10. Non-parametric hypothesis testing

Goodness-of-fit test

The hypothesis involve the *type of distribution*. The null hypothesis H_0 states that the random variable X has a probability distribution $f_0(x)$, and the alternative hypothesis states that it is not: $H_0\big(f(x) = f_0(x)\big)$ and $H_1\big(f(x) \neq f_0(x)\big)$.

H_0 is tested based on a random sample $\boldsymbol{v} = (x_1, x_2, \cdots, x_n)$. The sample value span is divided into r bins or *class intervals* and the frequencies n_i of the sample measurements that correspond to each class interval are determined. H_0 is then tested by comparing the sample frequencies n_i with the assumed frequencies $n_{i_0} = p_{i_0} n$, which are determined based on the distribution, assumed by H_0. To determine the assumed probabilities p_{i_0}, the necessary parameters of the distribution $f_0(x)$ are estimated by point estimators from the given sample. The comparison is realized by a test statistics, formed as a weighted sum of squares of differences of the relative frequencies:

$$\chi^2 = \sum_{i=1}^{r} \left(\frac{n_i - n_{i_0}}{n_{i_0}} \right)^2 n_{i_0} = \sum_{i=1}^{r} \frac{(n_i - n_{i_0})^2}{n_{i_0}} = n \sum_{i=1}^{r} \frac{(p_i - p_{i_0})^2}{p_{i_0}} = \left(\sum_{i=1}^{r} \frac{n_i^2}{n_{i_0}} \right) - n.$$

The distribution of the test statistics χ^2 asymptotically approaches the χ^2_{r-l-1} distribution, where l is the number of the assumed distribution parameters which had to be estimated from the sample to determine the frequencies n_{i_0}. For normal distribution $l = 2$, for exponential and Poisson $l = 1$, and for uniform distribution $l = 0$. For a large n and $n_i \geq 5$ for each class interval, the distribution of the test statistics χ^2 is very close to χ^2_{r-l-1} distribution. In case the test statistics value exceeds the critical value $\chi^2_{r-l-1;\,\alpha}$, the H_0 is rejected. To facilitate the determination of the test statistics, a table with columns x_i, n_i, n_i^2, $n_{i_0} = p_{i_0} n$ and n_i^2 / n_{i_0} is constructed. The value of the test statistics equals the sum of the values in the last column minus n.

Independence test

The hypothesis involve *(in)dependence of two random variables* or influences X and Y whose values can be divided into r and c class intervals, respectively. Based on a sample of n observations, frequencies n_{ij} are determined for all class interval pairs (x_i, y_j) and gathered in a **contingency table** where sign $*$ in place of an index denotes a sum according to that index: $n_{i*} = \sum_{j=1}^{c} n_{ij}$ or $n_{*j} = \sum_{i=1}^{r} n_{ij}$:

		Y				
		y_1	y_2	$\cdots$	y_c	n_{i*}
X	x_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1*}
	x_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2*}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r*}
	n_{*j}	n_{*1}	n_{*2}	$\cdots$	n_{*c}	$n_{**} = n$

The null hypothesis states the variables X and Y are independent while the alternative hypothesis states they are not:

$$H_0\left(p_{ij} = p_i \cdot p_j, \text{for each pair } (i,j)\right), \qquad H_1\left(p_{ij} \neq p_i \cdot p_j, \text{for at least one pair } (i,j)\right).$$

The assumed joint probabilities p_{ij_0} are determined based on H_0 as products of the marginal probabilities which are estimated by marginal relative frequencies: $p_{ij_0} = p_{i_0}p_{j_0} = n_{i*}n_{*j}/n^2$. The null hypothesis is tested using the following test statistics:

$$\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(p_{ij} - p_{ij_0})^2}{p_{ij_0}} = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n_{i*}n_{*j}/n)^2}{n_{i*}n_{*j}} = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right).$$

The test statistics χ^2 is $\chi^2_{(r-1)(c-1)}$ distributed. If the test statistics value exceeds the critical value $\chi^2_{(r-1)(c-1);\,\alpha}$, the H_0 is rejected.

Homogeneity test

The hypothesis involve *(in)homogeneity of v populations* (groups) with respect to some criteria having r possible values. For each of the populations we have a sample of n_i observations, which are arranged into r class intervals according to the criteria. This way, the frequencies n_{ij} are determined and gathered in a **contingency table** which is similar to the independence test table only that in the homogeneity case the marginal frequencies $n_i = n_{i*}$ are usually defined in advance:

		Criteria				
		y_1	y_2	$\cdots$	y_r	$n_{i*} = n_i$
Groups	x_1	n_{11}	n_{12}	$\cdots$	n_{1r}	$n_{1*} = n_1$
	x_2	n_{21}	n_{22}	$\cdots$	n_{2r}	$n_{2*} = n_2$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_v	n_{v1}	n_{v2}	$\cdots$	n_{vr}	$n_{v*} = n_v$
	n_{*j}	n_{*1}	n_{*2}	$\cdots$	n_{*r}	$n_{**} = n$

The null hypothesis states that the populations are homogeneous with respect to the criteria while the alternative hypothesis states that they are not:

$$H_0\left(p_{1j} = p_{2j} = \cdots = p_{vj} = p_{*j}, \text{for each } j\right), \qquad H_1\left(p_{ij} \neq p_{kj}, \text{for at least one triple } (i,j,k)\right).$$

