Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eighth lecture, December 11, 2023

Proportion test

Proportion test

Proportion tests can be used for example when testing proportions of faulty products in a production process.

Let $x_1, \ldots, x_n$ be the observed values of a random variable x. Assume that the observed values are i.i.d. and come from the Bernoulli distribution with parameter p.[†]

The null hypothesis: $H_0 \colon p = p_0$.

Possible alternative hypotheses:

$$\begin{aligned} H_1 &\colon p > p_0 \text{ (one tailed)}, \\ H_1 &\colon p < p_0 \text{ (one tailed)}, \\ H_1 &\colon p \neq p_0 \text{ (two tailed)}. \end{aligned}$$

[†] Now $\mathbb{P}(x_i = 1) = p$, $\mathbb{P}(x_i = 0) = 1 - p$, $\mathbb{E}[p] = p$, and $\mathrm{Var}(x) = p(1-p)$.

Proportion test

- The test statistic $C = \sum_{i=1}^{n} x_i$.
- If the null hypothesis H_0 is true, then the test statistic follows the binomial distribution with parameters n and $p = p_0$.
- Under the null hypothesis H_0, the expected value of the test statistic is np_0 ($\mathbb{E}[C] = np_0$) and the variance of the test statistic is $np_0(1 - p_0)$.
- If the value of the test statistic is large or small compared to the expected value np_0, evidence against the null hypothesis is found.
- The null hypothesis is rejected if the p-value is small enough.
- Python:

```
C_stat = sum(x) # x is a (0,1) vector of length n
                # containing the outcomes of
                # Bernoulli trials
p_value = scipy.stats.binom_test(C_stat,n,p=p0)
```

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

Proportion test, p-value

The distribution of the test statistic C is tabulated and statistical softwares calculate the p-values of the test.

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

- If the alternative hypothesis is $H_1 : p > p_0$, then the p-value is $p = \mathbb{P}(C \geq c)$.
- If the alternative hypothesis is $H_1 : p < p_0$, then the p-value is $p = \mathbb{P}(C \leq c)$.
- If the alternative hypothesis is $H_1 : p \neq p_0$, then the p-value is *usually*[†] defined in the literature as $p = 2\min(\mathbb{P}(C \geq c), \mathbb{P}(C \leq c))$.

The probabilities $\mathbb{P}(C \geq c)$ and $\mathbb{P}(C \leq c)$ are calculated under H_0.

[†] Statistical software such as R or the Scipy library use instead the formula $p = \sum_{k : p_C(k) \leq p_C(c)} p_C(k)$ for the two tailed p-value, where p_C denotes the PMF of $\mathrm{Bin}(n, p_0)$.

Asymptotic proportion test

If the sample size is large, then under the null hypothesis H_0, the standardized test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

where $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the unbiased estimator of the parameter p, approximately follows the standard normal distribution.

The approximation is usually good enough if $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$. For smaller samples, the test relies on the exact distribution of the test statistic.

Numerical example

In anticipation of an upcoming election, an opinion poll was conducted. In the poll, the sample size was 1000 and 420 out of the 1000 eligible voters reported that they support the mayor. We want to test on significance level 5% whether the true support is less than 50% of the population.

Null hypothesis: $H_0 \colon p = 0.5$.

Alternative hypothesis: $H_1 \colon p < 0.5$.

Since $n = 1000$ and $\hat{p} = \frac{420}{1000} = 0.42$ satisfy $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$, we can use normal approximation. The observed value of the Z-statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/1000}} = \frac{0.42 - 0.50}{\sqrt{0.5^2/1000}} \approx -5.06.$$

The p-value is $p = \mathbb{P}(Z \le z) = \Phi(-5.06) \approx 2.10 \cdot 10^{-7}$. H_0 is rejected.
Python:

```
>>>scipy.stats.binom_test(420,1000,p=0.5,alternative='less')
2.348554631632085e-07 # exact binomial test
>>>scipy.stats.norm.cdf((0.42-0.50)/numpy.sqrt(0.5*0.5/1000))
2.1001969880109918e-07 # normal approximation
```

Two sample proportion test

In the two sample proportion test, parameters of two different Bernoulli distributed samples are compared.

