

Goodness of fit

November 3, 2018

1 Introduction

The problem we will examine in this chapter is of great practical importance: Given an observed sample x_1, x_2, \dots, x_n is a numerical sequence of independent measurements of a random phenomenon whose probability law is not precisely known, we want to test if this sample comes from a given law F , for example from the law $\mathcal{N}(0, 1)$. The methods to be used are called methods for adjusting the observed sample to the theoretical law F . The principle of these methods is as follows: The hypothesis H_0 is that the observed sample is derived from the law F . The method consists in transforming the observed values in a certain way (either into a number for the goodness of fit, or into a function for the goodness of fit of Kolmogorov-Smirnov) so that, depending on the result obtained, we can decide with a confidence level of $1 - \alpha$ (where $\alpha \in]0, 1[$ is given) either to reject the H_0 hypothesis or to accept it.

The generally small number α of order 0.05 or 0.01 is the probability of accepting the hypothesis when it is false; it is the first type risk that we already saw for hypothesis testing. But it is important to understand that it is the experimenter who sets the risk α that he/she accepts to take, the method takes this risk into account and gives a result that ends up to the decision to accept or reject the hypothesis.

2 Chi square test

2.1 Framework

Let us first (re)introduce the multinomial distribution:

Any probability measure concentrated on the set $\{1, 2, \dots, k\}$ of integers between 1 and k , is written

$$\sum_{i=1}^k p_i \delta_i,$$

where δ_i is the Dirac measure in i , $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Such a law will be represented by a line vector $p = (p_1, \dots, p_k)$.

Let $n \geq 1$ be a fixed integer and (X_1, \dots, X_n) a n -sample of a law $p = (p_1, \dots, p_k)$ on $\{1, 2, \dots, k\}$. For each $1 \leq i \leq k$, the random variable

$$N_i = \sum_{j=1}^n \mathbb{1}_{\{X_j=i\}} \tag{1}$$

represents the number of variables in the observed sample that is equal to i .

The law of the random vector $N = (N_1, \dots, N_k)$ is called the multinomial law of parameters $(n; p_1, \dots, p_k)$ and noted $M(n, p)$. It is such that

$$\mathbb{P}(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} \quad (2)$$

where n_1, \dots, n_k are integers such that $n_1 + \dots + n_k = n$.

Let $\bar{p}_n = N/n$. Then, $\mathbb{E}[\bar{p}_n] = p$ and $\lim_{n \rightarrow \infty} \bar{p}_n = p$ almost surely. This means that \bar{p}_n is the empirical estimator of the law p and it is unbiased.

Let us consider the following vector:

$$\left(\frac{N_1 - np_1}{\sqrt{np_1}}, \dots, \frac{N_k - np_k}{\sqrt{np_k}} \right). \quad (3)$$

Theorem 1. $\sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i}$ converges in distribution to a χ_{k-1}^2 as $n \rightarrow \infty$.

Note that the previous variable is quasi-reduced because the true reduced centred variable associated to N_i is $\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}}$.

It should also be noted that the central limit theorem implies that the i -th component of the centered and quasi-reduced vector converges in law towards a normal random variable $\mathcal{N}(0, 1-p_i)$ when $n \rightarrow \infty$. The fact that we find a χ_{k-1}^2 as a limit in the theorem, is related to this and to the non-independence of the components (because $N_1 + \dots + N_k = n$).

2.2 χ^2 test

Let $p = (p_1, \dots, p_k)$ defined as above a reference probability law on the set $\{1, \dots, k\}$. If $q = (q_1, \dots, q_k)$ is another probability law, we define the χ^2 distance between p and q as:

$$\chi^2(p, q) = \sum_{i=1}^k \frac{(p_i - q_i)^2}{p_i}. \quad (4)$$

Care should be taken with this terminology because $\chi^2(p, q)$ is not the square of a true distance and terms $(p_i - q_i)^2$ are more important in Eq 4 when the p_i value is low.

If we take as q the empirical distribution \bar{p}_n defined as above of an n -sample of the law p , we rather measure the difference between p and \bar{p}_n as follows:

$$\chi_n^2(p, \bar{p}_n) = n\chi^2(p, \bar{p}_n) = \sum_{i=1}^k \frac{(np_i - N_i)^2}{np_i}. \quad (5)$$

Note that:

$$\chi_n^2(p, \bar{p}_n) = \sum_{i=1}^k \frac{(N_i - e_i)^2}{e_i}, \quad (6)$$

where $e_i = np_i = \mathbb{E}(N_i)$ is the expected value of N_i .