The assumed probability for each value of the criteria is estimated by $p_{*j_0} = n_{*j}/n$. The assumed probabilities p_{ij_0} are estimated by $p_{ij_0} = n_i p_{*j_0}/n = n_{i*}n_{*j}/n^2$. Since the estimator of p_{ij_0} is the same as in the independency test, the test statistics is also the same:

$$\chi^2 = n\sum_{i=1}^{v}\sum_{j=1}^{r}\frac{(p_{ij} - p_{ij_0})^2}{p_{ij_0}} = n\sum_{i=1}^{v}\sum_{j=1}^{r}\frac{(n_{ij} - n_{i*}n_{*j}/n)^2}{n_{i*}n_{*j}} = n\left(\sum_{i=1}^{v}\sum_{j=1}^{r}\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right).$$

The test statistics is $\chi^2_{(v-1)(r-1)}$ distributed. If the test statistics value exceeds the critical value $\chi^2_{(v-1)(r-1);\,\alpha}$, the H_0 is rejected.

10. Non-parametric hypothesis testing – problems

1. In the research of the grinding wheel wear we are interested in the volume of the removed material before the grinding wheel is worn down. The measurements are gathered in the table. Can we claim that the volume of the removed material is normally distributed?
   R: Yes. $\chi^2 = 4.81$, $p = 0.09$

Volume [cm^3]	[5, 8)	[8, 10)	[10, 11)	[11, 13)	[13, 16]
Nr. of pieces	8	17	18	15	7

2. The distribution of the highway accidents is studied. The table shows the number of accidents for different sections of a highway. Can we claim that the number of accidents is uniformly distributed? R: Yes. $\chi^2 = 2.34$, $p = 0.67$

Section [km]	[0, 20)	[20, 35)	[35, 45)	[45, 60)	[60, 80]
Nr. of accidents	21	13	15	17	24

3. The table shows frequencies of the daily average wind velocities at the Brnik airport. Can we claim that the daily average wind velocity is exponentially distributed? R: Yes. $\chi^2 = 2.03$, $p = 0.57$

Wind velocity [m/s]	[0, 1)	[1, 2)	[2, 3)	[3, 4)	[4, 6]
Frequency [day]	58	23	11	6	5

4. In a repair shop the frequency of failures for different types of products is monitored. A sample of 130 products with failures is sorted with respect to the type of product and the type of failure. The data is gathered in the table. Can we assume that the type of failure is independent of the type of product? R: No. $\chi^2 = 13.16$, $p = 0.011$

		Failure		
		1	2	3
roduct	A	22	11	8
	B	12	14	7
	C	16	14	26

5. In clinical tests of medication A we studied whether patients treated with the medication A recover quicker than patients who did not receive this medication. The results of the study are gathered in the table. Can we claim that the medication A significantly affects the time of recovery?
   R: Yes. $\chi^2 = 11.61$, $p = 6.6 \cdot 10^{-4}$

	Recovers quicker	Does not recover quicker
Medication	85	15
Placebo	64	36

10. Non-parametric hypothesis testing – additional problems

1. The diameters of 80 spruce trees in the forest parcel were measured. Could the distribution of the spruce diameters be described by a normal distribution? R: Yes. $\chi^2 = 4.50$, $p = 0.105$

Spruce diameter [cm]	[8, 12)	[12, 16)	[16, 20)	[20, 24)	[24, 28]
Nr. of trees	7	14	37	16	6

2. In the game of Loto the winning numbers between 1 and 39 are drawn. The table shows the frequency of the drawn numbers by decades in the past 12 years. Can we claim that the numbers drawn in that period are evenly distributed? R: Yes. $\chi^2 = 2.15$, $p = 0.541$

Decade	[1, 9]	[10, 19]	[20, 29]	[30, 39]
Nr. of draws	1019	1200	1147	1154

3. The tubes are welded longitudinally. In the quality check, the length of pipe between two defects of the weld is measured. The data is gathered in the table. Could the distribution of pipe length without defects be described by an exponential distribution? R: No. $\chi^2 = 11.62$, $p = 0.0088$

Pipe length [cm]	[0, 100)	[100, 200)	[200, 300)	[300, 400)	[400, 600]
Nr. of products	40	32	25	13	10

4. In a telephone survey citizens were asked about their support of a certain governmental decision. In the survey also the level of respondent education was recorded. The data is gathered in the table. Can we claim that the support for the governmental decision is independent on the level of education? R: Yes. $\chi^2 = 0.78$, $p = 0.677$

		Support	
		Yes	No
Ed. level	Primary	33	26
	Secondary	144	107
	University	101	89

5. In the process of checking the work of physicians it was also determined how many patients a physician has advised on stopping smoking. For each of the three age groups of physicians a sample of patients was selected and sorted according to the advice of their physician. The data is gathered in the table. Can we claim that doctors of those age groups similarly advise their patients in relation to smoking? R: Yes. $\chi^2 = 0.13$, $p = 0.937$

		Advice	
		Yes	No
Age	<30	88	128
	30–50	28	37
	>50	12	18