Pearson's chi-squared test

Pearson's chi-squared test (χ^2) is the best-known of many chi-squared tests (Yates, likelihood ratio, portmanteau test in time series, etc.) – statistical procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were first investigated by Karl Pearson in 1900.[1] In contexts where it is important to improve a distinction between the test statistic and its distribution, names similar to **Pearson X-squared** test or statistic are used.

It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair", i. e., all six outcomes are equally likely to occur.

Definition

Pearson's chi-squared test is used to assess two types of comparison: tests of goodness of fit and tests of independence.

- A test of **goodness of fit** establishes whether or not an observed frequency distribution differs from a theoretical distribution.
- A **test of independence** assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

The procedure of the test includes the following steps:

1. Calculate the chi-squared test statistic, χ^2, which resembles a normalized sum of squared deviations between observed and theoretical frequencies (see below).
2. Determine the degrees of freedom, *d*, of that statistic, which is essentially the number of frequencies reduced by the number of parameters of the fitted distribution.
3. Compare χ^2 to the critical value from the chi-squared distribution with *d* degrees of freedom, which in many cases gives a good approximation of the distribution of χ^2.

Test for fit of a distribution

Discrete uniform distribution

In this case N observations are divided among n cells. A simple application is to test the hypothesis that, in the general population, values would occur in each cell with equal frequency. The "theoretical frequency" for any cell (under the null hypothesis of a discrete uniform distribution) is thus calculated as

$$E_i = \frac{N}{n},$$

and the reduction in the degrees of freedom is $p = 1$, notionally because the observed frequencies O_i are constrained to sum to N.

Other distributions

When testing whether observations are random variables whose distribution belongs to a given family of distributions, the "theoretical frequencies" are calculated using a distribution from that family fitted in some standard way. The reduction in the degrees of freedom is calculated as $p = s + 1$, where s is the number of co-variates used in fitting the distribution. For instance, when checking a three-co-variate Weibull distribution, $p = 4$, and when checking a normal distribution (where the parameters are mean and standard deviation), $p = 3$. In other words, there will be $n - p$ degrees of freedom, where n is the number of categories.
It should be noted that the degrees of freedom are not based on the number of observations as with a Student's t or F-distribution. For example, if testing for a fair, six-sided dice, there would be five degrees of freedom because there are six categories/parameters (each number). The number of times the die is rolled will have absolutely no effect on the number of degrees of freedom.

Calculating the test-statistic

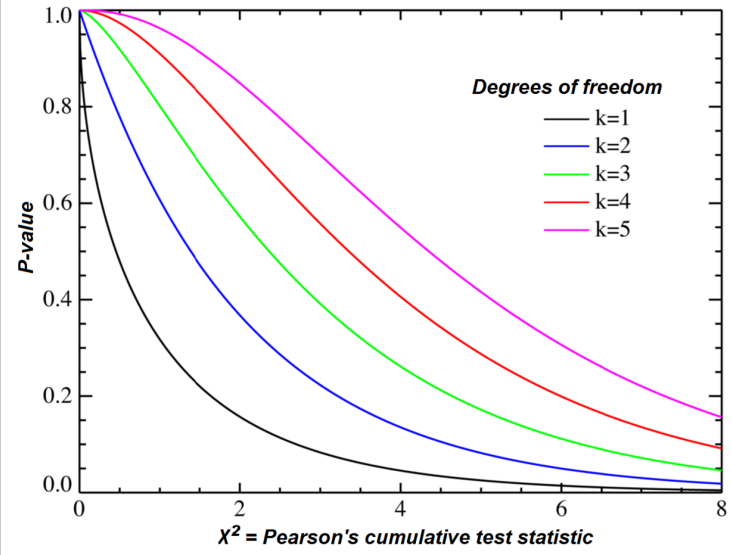
The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where

χ^2 = Pearson's cumulative test statistic, which asymptotically approaches a χ^2 distribution.
O_i = an observed frequency;
E_i = an expected (theoretical) frequency, asserted by the null hypothesis;
n = the number of cells in the table.

The chi-squared statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-squared distribution. The number of degrees of freedom is equal to the number of cells n, minus the reduction in degrees of freedom, p.



Chi-squared distribution, showing X^2 on the x-axis and P-value on the y-axis.

The result about the number of degrees of freedom is valid when the original data are multinomial and hence the estimated parameters are efficient for minimizing the chi-squared statistic. More generally however, when maximum likelihood estimation does not coincide with minimum chi-squared estimation, the distribution will lie somewhere between a chi-squared distribution with $n - 1 - p$ and $n - 1$ degrees of freedom (See for instance Chernoff and Lehmann, 1954).