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 Bernoulli distribution with parameter p_x, and assume that the observed values $y_1, \ldots, y_m$ are i.i.d. and come from the Bernoulli distribution with parameter p_y. Furthermore, assume that x_i and y_j are independent for all i, j.

The null hypothesis: $H_0\colon p_x = p_y$.

Possible alternative hypotheses:

$$\begin{aligned} H_1 &\colon p_x > p_y \text{ (one tailed)},\\ H_1 &\colon p_x < p_y \text{ (one tailed)},\\ H_1 &\colon p_x \neq p_y \text{ (two tailed)}, \end{aligned}$$

Two sample proportion test

- Calculate the sample proportions $\hat{p}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{p}_y = \frac{1}{m}\sum_{i=1}^{m} y_i$, and $\hat{p} = \frac{n\hat{p}_x + m\hat{p}_y}{n+m}$.
- Calculate the test statistic

$$Z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n}+\frac{1}{m})}}.$$

- If the sample size is large, then under the null hypothesis H_0, the test statistic Z approximately follows the standard normal distribution. The approximation is usually good enough if $n\hat{p}_x > 5$, $n(1-\hat{p}_x) > 5$, $m\hat{p}_y > 5$, and $m(1-\hat{p}_y) > 5$.
- If the value of the test statistic has large absolute value, then evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p-value is small enough.

Testing general statistical assumptions

In statistics, we very often make assumptions about the underlying distribution. Most statistical methods become ineffective or give false results if these assumptions do not hold. Hence it is very important to test the distributional assumptions separately.

Testing normality

Normality assumption

The normal distribution has a central role in statistics. Multiple methods for testing the normality of observations have been developed. Here, we take a look at a couple of them.

In what follows, let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x.

The null hypothesis is H_0: "the random variable x is normally distributed."

The alternative hypothesis is H_1: "the random variable x is not normally distributed."

The Bowman and Shenton normality test

The Bowman and Shenton normality test is a function of skewness and kurtosis:

$$BS = n\left(\frac{v^2}{6} + \frac{k^2}{24}\right),$$

where v is the sample skewness coefficient and k is the sample kurtosis coefficient.

If the skewness or kurtosis differ a lot from the skewness and/or kurtosis of the normal distribution, the test statistic gets large values.

Bowman and Shenton normality test

- If n is large, then under the null hypothesis H_0, the test statistic BS follows approximately the $\chi^2(2)$ distribution.
- The expected value of the test statistic under the null hypothesis H_0 is $\mathbb{E}[BS] = 2$.
- Large values of the test statistic compared to the expected value suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p-value is small enough.
- If one uses the formulae $\hat{v} = \frac{m_3}{\hat{s}^3}$ and $\hat{k} = \frac{m_4}{\hat{s}^4} - 3$, where $\hat{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is the *biased* sample standard deviation, then one obtains the closely related Jarque–Bera test statistic

  $$JB = n\left(\frac{\hat{v}^2}{6} + \frac{\hat{k}^2}{24}\right),$$

  also used to assess normality. This test statistic is implemented in the Python Scipy library as `scipy.stats.jarque_bera`
- Note that the Bowman and Shenton (resp. Jarque–Bera) normality test is suitable for large samples only!