Another important fact is that the exact law of $\chi_n^2(p, \bar{p}_n)$ is not known but according to the result of Theorem 1, if n is large, we can consider that $\chi_n^2(p, \bar{p}_n)$ follows the χ_{k-1}^2 distribution.

The convergence in law of $\chi_n^2(p, \bar{p}_n)$ to a χ_{k-1}^2 is very sensitive to the fact that \bar{p}_n is the empirical law of p . Indeed, let us suppose that we were wrong and that \bar{p}_n is in fact the empirical law of a law $q \neq p$. So according to the strong law of large numbers, $N_i/n \rightarrow q_i$ a.e. and therefore

$$\frac{1}{n}\chi_n^2(p, \bar{p}_n) = \chi^2(p, \bar{p}_n) = \sum_{i=1}^k \frac{(p_i - N_i/n)^2}{p_i} \rightarrow \chi^2(p, q) > 0 \text{ a.s.} \quad (7)$$

This implies that $\chi_n^2(p, \bar{p}_n) \rightarrow \infty$ a.s.

Thus, if n is large, the observed values of $\chi_n^2(p, \bar{p}_n)$ will be very large. This observation is the basis for the following test:

THE CHI-2 TEST PROCEDURE:

From a n -sample of a discrete law on the whole $\{1, \dots, k\}$, we want to test the hypothesis:

H_0 : the distribution of this sample is p .

In view of the above remark, the quantity $\chi_n^2(p, \bar{p}_n)$ is formed from the observed N_i values and the expected e_i values resulting from the H_0 hypothesis. Following the idea explained above (i.e. if the observed value of $\chi_n^2(p, \bar{p}_n)$ is too large, the H_0 hypothesis is not very credible but it is necessary to specify what is meant by it. The easiest way is to be able to define a bound beyond which H_0 will be rejected. To do this, proceed as follows

- Choose a first type risk α
- With the table of the law of χ_{k-1}^2 , we determine the quantile b_α (i.e. $\mathbb{P}(X > b_\alpha) = \alpha$) for $X \sim \chi_{k-1}^2$.
- Reject H_0 if $\chi_n^2(p, \bar{p}_n) > b_\alpha$.

Justification of the test: Under hypothesis H_0 , and if n is large enough, we know from Theorem 1 that the random variable $\chi_n^2(p, \bar{p}_n)$ follows a χ_{k-1}^2 law. Therefore

$$\mathbb{P}_{H_0}(\text{reject } H_0) = \mathbb{P}(\chi_n^2(p, \bar{p}_n) > b_\alpha) \approx \alpha \quad (8)$$

which shows that the test is justified.

Note: The χ^2 test is only an asymptotic test and n must be large. Let us just say that it is better to avoid using this test if the sample size is less than 50.

2.3 Independence test

Derive the previous test for independence between two random samples.

3 Kolmogorov-Smirnov test

We now examine the goodness of fit of an observed sample x_1, \dots, x_n to a continuous distribution F . The method is completely different from the one discussed above. It is based on the construction of a step function called the empirical distribution function of the observed sample.

3.1 Empirical distribution of a n -sample

Definition 1. Let (X_1, \dots, X_n) be a n -sample of law F on \mathbb{R}^d , $d \geq 1$. We define the empirical distribution of this sample the following random measure:

$$dF_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \quad (9)$$

In the case of $d = 1$, the distribution function F_n of the dF_n measure is called the n-sample empirical distribution function of (X_1, \dots, X_n) . Ranking the observations such as $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ and assuming that they are distinct two by two, the jumps of the function are all equal to $1/n$.

When not all $x_{(i)}$ values are distinct, for example if there is k times the $x_{(i)}$ value in the sample, the jump from F_n to point $x_{(i)}$ is equal to k/n .

The interest of the empirical law F_n is to approach the law F when the sample size n is big enough.

3.2 Fundamental theorem of statistics

Let us continue the previous study. Let F_n be the empirical distribution function of an n-sample (X_1, \dots, X_n) of real random variables of distribution function F .