Bayesian method

In Bayesian statistics, one would instead use a Dirichlet distribution as conjugate prior. If one took a uniform prior, then the maximum likelihood estimate for the population probability is the observed probability, and one may compute a credible region around this or another estimate.

Test of independence

In this case, an "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each observation is allocated to one cell of a two-dimensional array of cells (called a contingency table) according to the values of the two outcomes. If there are *r* rows and *c* columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = \frac{\left(\sum_{n_c=1}^{c} O_{i,n_c}\right) \cdot \left(\sum_{n_r=1}^{r} O_{n_r,j}\right)}{N},$$

where *N* is the total sample size (the sum of all cells in the table). The value of the test-statistic is

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

Fitting the model of "independence" reduces the number of degrees of freedom by $p = r + c - 1$. The number of degrees of freedom is equal to the number of cells *rc*, minus the reduction in degrees of freedom, *p*, which reduces to $(r - 1)(c - 1)$.

For the test of independence, also known as the test of homogeneity, a chi-squared probability of less than or equal to 0.05 (or the chi-squared statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is independent of the column variable.[2] The alternative hypothesis corresponds to the variables having an association or relationship where the structure of this relationship is not specified.

Assumptions

The chi-squared test, when used with the standard approximation that a chi-squared distribution is applicable, has the following assumptions:[*citation needed*]

- Simple random sample – The sample data is a random sampling from a fixed distribution or population where every collection of members of the population of the given sample size has an equal probability of selection. Variants of the test have been developed for complex samples, such as where the data is weighted.
- Sample size (whole table) – A sample with a sufficiently large size is assumed. If a chi squared test is conducted on a sample with a smaller size, then the chi squared test will yield an inaccurate inference. The researcher, by using chi squared test on small samples, might end up committing a Type II error.
- Expected cell count – Adequate expected cell counts. Some require 5 or more, and others require 10 or more. A common rule is 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables, but no cells with zero expected count. When this assumption is not met, Yates's Correction is applied.
- Independence – The observations are always assumed to be independent of each other. This means chi-squared cannot be used to test correlated data (like matched pairs or panel data). In those cases you might want to turn to McNemar's test.

A test that relies on different assumptions is Fisher's exact test; if its assumption of fixed marginal distributions is met it is substantially more accurate in obtaining a significance level, especially with few observations. In the vast majority of applications this assumption will not be met, and Fisher's exact test will be over conservative and not have correct coverage.[*citation needed*]

Examples

Goodness of fit

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women. If there were 44 men in the sample and 56 women, then

$$\chi^2 = \frac{(44-50)^2}{50} + \frac{(56-50)^2}{50} = 1.44.$$

If the null hypothesis is true (i.e., men and women are chosen with equal probability), the test statistic will be drawn from a chi-squared distribution with one degree of freedom. If the male frequency is known, then the female frequency is determined.

Consultation of the chi-squared distribution for 1 degree of freedom shows that the probability of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.23. This probability is higher than conventional criteria for statistical significance (0.001–0.05), so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women (i.e., we would consider our sample within the range of what we'd expect for a 50/50 male/female ratio.)

Problems

The approximation to the chi-squared distribution breaks down if expected frequencies are too low. It will normally be acceptable so long as no more than 20% of the events have expected frequencies below 5. Where there is only 1 degree of freedom, the approximation is not reliable if expected frequencies are below 10. In this case, a better approximation can be obtained by reducing the absolute value of each difference between observed and expected frequencies by 0.5 before squaring; this is called Yates's correction for continuity.

In cases where the expected value, E, is found to be small (indicating a small underlying population probability, and/or a small number of observations), the normal approximation of the multinomial distribution can fail, and in such cases it is found to be more appropriate to use the G-test, a likelihood ratio-based test statistic. Where the total sample size is small, it is necessary to use an appropriate exact test, typically either the binomial test or (for contingency tables) Fisher's exact test; but note that this test assumes fixed and known totals in all margins, an assumption which is typically false.

Distribution

The null distribution of the Pearson statistic with j rows and k columns is approximated by the chi-squared distribution with $(k-1)(j-1)$ degrees of freedom.[3]

This approximation arises as the true distribution, under the null hypothesis, if the expected value is given by a multinomial distribution. For large sample sizes, the central limit theorem says this distribution tends toward a certain multivariate normal distribution.

Two cells

In the special case where there are only two cells in the table, the expected values follow a binomial distribution,

$$E \sim \mathrm{Bin}(n, p),$$

where

p = probability, under the null hypothesis,

n = number of observations in the sample.

In the above example the hypothesised probability of a male observation is 0.5, with 100 samples. Thus we expect to observe 50 males.

