Anna Janicka

Mathematical Statistics 2021/2022
Lecture 12

1. Hypothesis Testing – Non-parametric tests

The tests discussed in the previous lectures all fell into the category of parametric tests, i.e. tests where the hypotheses concern the values of the parameters of distributions of random variables. During this lecture, we will discuss a different category of tests, where parameter values are not the main topic of interest. This happens for example if we want to check whether a random variable fits (comes from) a specified distribution (when we will perform so-called goodness-of-fit tests), when we want to check whether random variables have the same distribution or when we want to check whether variables or characteristics are independent (when we will perform so-called tests of independence).

1.1. **Goodness of fit tests.** Let us first look at tests that may be used if we want to verify a hypothesis about the distribution of an observed random variable. We will have different categories of tests, depending on whether the specified distribution is continuous or discrete.

1.1.1. *Tests for continuous distributions – Kolmogorov type tests.* Let us assume that we have a sample $X_1, X_2, \ldots, X_n$ from a continuous distribution with cumulative distribution function F, and we want to verify the null hypothesis $H_0 : F = F_0$ (for a specific CDF F_0) against the alternative that the CDF is different. In such a case we may use a test from a class of tests connected with the name of Kolmogorov. These tests are based on theorems which state that regardless of the true form of the cumulative distribution function F_0, if we look at the highest possible difference between F_n – the empirical CDF (based on the sample $X_1, X_2, \ldots, X_n$) – and F_0, the distribution of this difference does not depend on the exact form of F_0 (assuming that the null hypothesis is true). In other words, if we compare the stair-like empirical distribution function F_n with the continuous true cumulative distribution function F_0, the difference between these two functions is s random variable, whose distribution depends only on the number of observations on which the empirical distribution function is based on.

Formally, in the testing procedure we will use a test statistic

$$D_n = \sup_{t\in\mathbb{R}} |F_n(t) - F_0(t)| = \max\{D_n^+, D_n^-\},$$

where

$$D_n^+ = \max_{i=1,\ldots,n} |\frac{i}{n} - F_0(X_{i:n})| \text{ and } D_n^- = \max_{i=1,\ldots,n} |\frac{i-1}{n} - F_0(X_{i:n})|,$$

and reject the null hypothesis if the value of this test statistic is too large (larger than an appropriate quantile of the distribution of D_n, under the null hypothesis). The appropriate quantiles of the resulting Kolmogorov distribution may be found in tables. At this point we will just signal that this distribution formally requires tables for all specific values of n. It can be shown, however, that

$$\mathbb{P}(\sqrt{n}D_n \le d) \xrightarrow{n\to\infty} K(d) = \sum_{i=-\infty}^{+\infty} (-1)^k e^{-2i^2d^2},$$

which means that the function $K(d)$ may be used to compute approximate values of the distribution quantiles for specific values of n. This approximation may be used for $n \ge 100$, and in such cases we have:

$1-\alpha$	0.8	0.9	0.95	0.99
quantile of $K(d)$	1.07	1.22	1.36	1.63
critical value $c(n,\alpha)$ for $n \ge 100$	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

1.1.2. *Tests for discrete distributions – Chi-square type tests.* A totally different class of distributions is used for the verification of hypotheses for discrete distributions. Let us assume that we have a sample of observations from a discrete distribution with k possible values. For simplicity, we will denote these values by $1, \ldots, k$. These values may be treated as value labels – the exact values are not used in the testing procedure (only their probabilities are considered). Let us now assume that we wish to test the null hypothesis H_0 that the distribution probabilities are equal to

i	1	2	3	...	k
$\mathbb{P}(X = i)$	p_1	p_2	p_3	...	p_k

against the alternative that they are not. Let us assume that the observed outcomes in reality are

i	1	2	3	...	k
N_i	N_1	N_2	N_3	...	N_k

where N_i denotes the number of outcomes equal to i, and $N_1 + N_2 + \ldots + N_k = n$. We will use a chi-square test statistic, which has a general form that may be described as

$$\chi^2 = \sum \frac{(\text{observed value - expected value})^2}{\text{expected value}},$$

and in this particular case has the form

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n \cdot p_i)^2}{n \cdot p_i}.$$

Please note that if the empirical distribution shows a perfect fit to the theoretical distribution, the test statistic amounts to zero. The larger the difference between the observed and the expected values, the larger the value of the test statistic. Therefore, we will reject the null hypothesis if the value of the test statistic is larger than an appropriate critical value. It may be shown that if the null hypothesis is true, the distribution of the test statistic χ^2 converges, as n tends to infinity, to $\chi^2(k-1)$ – a chi-square distribution with $k-1$ degrees of freedom (where k is the number of values of the discrete distribution considered). Therefore, we should reject the null hypothesis in favor of the alternative if the calculated test statistic is larger than $c = \chi^2_{1-\alpha}(k-1)$, where $\chi^2_{1-\alpha}(k-1)$ is a quantile of rank $1-\alpha$ of the chi-square distribution with $k-1$ degrees of freedom.

