TO NON-PARAMETRIC TESTS, χ^2-test

Marianna Bolla, DSc

Institute of Mathematics, BME

April 21, 2020

1 Asymptotically χ^2 test statistic

Let $A_1, \dots, A_r$ be a complete set of mutually disjoint events and

$$H_0 : \mathbb{P}(A_i) = p_i \quad (i = 1, \dots, r),$$

where $p_i > 0$, $\sum_{i=1}^r p_i = 1$ are given (sometimes estimated). We make n independent trials, out of which $\nu_1, \dots, \nu_r$ denote the frequencies of $A_1, \dots, A_r$, respectively. As $\sum_{i=1}^r \nu_i = n$, the rv's $\nu_1, \dots, \nu_r$ are not independent! Under H_0, $(\nu_1, \dots, \nu_r)$ follows *multinomial distribution*:

$$\mathbb{P}(\nu_1 = n_1, \dots, \nu_r = n_r \,|\, H_0) = \frac{n!}{n_1! \cdots n_r!} p_1^{n_1} \cdots p_r^{n_r}, \quad \text{if} \quad n_1 + \cdots + n_r = n.$$

A theorem guarantees that in this case

$$\chi^2 = \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i} \to \chi^2(r-1) \tag{1}$$

in distribution as $n \to \infty$.
Remarks:

- The limit distribution does not depend on p_is, it only depends on the number of categories r. Therefore, the χ^2 test is called non-parametric.
- Above n "large" (usually much larger than 30, it depends on r), it is also required that $\nu_i \geq 3$ $(i = 1, \dots, r)$, so there cannot be empty or too "sparse" *cells*.
- Denoting the relative frequencies by $r_i = \nu_i/n$, Equation (1) becomes

  $$n \sum_{i=1}^r \frac{(r_i - p_i)^2}{p_i}$$

  that increases with n. However, with large n, the laws of large numbers guarantee that r_i gets closer and closer to p_i.
- If we estimate number e of parameters, then the degree of freedom of the χ^2-statistic becomes $df = r - 1 - e$.

- Equation (1) is easy to remember as

$$\sum \frac{(O-E)^2}{E}, \tag{2}$$

  where O is the observed and E is the expected frequency.

- In a table, you can see the quantile values of the χ^2-distribution.

We will use the χ^2-test in the following situations.

1.1 Testing Goodness of Fit with χ^2-test

Here the categories A_is either correspond to taking on different values of a discrete rv, or taking on values within disjoint intervals of a continuous rv. Even in the discrete case, some values (for example, above a threshold) are amalgamated, like in the forthcoming Poisson example.

Example (goodness of fit to a discrete population distribution): H_0 is that the number (X) of insurance claims for in-hospital medical care among families with two children in a four-year period follows Poisson distribution. Our data about $n = 200$ such families are here:

Number of claims	0	1	2	3	4	5	6	7	Total
Frequency	22	53	58	39	20	5	2	1	200

Table 1: Number of claims for in-hospital medical care among families with two children in a four-year period

The last three categories are amalgamated as $\{X \geq 5\}$ and we have the following results:

Number of claims	0	1	2	3	4	at least 5	Total
ν_i	22	53	58	39	20	8	200
np_i	27.0	54.2	54.2	36.0	18.0	10.6	200

Here p_is are Poisson probabilities:

$$p_i = \frac{\hat{\lambda}^i}{i!} e^{-\hat{\lambda}}, \quad i = 0, 1, 2, 3, 4$$

and $p_5 = 1 - \sum_{i=0}^{4} p_i$, where

$$\hat{\lambda} = \bar{X} = \frac{0 \times 22 + 1 \times 53 + \cdots + 7 \times 1}{200} = 2.05.$$

The test-statistic is $\chi^2 = 2.33$ with $df = 6 - 1 - 1 = 4$. Since $2.33 < \chi^2_{0.5}(4)$, the P-value at which we could reject H_0 is too large, so we cannot reject it. We accept that the number of claims has Poisson distribution.

Example (goodness of fit to a continuous population distribution): H_0 is that the monthly spending (X) of the customers in a supermarket is normally

(0 - 84.99)	(85.00 - 107.39)	(107.4 - 124.99)	(125 - 142.59)	(142.6 - 164.99)	(165,∞)
14	20	16	19	16	15

Table 2: Number of customers in the spending categories

distributed. They asked 100 customers and obtained $\bar{x} = 125$ and $s^* = 40$ (USD).

Under H_0, X has normal distribution with the estimated parameters $\bar{x}$ and s^*. Therefore,

$$H_0 : Y = \frac{X - 125}{40} \sim \mathcal{N}(0, 1).$$

The observed and expected frequencies in the transformed categories are as follows:

Categories	(-3.125,-1)	(-1,-0.44)	(-0.44,0)	(0,0.44)	(0.44,1)	(1,∞)
ν_i	14	20	16	19	16	15
np_i	15.87	17.13	17.00	17.00	17.13	15.87

Here $p_1 = \mathbb{P}(Y < -1) = \Phi(-1) = 1 - \Phi(1)$, $p_2 = \mathbb{P}(-1 \le Y < -0.44) = \Phi(-0.44) - \Phi(-1) = -\Phi(0.44) + \Phi(1)$, ..., $p_5 = \mathbb{P}(0.44 \le Y < 1) = \Phi(1) - \Phi(0.44)$; eventually, $p_6 = \mathbb{P}(Y \ge 1) = 1 - \sum_{i=1}^{5} p_i$.

The test statistic with Equation (1) is

$$\chi^2 = \frac{(14 - 15.87)^2}{15.87} + \frac{(20 - 17.13)^2}{17.13} + \frac{(16 - 17.00)^2}{17.00} +$$

$$+ \frac{(19 - 17.00)^2}{17.00} + \frac{(16 - 17.13)^2}{17.13} + \frac{(15 - 15.87)^2}{15.87} = 1.12.$$

This value is too "small", not significant at any small α with $df = 6 - 1 - 2 = 3$. Hence, we accept H_0.

