Statistics
MDI220
5. Part 2: χ^2 Tests

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

We present χ^2 tests, used to test the adequation of a statistical model or an independence assumption. The name comes from the fact that, in both cases, the statistic used to check the hypothesis tends to a χ^2 distribution for a large number of observations.

1 Test for fit

Let $p = (p_1, \ldots, p_K)$ be a discrete distribution over $\{1, \ldots, K\}$, with $K \geq 2$ and $p_1, \ldots, p_K > 0$. We would like to test the fit of n i.i.d. observations $X_1, \ldots, X_n$ to this distribution. That is, we would like to test the null hypothesis $H_0 = \{X \sim p\}$ against $H_1 = \{X \not\sim p\}$.

Counting. The counts of each category are:

$$\forall i = 1, \ldots, K, \quad N_i = \sum_{t=1}^{n} 1_{\{X_t = i\}}.$$

We have:

$$\mathrm{E}(N_i) = np_i, \quad \mathrm{var}(N_i) = np_i(1 - p_i),$$

so that for large n,

$$\frac{N_i - np_i}{\sqrt{np_i(1 - p_i)}} \approx \mathcal{N}(0, 1).$$

Equivalently,

$$\frac{(N_i - np_i)^2}{np_i(1 - p_i)} \approx \chi^2(1).$$

The following statistic is a measure of the dispersion between expected values and observed values.

χ^2 statistic

$$T(X) = \sum_{i=1}^{K} \frac{(N_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2(K - 1) \quad \text{when } n \to +\infty.$$

Observe that the limiting distribution is not a $\chi^2(K)$ distribution because the random variables $N_1, \ldots, N_K$ are *not* mutually independent. Their sum being equal to 1, there are $K - 1$ degrees of freedom instead of K. The proof of the above result is deferred to the appendix.

Statistical test. The χ^2 test for fit is based on the following decision function:

$$\delta(x) = 1_{\{T(x)>c\}},$$

for some constant c. That is, the null hypothesis is rejected whenever $T(x) > c$. For a test at level α, we have (under the null hypothesis):

$$\alpha = P_0(T(X) > c).$$

The χ^2 test at level α is $\delta(x) = 1_{\{T(x)>c\}}$ with $c = Q(1-\alpha)$, quantile of the $\chi^2(K-1)$ distribution.

Although the result on the χ^2 statistic is asymptotic, the approximation is good whenever $np_k \geq 5$ for all $k = 1, \dots, K$.

Non parametric model. Let P be some distribution. To test the fit of n i.i.d. observations $X_1, \dots, X_n$ to P, we choose a partition $A_1, \dots, A_K$ of the value space. Then we can apply the same test with

$$\forall i = 1, \dots, K, \quad N_i = \sum_{t=1}^{n} 1_{\{X_t \in A_i\}}$$

and

$$p_i = P_0(X_t \in A_i).$$

The partition $A_1, \dots, A_K$ should be chosen so that $np_i \geq 5$ for all $i = 1, \dots, K$.

2 Test for independence

Consider n i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$, each with the same distribution as (X, Y). We would like to test the independence between X and Y, i.e., we would like to test the null hypothesis $H_0 = \{X \perp Y\}$ against $H_1 = \{X \not\perp Y\}$.

We choose two partitions $A_1, \dots, A_K$ and $B_1, \dots, B_L$ of the corresponding value spaces. Let:

$$N_{ij} = \sum_{t=1}^{n} 1_{\{X_t \in A_i, Y_t \in B_j\}}, \quad N_i = \sum_{t=1}^{n} 1_{\{X_t \in A_i\}}, \quad N_j = \sum_{t=1}^{n} 1_{\{Y_t \in B_j\}}.$$

χ^2 statistic for independence

$$T(X,Y) = \sum_{i,j} \frac{(N_{ij} - \frac{N_i N_j}{n})^2}{\frac{N_i N_j}{n}} \xrightarrow{d} \chi^2((K-1)(L-1)) \quad \text{when } n \to +\infty.$$

The test is then:

$$\delta(x,y) = 1_{\{T(x,y)>c\}},$$

i.e., the independence is rejected whenever $T(x,y) > c$. For a test at level α, we have (under the null hypothesis):

$$\alpha = P_0(T(X,Y) > c).$$

For large n, we can take $c = Q(1-\alpha)$ where Q is the quantile function of the $\chi^2((K-1)(L-1))$ distribution.

Appendix

Proof of the main result

Let Y_t be the binary vector of dimension K with component i equal to 1 if and only if $X_t = i$, for all $t = 1, \dots, n$. Then the count vector $N = (N_1, \dots, N_K)$ is given by:

$$N = \sum_{t=1}^{n} Y_t.$$

Let $Y = Y_1$ be the random vector with the same distribution as $Y_1, \dots, Y_n$. We have:

$$\mathrm{E}(Y) = p, \quad \Gamma \overset{d}{=} \mathrm{cov}(Y) = \mathrm{E}(YY^T) - \mathrm{E}(Y)\mathrm{E}(Y)^T = \mathrm{diag}(p) - pp^T,$$

so that, by the Central Limit Theorem,

$$N \approx \mathcal{N}(np, n\Gamma), \quad \text{for } n \to +\infty,$$

that is

$$\frac{N - np}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Gamma).$$

Let $Z \sim \mathcal{N}(0, \Gamma)$ and $D = \mathrm{diag}(1/\sqrt{p})$. Then,

$$T(X) = ||\frac{1}{\sqrt{n}} D(N - np)||^2 \xrightarrow{d} ||DZ||^2.$$

We have:

$$\mathrm{cov}(DZ) = D\Gamma D = I - \sqrt{p}\sqrt{p}^T.$$

Let Q be an orthogonal matrix such that $Q\sqrt{p} = e_1$ (unit vector on the first component); such a matrix exists since $||\sqrt{p}|| = 1$. We have:

$$||DZ||^2 = ||QDZ||^2$$

Moreover,

$$\mathrm{E}(QDZ) = 0, \quad \mathrm{cov}(QDZ) = Q\mathrm{cov}(DZ)Q^T = I - Q\sqrt{p}\sqrt{p}^T Q^T = I - e_1 e_1^T,$$

so that $QDZ \sim \mathcal{N}(0, I - e_1 e_1^T)$. In particular, $||QDZ||^2$ is the sum of the squares of $K - 1$ independent standard normal random variables. We conclude that:

$$T(X) \xrightarrow{d} ||DZ||^2 = ||QDZ||^2 \sim \chi^2(K - 1).$$

$\square$