The chi-square test may also be applied in cases where we do not have an exact, single distribution, but a family of different distributions for different values of a parameter (i.e., the model assumed in the null hypothesis is not probabilistic but rather stochastic). In such a case, the test statistics are constructed just like in the above procedure, with the expected values calculated using maximum likelihood estimators of the unknown parameter. Only the number of degrees of freedom changes: it can be shown that if the null hypothesis is true, the distribution of the chi-square test statistic converges as n tends to infinity to a chi-square distribution with $k-d-1$ degrees of freedom, where d is the dimension of the parameter being estimated. The philosophy behind this property is the following: if we compare an empirical distribution with a 'theoretical' distribution, which is not purely theoretical anymore but has been adjusted to fit the data (a ML estimator for the data has been calculated), then for the same sample size we expect better accuracy (smaller errors) than in cases where the empirical distribution is compared to a benchmark which has nothing to do with empirical data. This is equivalent to subtracting degrees of freedom in the chi-square distribution, one degree of freedom per each parameter that has been fit to data.

Example. If we wished to verify whether a collected sample comes from a Poisson distribution with some parameter λ, we would first calculate the maximum likelihood estimator of the value of λ (i.e., the empirical average in this case), and then substitute the estimator value in the formulas when calculating expected counts. The parameter λ is single-dimensional, therefore we would subtract 1 from the number of degrees of freedom of the chi-square distribution that we would treat as a benchmark. Additionally, we would need to pay attention to the number of outcomes, k. A Poisson distribution has infinitely many possible outcomes. However, for the majority of the outcomes (infinitely many of them, in fact), the expected counts will be extremely small. Such cases should be aggregated, in order to avoid bias resulting from division by very small expected counts. Generally, one may apply a rule of thumb stating that if a category has an expected count lower than 5 (taking into account sample size and the value of the ML estimator), it should be merged with adjacent categories. The same rule applies to the probabilistic models described previously: categories with low expected counts should be merged.

It is worth noting that the chi-square goodness of fit test may also be applied to continuous distributions. It suffices to divide the range of values of the studied random variable into classes and count the observations which fall into these classes. The expected probabilities of falling into each class are known (they result from the distribution and may be calculated based on the cumulative distribution function F_0). Once we have the expected and observed counts for the k categories, we may apply the chi-square test just like for the discrete case. It is worth noting, however, that the chi-square test requires larger sample sizes.

1.2. **Tests of independence.** Let us now consider the case where we wish to verify whether two dimensions of a phenomenon under study are independent (for example, whether preference for cakes depends on age or whether income is independent from gender). Using the observation made above, we may assume that the considered distributions are discrete (if they are not, we may divide the value ranges into classes and proceed based on these classes), and that the first dimension has r values, $1, \ldots, r$, while the second dimension has s values, $1, \ldots, s$. In such a case, the two-dimensional random variable has $r \cdot s$ values. Let the theoretical distribution be

$$p_{ij} = \mathbb{P}(X = i, Y = j), \text{ for } i = 1, \ldots, r, \quad j = 1, \ldots, s.$$

Let us introduce the following notation:

$$p_{\bullet j} = \sum_{i=1}^{r} p_{ij}, \quad p_{i\bullet} = \sum_{j=1}^{s} p_{ij}.$$

If we want to verify the independence of the two dimensions, we may write the null hypothesis as

$$H_0 : p_{ij} = p_{i\bullet} \cdot p_{\bullet j}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, s.$$

We may test this null against the alternative that H_0 is not true using a version of the chi-square goodness of fit test. Note that in this case, we wish to verify whether the two-dimensional random vector under study has the required distribution with $r \cdot s$ values, and we have $(r-1) + (s-1)$ unknown parameters to be estimated. These unknown parameters are the probabilities $p_{i\bullet}$ and $p_{\bullet j}$ for $r-1$ and $s-1$ categories, respectively (the last categories may be found using the property that the probabilities in a discrete distribution add up to 1, and that is why there are only $r-1+s-1$ and not $r+s$ unknown parameters).

As far as the estimates of the unknown parameters $p_{i\bullet}$ and $p_{\bullet j}$ are concerned, the ML estimators are the same as the frequency estimators, and can be calculated as:

$$\hat{p}_{i\bullet} = \frac{N_{i\bullet}}{n} \text{ and } \hat{p}_{\bullet j} = \frac{N_{\bullet j}}{n},$$

where $N_{i\bullet}$ and $N_{\bullet j}$ are empirical counts for the aggregate categories

$$N_{\bullet j} = \sum_{i=1}^{r} N_{ij}, \quad N_{i\bullet} = \sum_{j=1}^{s} N_{ij}.$$

Under the null hypothesis of independence, the expected counts np_{ij} can therefore be estimated to be

$$n\hat{p}_{ij} = n \cdot \frac{N_{i\bullet}}{n} \cdot \frac{N_{\bullet j}}{n} = \frac{N_{i\bullet} N_{\bullet j}}{n}.$$

In such a setting, we may use the following test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - N_{i\bullet}N_{\bullet j}/n)^2}{N_{i\bullet}N_{\bullet j}/n},$$

which, under the null hypothesis, has a chi-square distribution with $rs-(r-1)-(s-1)-1 = (r-1)(s-1)$ degrees of freedom, and proceed exactly like in the chi-square goodness of fit testing procedure. Namely, we will reject the null if the value of the test statistic exceeds the quantile of rank $1-\alpha$ of the chi-square distribution with $(r-1)(s-1)$ degrees of freedom.