1.2 Testing for homogeneity of two samples

Here we want to check the null-hypothesis that two (independent) samples have the same distribution. Irrespective, whether the samples are from discrete or continuous population distribution (both from the same type), we form r categories by dividing the sample space into disjoint values or intervals. The categories are the same for the two samples, and their frequencies are $\nu_1, \ldots, \nu_r$ and $\mu_1, \ldots, \mu_r$ in the two samples, respectively (these are positive integers, usually at least 3).

Example (testing for homogeneity of two diets) 80 and 70 children are put into two different diets (A and B), respectively. After a while, the health condition of the children is recorded:

	Excellent	Medium	Poor	Total
Diet A	37	24	19	80
Diet B	17	33	20	70
Total	54	57	39	150

We wonder whether there is significant difference between the two diets as for the health condition of the children.

$$H_0: \text{ no significant difference.}$$

Under H_0, we can estimate the probability of the three health conditions by the population proportions, merged for the two samples:

$$\hat{p}_1 = \frac{54}{150}, \quad \hat{p}_2 = \frac{57}{150}, \quad \hat{p}_3 = \frac{39}{150}.$$

So under H_0, the expected cell frequencies are $\hat{p}_i n$ and $\hat{p}_i m$, $(i = 1, \dots, r)$, where $r = 3$, $n = 80$, $m = 70$. Wit these, Equation (2) gives for

$$\chi^2 = \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

the following sum over the $2 \times r$ cells:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - \hat{p}_i n)^2}{\hat{p}_i n} + \sum_{i=1}^{r} \frac{(\mu_i - \hat{p}_i m)^2}{\hat{p}_i m} = \cdots = nm \sum_{i=1}^{r} \frac{(\frac{\nu_i}{n} - \frac{\mu_i}{m})^2}{\nu_i + \mu_i}. \quad (3)$$

Here we used that $\hat{p}_i = \frac{\nu_i + \mu_i}{n+m}$.

The above χ^2-statistic under H_0 asymptotically (for n, m 'large') follows χ^2-distribution with $df = r-1$ (indeed, the number $2(r-1)$ of free cells is decreased by the number $r-1$ of the estimated parameters, which are $\hat{p}_1, \dots, \hat{p}_{r-1}$ only, in view of $\sum_{i=1}^{r} \hat{p}_i = 1$).

In the above example, $\chi^2 = 8.224 > \chi^2_{0.025}(2)$, therefore with significance $\alpha = 0.025$ we can reject H_0, which is strong enough evidence for the difference of the two diets.

Note that this was a PROSPECTIVE study, when we decided in advance the sample sizes n and m. It is also called contingency table with ONE MARGIN FIXED.

1.3 Testing for independence of two rv's

Here we have a 2-dimensional sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from the joint distribution of X and Y.

$$H_0: X \quad \textit{and} \quad Y \quad \text{are independent rv's.}$$

For example, whether hair and eye colors are independent, or body hight and weight are independent of each other.

Example (testing independence of political affiliation and attitude toward an energy-rationing program). A random sample of 500 persons is questioned regarding their political affiliation and attitude toward an energy-rationing program:

	Favor	Indifferent	Opposed	Total
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

We divide the sample spaces of X and Y into r and s categories, say $A_1, \ldots, A_r$ and $B_1, \ldots, B_s$, respectively. The above contingency table contains the counts ν_{ij} for $i = 1, \ldots, r$ and $j = 1, \ldots, s$; $\sum_{i=1}^{r} \sum_{j=1}^{n} \nu_{ij} = n$. In this way

$$H_0 : \mathbb{P}(A_i B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j), \quad i = 1, \ldots, r; \quad j = 1, \ldots, s.$$

The observed frequencies are the ν_{ij}s, whereas, the corresponding expected frequencies are

$$n \cdot \frac{\nu_{i\cdot}}{n} \cdot \frac{\nu_{\cdot j}}{n},$$

where $\nu_{i\cdot} = \sum_{j=1}^{s} \nu_{ij}$ and $\nu_{\cdot j} = \sum_{i=1}^{r} \nu_{ij}$ are the two margins.

Therefore,

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

that asymptotically (for n 'large'), under H_0, follows $\chi^2(df)$-distribution with

$$df = rs - 1 - [(r-1) + (s-1)] = (r-1)(s-1).$$

In the above example, $n = 500$ and $df = 2$, since $r = 2$ and $s = 3$. The test statistic is $\chi^2 = 22.153 > \chi^2_{0.005}(2)$, therefore with $\alpha = 0.005$ we can reject H_0, and say that political affiliation and the attitude toward the energy-rationing program are far not independent.

Note that this study was RETROSPECTIVE with NEITHER MARGIN FIXED. Medical investigations are usually such. However, if in the exercise we asked 285 and 215 democrats and republicans (sitting in the senate, for example), then we would test the homogeneity of the attitudes in the two parties. The χ^2-statistic were the same, asymptotically following χ^2-distribution with the same degree of freedom $(2-1)(s-1) = s - 1 = 2$.

2 Bonus questions

1. Prove Equation (3) by accomplishing the calculations in the ... part.

2. Prove that the independence test for a 2×2 contingency table has the χ^2-statistic in the following simple form:

$$\chi^2 = n\frac{(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\cdot}\nu_{2\cdot}\nu_{\cdot 1}\nu_{\cdot 2}}.$$

   Note that $\chi^2 = n\Phi^2$, where $-1 \leq \Phi \leq 1$ is a measure of association between two binary rv's.