Implementation using Python

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

def BowmanShentonTest(x):
    n = len(x)
    xbar = np.mean(x)
    std = np.std(x,ddof=1)
    v = (1/n)*sum((x-xbar)**3)/std**3
    k = (1/n)*sum((x-xbar)**4)/std**4-3
    BS = n*(v**2/6+k**2/24)
    q = chi2.cdf(BS,2)
    return BS,1-q # corrected 19.12.2023
# Note: if the distribution has 0 skewness and 0 kurtosis
# (ideal case for the normal distribution), then the test
# statistic BS == 0. Thus we choose a one sided alternative
# hypothesis of type 'greater' since only large values of BS
# would be evidence of non-normality.
```

Rank plot

Let $x_1, \dots, x_n$ be i.i.d. observations from some distribution F_x. Let $z_1 \leq \cdots \leq z_n$ be the observations $x_1, \dots, x_n$ ordered from the smallest to the largest one. Let $y_1 \leq \cdots \leq y_n$ be the ordered values of n i.i.d. observations from the standard normal distribution $\mathcal{N}(0, 1)$ and let $\mathbb{E}[y_i]$ be the expected value of y_i.

Plot the pairs $(\mathbb{E}[y_i], z_i)$, $i = 1, \dots, n$. If the x_i come from a normal distribution, then the points $(\mathbb{E}[y_i], z_i)$ should approximately lie on a line. If the points do not lie on a line, the sample differs from the normal distribution. The plot can be used in detecting skewness of the distribution and in finding outliers.

Rank plots are useful for quick visual assessment of the distribution of the data: cf., e.g., the excellent StackExchange post
https://stats.stackexchange.com/a/101290

Shapiro–Wilk normality test

- The Shapiro–Wilk normality test statistic is the squared value of the Pearson sample correlation coefficient calculated from the rank plot points $(\mathbb{E}[y_i], z_i)$, $i = 1, \ldots, n$.
- Small values of the test statistic suggest that the assumption of normality does not hold. Large values of the test statistic are in line with the null hypothesis.
- The null hypothesis is rejected if the p-value is small enough. The test requires a large sample.
- Statistical software can be used to calculate the p-value of the test. Python: `scipy.stats.shapiro(x)`

Numerical example

During the previous lecture, we considered an example where we compared Brand X and Brand Y cookies. In the example, the price differences were assumed to be symmetrically distributed. The data consisted of the cookie prices in 10 randomly selected stores. We now wish to test the normality of the price differences. The price differences are given below.

Difference:	**0.04**	**0.19**	**-0.23**	**0.27**	**0.19**	**-0.13**	**0.03**	**-0.06**	**-0.24**	**0.01**

Table: The differences of Brand X and Brand Y cookie prices.

The Bowman and Shenton test: In order to calculate the test statistic, the sample skewness and kurtosis coefficients v and k are needed. The sample standard deviation is $s \approx 0.176$ and the sample mean is $\overline{x} \approx 0.07$. Now

$$
\begin{aligned}
v &= \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3} \approx -0.0139 \\
k &= \frac{m_4}{s^4} - 3 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4}\right) - 3 \approx -1.506.
\end{aligned}
$$

The value of the test statistic is

$$
BS = n\left(\frac{v^2}{6} + \frac{k^2}{24}\right) \approx 0.945.
$$

Under the null hypothesis, the test statistic follows the $\chi^2(2)$ distribution. We decide to use the significance level 0.05. The critical values are then 0.051 and 7.378. Since $0.051 < 0.945 < 7.378$, evidence of non-normality was not found.

Rank plot (Q-Q plot):

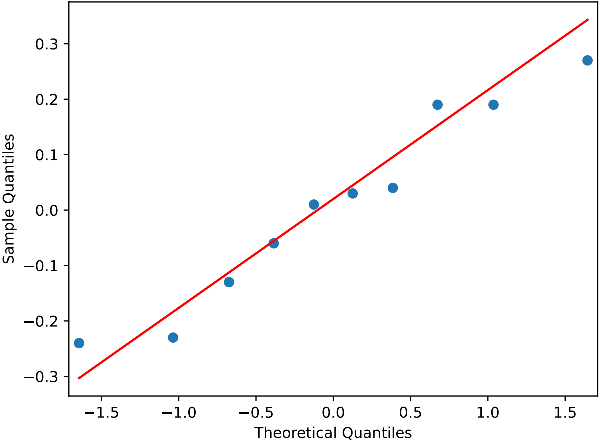

Figure: Rank plot of the price differences.

Shapiro–Wilk test: calculated in Python using the function `scipy.stats.shapiro`

data: differences

$W = 0.9439$, p-value $= 0.5966$

The p-value is large and thus evidence of non-normality was not found.

Can these results be trusted? Were all the required assumptions fulfilled? What was the type 2 error?

χ^2 tests

Multinomial distribution