Let us define the uniform deviation as:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|. \quad (10)$$

Of course, D_n is a random variable since it depends on each realization of the n-sample. A priori the law of D_n seems to depend on F but, (and this is particularly remarkable), this is not the case if we restrict ourselves to continuous distributions F .

Theorem 2. *If F is continuous, the probability law of the random variable D_n is an intrinsic law i.e. it does not depend on F .*

The probability law of D_n has been tabulated for different values of the integer n and its asymptotic behaviour is known for "n large" (see next paragraph). Let us now look at the fundamental result obtained by Glivenko and Cantelli:

Theorem 3 (Fundamental theorem of Statistics). *\mathbb{P} - almost surely , we have*

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0, \quad (11)$$

i.e. $D_n \rightarrow 0$, \mathbb{P} a.e. In other words, the empirical distribution function converges uniformly to F \mathbb{P} - almost surely.

The empirical distribution function is therefore an estimator of the distribution function. The result of the theorem justifies the principle of statistical methods. This principle says that the (unknown) probability law of a random variable can be determined from a (fairly large) sample of drawings made according to this empirical distribution.

3.3 Kolmogorov-Smirnov test procedure

We saw, in the previous paragraph, that the empirical distribution function F_n of an n-sample from an unknown distribution F , provides a very good approximation of the distribution function itself. We will see how this result is used in practice.

We work with a fixed value of n . The law of the random variable D_n can be found in statistical table collections. Thus for all $0 < \alpha < 1$, we can determine the value d_n^α such that

$$\mathbb{P}(D_n \leq d_n^\alpha) = 1 - \alpha, \quad (12)$$

that is to say that $\mathbb{P}(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq d_n^\alpha) = 1 - \alpha$, which is also equivalent to

$$\mathbb{P}(\forall x \in \mathbb{R}, F_n(x) - d_n^\alpha \leq F(x) \leq F_n(x) + d_n^\alpha) = 1 - \alpha. \quad (13)$$

The quantity $1 - \alpha$ is the confidence level and α is the first type risk as before for our function F to be in the interval given by

$$F_n(x) - d_n^\alpha \leq F(x) \leq F_n(x) + d_n^\alpha. \quad (14)$$

This enable to define a confidence tube around F as

Definition 2. $\forall 0 < \alpha < 1$, the random set

$$\{(x, y \in \mathbb{R}^2), F_n(x) - d_n^\alpha \leq y \leq F_n(x) + d_n^\alpha\} \quad (15)$$

is called the confidence band of the distribution function F at confidence level $1 - \alpha$.

Note that the confidence band is obviously sensitive to the value of n chosen initially. Indeed, at a given confidence level $1 - \alpha$, the bandwidth is determined by the value d_n^α which decreases when n increases as shown in the tables. Thus, if we want precision on the estimation of F , the confidence band must be narrow, so that n is large. To have a given width band, the size n that the sample should have will be determined from the tables of the distribution of D_n .

This leads us to the KS test:

We aim at testing, with confidence level $1 - \alpha$ the following hypothesis: $H_0 : "F = F_0"$.

Implementation of the test:

From an observed n -sample of the F distribution, the confidence band is constructed at confidence level $1 - \alpha$. If this band does not entirely contain the graph of the c.d.f. F_0 , H_0 is rejected. In the case where the tape contains the entire F_0 graph, H_0 is not rejected at confidence level $1 - \alpha$.

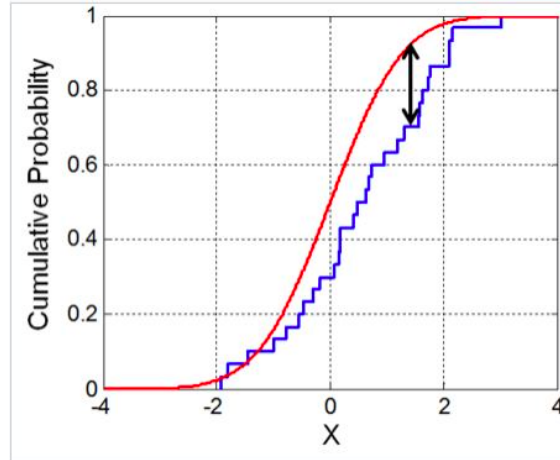



Illustration of the Kolmogorov-Smirnov statistic. 
Red line is CDF, blue line is an ECDF, and the black arrow is the K-S statistic.

Figure 1: caption