Non-parametric hypothesis testing

Goodness-of-fit test: we do not know the distribution of population and the hypothesis involve the *type of distribution*. The null hypothesis states that the random variable X has a probability distribution $f_0(x)$ and the alternative hypothesis states that it is not:

$$\begin{aligned} &\mathrm{H}_0 : f(x) = f_0(x)\,, \\ &\mathrm{H}_1 : f(x) \neq f_0(x)\,. \end{aligned} \tag{1}$$

The test procedure requires a random sample of n observations from the population. The observations are arranged in a frequency histogram, having r bins or class intervals with bin frequencies n_i. The probabilities p_i of X falling in i-th bin can be estimated by $p_i = n_i/n$. The frequencies n_i are then compared to the expected bin frequencies $n_{i_0} = p_{i_0} n$ which are calculated from the hypothesized probability distribution. The comparison is done using the following test statistics:

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

The distribution of the test statistics χ^2 is asimptotically approaching the χ^2_{r-l-1} distribution, where l is number of hypothesized distribution parameters which were estimated from the sample in order to calculate the frequencies n_{i_0}. For normal distribution $l = 2$, for exponential and Poisson $l = 1$, and for uniform distribution $l = 0$. If n is large 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. If the test statistics value exceeds the critical value $\chi^2_{r-l-1;\alpha}$, the H_0 is rejected. To facilitate the calculation of test statistics value 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 test statistics value equals the sum of the values in the last column.

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. From the sample of n observations, frequencies n_{ij} are determined for all of the class intervals pairs (x_i, y_j). The determined frequencies are gathered in a $r \times c$ contingency table:

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

The sign $\star$ on the place of an index means a sum according to that index:

$$n_{i\star} = \sum_{j=1}^{c} n_{ij} \qquad \text{and} \qquad n_{\star j} = \sum_{i=1}^{r} n_{ij}\,. \tag{3}$$

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

$$\begin{aligned} &\mathrm{H}_0 : p_{ij} = p_i \cdot p_j\,, \qquad \text{for any pair } (i,j) \\ &\mathrm{H}_1 : p_{ij} \neq p_i \cdot p_j\,, \qquad \text{for at least one pair } (i,j)\,. \end{aligned} \tag{4}$$

The expected joint probabilities p_{ij_0} are calculated on the basis of H_0 as products of the corresponding marginal probabilities: $p_{ij_0} = p_{i_0} p_{j_0} = n_{i\star} n_{\star j}/n^2$. The null hypothesis is tested using the following test statistics:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(p_{ij} - p_{ij_0})^2}{p_{ij_0}} = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n_{i\star} n_{\star j}/n)^2}{n_{i\star} n_{\star j}} = n \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i\star} n_{\star j}} - 1 \right) , \qquad (5)$$

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* that has r different 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. They are gathered in a $v \times r$ contingency table:

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

The table is similar to the independence test contingency table, only that the frequencies n_i are determined before the test by sizes of the populations samples. The null hypothesis states that the populations are homogeneous with respect to the criteria, the alternative hypothesis states that they are not:

$$\begin{aligned} \mathrm{H}_0 &: p_{1j} = p_{2j} = \cdots = p_{vj} = p_{\star j} , \qquad \text{for any } j \\ \mathrm{H}_1 &: p_{ij} \neq p_{kj} , \qquad \text{for at least one combination of } (i, j, k) . \end{aligned} \qquad (6)$$

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

$$\chi^2 = n \sum_{i=1}^{v} \sum_{j=1}^{r} \frac{(p_{ij} - p_{ij_0})^2}{p_{ij_0}} = \sum_{i=1}^{v} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i\star} \cdot n_{\star j}/n)^2}{n_{i\star} \cdot n_{\star j}/n} = n \left(\sum_{i=1}^{v} \sum_{j=1}^{r} \frac{n_{ij}^2}{n_{i\star} n_{\star j}} - 1 \right) , \qquad (7)$$

which 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.