Consider a situation, where a random experiment has k mutually exclusive outcomes and consider n independent runs of that experiment. The multinomial distribution models the frequency distribution of the outcome of these n independent random experiments.

The random variables $x_1, \ldots, x_k$ follow the multinomial distribution with parameters $n, p_1, \ldots, p_k$, if the probability mass function is

$$p(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where

$$\sum_{i=1}^{k} x_i = n \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1.$$

Assume that $x_1, \ldots, x_k$ follow multinomial distribution with parameters $n, p_1, \ldots, p_k$. If n is large, then

$$\sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$$

approximately follows the $\chi^2(k-1)$ distribution.

χ^2 goodness-of-fit test

The χ^2 goodness-of-fit test examines the discrepancy between observed values and the values expected under some particular distribution of a random variable x.

The null hypothesis H_0: "The random variable x follows distribution F_x (with or without unknown parameters)."

The alternative hypothesis H_1: "The random variable x does not follow distribution F_x (with or without unknown parameters)."

χ^2 goodness-of-fit test

Let $x_1, \ldots, x_n$ be i.i.d. observations of a random variable x.

- Categorize the n observations into k categories.
- Calculate the frequencies O_i, $i = 1, \ldots, k$, where O_i is the observed frequency of the category i. Note that $\sum_{i=1}^{k} O_i = n$.
- Let p_i be the probability that, under the null hypothesis, the random variable x belongs to the category i. Calculate the expected frequencies $E_i = np_i$ of the observations in category i. Note that $\sum_{i=1}^{k} p_i = 1$.
- Now, under the null hypothesis, the random variables $O_1, \ldots, O_k$ follow the multinomial distribution with parameters $n, p_1, \ldots, p_k$.

χ^2 goodness-of-fit test

- Calculate the test statistic
$$\chi_g^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$
- If n is large, then under the null hypothesis, the test statistic χ_g^2 approximately follows $\chi^2(k-1-e)$ distribution, where e is the number of estimated parameters.
- The expected value of the test statistic, under the null hypothesis, is $\mathbb{E}[\chi_g^2] = k - 1 - e$.
- Large values of the test statistic (compared to the expected value) suggest that the null hypothesis H_0 does not hold.
- If the p-value is small enough, then the null hypothesis H_0 is rejected.
- If the value of the test statistic is large, the sample frequencies differ greatly from the expected value and it is clear that the null hypothesis should be rejected. However, if the value is very small, then the sample frequencies differ less than expected. This is called overfitting – usually, we are not concerned about this, so typically a one tailed alternative hypothesis (of type `alternative='greater'`) is used.

Goodness-of-fit test, Example 1

Let us examine the quality of giant mugs made in a ceramics factory. The null hypothesis is that:

- an error in the shape of the mug occurs with probability $2/14$,
- a color error occurs with probability $2/14$,
- both errors occur simultaneously with probability $1/14$,
- the probability of an error-free product is $9/14$.

Consider a sample of 200 randomly selected mugs such that

- 40 mugs have an error in the shape,
- 44 have a color error,
- 26 mugs have both errors,
- 90 mugs are error-free.

Now $O_1 = 40$, $O_2 = 44$, $O_3 = 26$, $O_4 = 90$

$E_1 = 200 \cdot \frac{2}{14}$, $E_2 = 200 \cdot \frac{2}{14}$, $E_3 = 200 \cdot \frac{1}{14}$, $E_4 = 200 \cdot \frac{9}{14}$

$\therefore \chi_g^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 34.08$. Under the null hypothesis, the test statistic approximately follows the $\chi^2(4-1) = \chi^2(3)$ distribution. Since $\mathbb{P}(\chi^2(3) \geq 34.08) < 0.00001$, the null hypothesis is rejected.

Goodness-of-fit test, Example 1

The χ^2 goodness-of-fit test is implemented in the Python Scipy library as `scipy.stats.chisquare`

For example, we can solve the previous example numerically as follows:

