chi-squared statistic

Canonical name	ChisquaredStatistic
Date of creation	2013-03-22 15:10:00
Last modified on	2013-03-22 15:10:00
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	9
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62F03
Classification	msc 62G10
Classification	msc 62H17
Synonym	χ^2 statistic
Synonym	chi-square statistic
Synonym	Pearson-chi-squared statistic
Synonym	Pearson-chi-square statistic
Related topic	ChiSquaredRandomVariable
Related topic	HypothesisTesting

Let X be a discrete random variable with m possible outcomes $x_1, \dots, x_m$ with probability of each outcome $\mathrm{P}(X = x_i) = p_i$.

n independent observations are obtained where each observation has the same distribution as X. Bin the observations into m groups, so that each group contains all observations having the same outcome x_i. Next, count the number of observations in each group to get $n_1, \dots, n_k$ corresponding to the outcomes $x_1, \dots, x_k$, so that $n = \sum n_i$. It is desired to find out how close the actual number of outcomes n_i are to their expected values np_i.

Intuitively, this "closeness" depends on how big the sample is, and how large the deviations are between the observed and the expected, for all categories. The value

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{1}$$

called the χ^2 *statistic*, or the *chi-squared statistic*, is such a measure of "closeness". It is also known as the *Pearson-chi-squared* statistic, in honor of the English statistician Karl Pearson, who showed that (1) has approximately a `http://planetmath.org/ChiSquaredRandomVariable`chi-squared distribution with $m - 1$ degrees of freedom. The degree of freedom depends on the number of free variables in χ^2, and is not always $m - 1$, as we will see in Example 3.

Usually, χ^2 statistic is utilized in hypothesis testing, where the null hypothesis specifies that the actual equals the expected. A large value of χ^2 means either the deviations from the expectations are large or the sample is small, and therefore, either the null hypothesis should be rejected or there is not enough information to give a meaningful interpretation. How large of a deviation, compared to the sample size, is enough to reject the null hypothesis depends on the degree of freedom of chi-squared distribution of χ^2 and the specified critical values.

Examples.

1. Suppose a coin is tossed 10 times and 7 heads are observed. We would like to know if the coin is fair based on the observations. We have the

following hypothesis:

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

Break up the observations into two groups: heads and tails. Then, according to H_0,

$$\chi^2 = \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} = 1.60.$$

Checking the table of critical values of chi-squared distributions, we see that at degree of freedom = 1, there is a 0.100 chance that the χ^2 value is higher than 2.706. Since $1.600 < 2.706$, we may not want to reject the null hypothesis. However, we may not want to outrightly accept it either simply because the sample size is not very large.

2. Now, a coin is tossed 100 times and 70 heads are observed. Using the same null hypothesis as above,

$$\chi^2 = \frac{(70-50)^2}{50} + \frac{(30-50)^2}{50} = 16.00.$$

Even at p-value = 0.005, the corresponding critical value of 7.879 is quite a bit smaller than 16. So we will reject the null hypothesis even at confidence level 99.5%(= 1−p-value).

3. χ^2 statistic can be used in non-parametric situations as well, particularly, in contingency tables. Three dice of varying sizes are each tossed 100 times and the top faces are recorded. The results of the count of each possible value of the top face, for each die is summarized in the following table:

Die\top face	1	2	3	4	5	6	all
Die 1	16	19	17	15	19	14	100
Die 2	17	18	14	13	22	16	100
Die 3	12	20	19	18	20	11	100
All dice	45	57	50	46	61	41	300

Let X_i = count of top face= i, and Y_j = Die j. Next, we want to test the following hypotheses:

$$H_0 : X_i \text{ is independent of } Y_j \qquad H_1 : \text{otherwise}.$$

Since we do not know the exact distribution of the top faces, we approximate the distribution by using the last row. For example, the (marginal) probability that top face = 1 is $\frac{45}{300} = 0.15$. This says that the probability that top face = 1 in Die $i = 0.15 \times \frac{1}{3} = 0.05$. Then, based on the null hypothesis, we have the following table of "expected count":

Die\top face	1	2	3	4	5	6
Die 1	15.0	19.0	16.7	15.3	20.3	13.7
Die 2	15.0	19.0	16.7	15.3	20.3	13.7
Die 3	15.0	19.0	16.7	15.3	20.3	13.7

For each die, we can compute the χ^2. For instance, for the first die,

$$\begin{aligned}\chi^2 &= \frac{(16-15.0)^2}{15.0} + \frac{(19-19.0)^2}{19.0} + \frac{(17-16.7)^2}{16.7} + \\ &\quad \frac{(15-15.3)^2}{15.3} + \frac{(19-20.3)^2}{20.3} + \frac{(14-13.7)^2}{13.7} \\ &= 0.176\end{aligned}$$

The results are summarized in the following

	χ^2	degrees of freedom
Die 1	0.176	5
Die 2	1.636	5
Die 3	1.969	0
All dice	3.781	10

Note that the degree of freedom for the last dice is 0 because the expected counts in the last row are completely determined by those in the first two rows (and the totals). Looking up the table, we see that there is a 90% that the value of χ^2 will be greater than 4.865, and since $3.781 < 4.865$, we accept the null hypothesis: the outcomes of the tosses have no bearing on which die is tossed.

Remark. In general, for a $p \times q$ 2-way contingency table, the χ^2 statistic is given by

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \qquad (2)$$

where n_{ij} and m_{ij} are the actual and expected counts in Cell (i, j). When the sample is large, χ^2 has a chi-squared distribution with $(p-1)(q-1)$ degrees of freedom. In particular, when testing for the independence between two categorical variables, the expected count m_{ij} is

$$m_{ij} = \frac{n_{i*}n_{*j}}{n}, \text{ where } n_{i*} = \sum_{j=1}^{q} n_{ij}, \; n_{*j} = \sum_{i=1}^{p} n_{ij}, \text{ and } n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}.$$