If n is sufficiently large, the above binomial distribution may be approximated by a Gaussian (normal) distribution and thus the Pearson test statistic approximates a chi-squared distribution,

$$\mathrm{Bin}(n, p) \approx \mathrm{N}(np, np(1-p)).$$

Let O_1 be the number of observations from the sample that are in the first cell. The Pearson test statistic can be expressed as

$$\frac{(O_1 - np)^2}{np} + \frac{(n - O_1 - n(1-p))^2}{n(1-p)},$$

which can in turn be expressed as

$$\left(\frac{O_1 - np}{\sqrt{np(1-p)}}\right)^2.$$

By the normal approximation to a binomial this is the squared of one standard normal variate, and hence is distributed as chi-squared with 1 degree of freedom. Note that the denominator is one standard deviation of the Gaussian approximation, so can be written

$$\frac{(O_1 - \mu)^2}{\sigma^2}.$$

So as consistent with the meaning of the chi-squared distribution, we are measuring how probable the observed number of standard deviations away from the mean is under the Gaussian approximation (which is a good approximation for large n).

The chi-squared distribution is then integrated on the right of the statistic value to obtain the P-value, which is equal to the probability of getting a statistic equal or bigger than the observed one, assuming the null hypothesis.

Two-by-two contingency tables

When the test is applied to a contingency table containing two rows and two columns, the test is equivalent to a Z-test of proportions.

Many cells

Similar arguments as above lead to the desired result. [*citation needed*] Each cell (except the final one, whose value is completely determined by the others) is treated as an independent binomial variable, and their contributions are summed and each contributes one degree of freedom.

Notes

[3] Statistics for Applications. *MIT OpenCourseWare*. Lecture 23 (http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2003/lecture-notes/lec23.pdf). Pearson's Theorem. Retrieved 21 March 2007.

References

- Chernoff, H.; Lehmann, E. L. (1954). "The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit". *The Annals of Mathematical Statistics* **25** (3): 579–586. doi: 10.1214/aoms/1177728726 (http://dx.doi.org/10.1214/aoms/1177728726).
- Plackett, R. L. (1983). "Karl Pearson and the Chi-Squared Test". *International Statistical Review* (International Statistical Institute (ISI)) **51** (1): 59–72. doi: 10.2307/1402731 (http://dx.doi.org/10.2307/1402731). JSTOR 1402731 (http://www.jstor.org/stable/1402731).
- Greenwood, P.E.; Nikulin, M.S. (1996). *A guide to chi-squared testing*. New York: Wiley. ISBN 0-471-55779-X.

Article Sources and Contributors

Pearson's chi-squared test *Source*: http://en.wikipedia.org/w/index.php?oldid=567048567 *Contributors*: A-k-h, AbsolutDan, Agüeybaná, Aljeirou, Andropod, Arcadian, Asqueella, Athaler, Avraham, Bender235, BlaiseFEgan, Bobo192, Btyner, Bubba73, Cherkash, Chuck Carroll, Connet, Cortonin, Czenek, Delirium, Den fjättrade ankan, Dlituiev, Doyoung, Dpbsmith, Egil, Ektodu, Fgnievinski, Free Software Knight, Funk17, Furrykef, Giftlite, Giuseppedn, Grotendeels Onschadelijk, Hirak 99, Horn.imh, Jcobb, Jfitzg, JimsMaher, Jmcclung711, Joel B. Lewis, John254, Jporitz, JustAGal, Kastchei, Kgwet, Kjtobo, Kmg90, KohanX, Kwamikagami, Lambiam, Lexor, LilHelpa, Loadmaster, Loodog, Mad Scientist, MarkSweep, Mathonius, Matt Crypto, Maxal, Maxbox51, Melcombe, Michael Hardy, Mikael Häggström, Mikko8, Mirko Horvacki, Motorneuron, Moverly, MrOllie, Muhali, MusikAnimal, N5iln, Navywings, Nbarth, Neffk, O18, Omicron1234, Paul August, Paulck, Piotrus, PowerWill500, Qartis, Quadduc, Qwfp, Ranger2006, Rar74B, Requestion, Retobaum, Rjwilmsi, Rks22, Robinh, Rvrocha, Sander123, Sayantan m, Sbmehta, Schwnj, Seglea, Shadow308b4, Skbkekas, Spangineer, Ssola, Talgalili, Tambal, Tayste, The Anome, Tim bates, TimBentley, TimBock, ToddDeLuca, Tomi, Triacylglyceride, Wtmitchell, 188 anonymous edits

Image Sources, Licenses and Contributors

File:Chi-square distributionCDF-English.png *Source*: http://en.wikipedia.org/w/index.php?title=File:Chi-square_distributionCDF-English.png *License*: Public Domain *Contributors*: Mikael Häggström

License