```
from scipy.stats import chisquare
O = [40,44,26,90]
E = [200*2/14,200*2/14,200*1/14,200*9/14]
chisquare(O,E)
```

The output is

```
Power_divergenceResult(statistic=34.08,
                       pvalue=1.905621048402571e-07)
```

Goodness-of-fit test, Example 2

Consider testing whether the monthly salary of Germans follows the normal distribution. Select randomly n Germans and document the salaries. The null hypothesis is that the observations come from a normal distribution with an unknown expected value and an unknown variance.

- Estimate the unknown parameters (μ and σ^2) from the sample.
- Discretize the continuous salary variable.
- Calculate the observed category frequencies $O_1, \dots, O_k$, i.e., calculate the number of observations in each category.
- Calculate the category probabilities from the normal distribution. For example,

$$\dots, \mathbb{P}(1900 < X \leq 2000),\ \mathbb{P}(2000 < X \leq 2100), \dots$$

- Calculate the expected category frequencies $E_1, \dots, E_k$.
- Calculate the test statistic. Under the null hypothesis, the test statistic approximately follows the $\chi^2(k-1-2) = \chi^2(k-3)$, where k is the number of the used categories and we estimated $e = 2$ parameters (μ and σ^2). Calculate the p-value and based on that, either reject or do not reject the null hypothesis.

χ^2 homogeneity test

In the χ^2 homogeneity test, several (r) samples are examined.

The null hypothesis H_0: "The samples come from (some) same distribution."

The alternative hypothesis H_1: "The samples do not come from the same distribution."

χ^2 homogeneity test

Consider several (r) independent samples. Assume that the observations of each sample are i.i.d. Assume that the sample i, $i \in \{1, \ldots, r\}$, has n_i observations.

- Categorize all the observations into c categories of size C_j.
- Calculate the frequencies O_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, where O_{ij} is the observed frequency of the observations of the sample i in category j

	1	2	$\cdots$	c	sum
1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
sum	C_1	C_2	$\cdots$	C_c	n

Table: The observed frequencies.

- Let $p_j = C_j/n$. Under the null hypothesis, for each sample i, the probability of the category j is the same p_j.
- Calculate the expected frequencies $E_{ij} = n_i p_j$.

χ^2 homogeneity test

	1	2	$\cdots$	c	sum
1	E_{11}	E_{12}	$\cdots$	E_{1c}	n_1
2	E_{21}	E_{22}	$\cdots$	E_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	E_{r1}	E_{r2}	$\cdots$	E_{rc}	n_r
sum	C_1	C_2	$\cdots$	C_c	n

Table: The expected frequencies.

χ^2 homogeneity test

- Calculate the value of the test statistic

$$\chi_h^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- If n is large, then under the null hypothesis, the test statistic χ_h^2 approximately follows the $\chi^2((r-1)(c-1))$ distribution.
- Under the null hypothesis, the expected value of the test statistic is $(r-1)(c-1)$. (That is, $\mathbb{E}[\chi_h^2] = (r-1)(c-1)$.)
- Large values of the test statistic compared to the expected value suggest that the null hypothesis H_0 is false. Small values of the test statistic compared to the expected value are indicative of overfitting – the data fits the model "too well". Usually, we are not too concerned about this, so typically a one tailed alternative hypothesis (of type `alternative='greater'`) is used.
- The null hypothesis H_0 is rejected if the p-value is small enough.

Homogeneity test, Example

A city council is about to make decisions about building a new library. There was a preliminary plan and 250 randomly selected men and 300 randomly selected women were asked to comment the plan. 169 men and 125 women thought that the plan was good, 52 men and 144 women did not like the plan, and 29 men and 31 women did not have an opinion about the plan.

	good plan	bad plan	no opinion	Total
Men	169	52	29	250
Women	125	144	31	300
Total	294	196	60	550

Table: Observed frequencies

	good plan	bad plan	no opinion	Total
Men	133.6	89.1	27.3	250
Women	160.4	106.9	32.7	300
Total	294	196	60	550

Table: Expected frequencies

Homogeneity test, Example

The value of the test statistic:

$$\chi_h^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 45.7105.$$

Under the null hypothesis, the test statistic approximately follows the $\chi^2((2-1)(3-1)) = \chi^2(2)$ distribution. Since $\mathbb{P}(\chi^2(2) \geq 45.7105) < 0.00001$, it can be concluded that the opinions about the preliminary plan do differ between men and women.

Solution using Python:

```
import pandas as pd
from scipy.stats import chisquare
O =  pd.DataFrame({'good plan':[169,125],'bad plan':
     [52,144],'no opinion':[29,31]},index=['Men','Women'])
tmp = O.values # create expected frequency table
E = pd.DataFrame((tmp.sum(0)*tmp.sum(1)[:,None])/tmp.sum(),
     columns=O.columns,index=O.index)
chisquare(O,E,ddof=(O.shape[0]-1)*(O.shape[1]-1),axis=None)
```

χ^2 test of independence

The χ^2 test of independence is applied to study whether two random variables (factors) are stochastically independent.

Null hypothesis H_0: "the variables are independent."

Alternative hypothesis: H_1: "the variables are not independent."

χ^2 test of independence

Consider a simple random sample of size n. Divide the observations to r classes with respect to a factor A and to c classes with respect to a factor B. Let R_i be the frequency of the observations in class i with respect to the factor A. Let C_j be the frequency of the observations in class j with respect to the factor B. Let O_{ij} be the observed frequency of the observations in class i with respect to the factor A and class j with respect to the factor B.

	1	2	$\cdots$	c	sum
1	O_{11}	O_{12}	$\cdots$	O_{1c}	R_1
2	O_{21}	O_{22}	$\cdots$	O_{2c}	R_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	R_r
sum	C_1	C_2	$\cdots$	C_c	n

Table: The observed frequencies.

- Let $P_j = C_j/n$. Under the null hypothesis, for each category i of the factor A, the probability of category j of the factor B has the same probability P_j.
- Calculate the expected frequencies $E_{ij} = R_i P_j$.

	1	2	$\cdots$	c	sum
1	E_{11}	E_{12}	$\cdots$	E_{1c}	R_1
2	E_{21}	E_{22}	$\cdots$	E_{2c}	R_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	E_{r1}	E_{r2}	$\cdots$	E_{rc}	R_r
sum	C_1	C_2	$\cdots$	E_c	n

Table: The expected frequencies.

χ^2 test of independence

- Calculate the value of the test statistic

$$\chi_I^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- If n is large, then under the null hypothesis, the test statistic approximately follows the $\chi^2((r-1)(c-1))$ distribution.
- The expected value of the test statistic is $(r-1)(c-1)$. That is, $\mathbb{E}[\chi_I^2] = (r-1)(c-1)$.
- Large values (compared to the expected values) of the test statistic suggest that the null hypothesis is false. Small values of the test statistic compared to the expected value are indicative of overfitting – the data fits the model "too well". Usually, we are not too concerned about this, so typically a one tailed alternative hypothesis (of type `alternative='greater'`) is used.
- The null hypothesis is rejected if the p-value is small enough.

Test of independence, Example

There was an interesting presidential election and we wish to examine the independence of the voting behavior of married men (M) and women (W). The sample consists of 120 married couples and the presidential candidates were A, B, C. In total, there are nine categories: AA, AB, AC, BA, BB, BC, CA, CB, CC.

	A, man	B, man	C, man	Total
A, woman	15	7	8	30
B, woman	20	25	5	50
C, woman	10	10	20	40
Total	45	42	33	120

Table: Observed frequencies

	A, man	B, man	C, man	Total
A, woman	11.25	10.50	8.25	30
B, woman	18.75	17.50	13.75	50
C, woman	15.00	14.00	11.00	40
Total	45	42	33	120

Table: Expected frequencies

The value of the test statistic

$$\chi_r^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 21.46.$$

Under the null hypothesis, the test statistic approximately follows the $\chi^2((3-1)(3-1)) = \chi^2(4)$ distribution. Since $\mathbb{P}(\chi^2(4) \geq 21.46) = 0.000257$, we conclude that the voting behavior of married men and women is not independent.

Solution using Python:

```
import pandas as pd
from scipy.stats import chisquare
O = pd.DataFrame({'A, man': [15,20,10],'B, man': [7,25,10],
'C, man': [8,5,20]},index=['A, woman','B, woman','C, woman'])
tmp = O.values # create expected frequency table
E = pd.DataFrame((tmp.sum(0)*tmp.sum(1)[:,None])/tmp.sum(),
    columns=O.columns,index=O.index)
chisquare(O,E,ddof=(O.shape[0]-1)*(O.shape[1]-1),axis=None)
```

Remark. The χ^2 test of independence and the χ^2 homogeneity test are very similar. The test statistics and the degrees of freedom are calculated identically. However, the tests measure very different phenomena.