alternative hypothesis (why?) and significance level 0.05. Since the samples are small, we take the critical values of the test statistic from tabulated values. The critical values are 12 and 32. Since $12 < 18 < 32$, we do not reject the null hypothesis.