Chi-square test for goodness-of-fit

The chi-square test for goodness-of-fit is an alternative to the G-test for goodness-of-fit. Most of the information on this page is identical to that on the G-test page. You should read the section on "Chi-square vs. G-test" near the bottom of this page, pick either chi-square or G-test, then stick with that choice for the rest of your life.

When to use it

Use the chi-square test for goodness-of-fit when you have one nominal variable with two or more values (such as red, pink and white flowers). The observed counts of numbers of observations in each category are compared with the expected counts, which are calculated using some kind of theoretical expectation (such as a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross).

If the expected number of observations in any category is too small, the chi-square test may give inaccurate results, and an exact test or a randomization test should be used instead. See the web page on small sample sizes for further discussion.

Null hypothesis

The statistical null hypothesis is that the number of observations in each category is equal to that predicted by a biological theory, and the alternative hypothesis is that the observed numbers are different from the expected. The null hypothesis is usually an extrinsic hypothesis, one for which the expected proportions are determined before doing the experiment. Examples include a 1:1 sex ratio or a 1:2:1 ratio in a genetic cross. Another example would be looking at an area of shore that had 59% of the area covered in sand, 28% mud and 13% rocks; if seagulls were standing in random places, your null hypothesis would be that 59% of the seagulls were standing on sand, 28% on mud and 13% on rocks.

In some situations, an intrinsic hypothesis is used. This is a null hypothesis in which the expected proportions are calculated after the experiment is done, using some of the information from the data. The best-known example of an intrinsic hypothesis is the Hardy-Weinberg proportions of population genetics: if the frequency of one allele in a population is *p* and the other allele is *q*, the null hypothesis is that expected frequencies of the three genotypes are p^2, $2pq$, and q^2. This is an intrinsic hypothesis, because *p* and *q* are estimated from the data after the experiment is done, not predicted by theory before the experiment.

How the test works

The test statistic is calculated by taking an observed number (*O*), subtracting the expected number (*E*), then squaring this difference. The larger the deviation from the null hypothesis, the larger the difference between observed and expected is. Squaring the differences makes them all positive. Each difference is divided by the expected number, and these standardized differences are summed. The test statistic is conventionally called a "chi-square" statistic, although this is somewhat confusing (it's just one of many test statistics that follows the chi-square distribution). The equation is

$$\text{chi}^2 = \sum(O-E)^2/E$$

As with most test statistics, the larger the difference between observed and expected, the larger the test statistic becomes.

The distribution of the test statistic under the null hypothesis is approximately the same as the theoretical chi-square distribution. This means that once you know the chi-square test statistic, you can calculate the probability of getting that value of the chi-square statistic.

The shape of the chi-square distribution depends on the number of degrees of freedom. For an extrinsic null hypothesis (the much more common situation, where you know the proportions predicted by the null hypothesis before collecting the data), the number of degrees of freedom is simply the number of values of the variable, minus one. Thus if you are testing a null hypothesis of a 1:1 sex ratio, there are two possible values (male and female), and therefore one degree of freedom. This is because once you know how many of the total are females (a number which is "free" to vary from 0 to the sample size), the number of males is determined. If there are three values of the variable (such as red, pink, and white), there are two degrees of freedom, and so on.

An intrinsic null hypothesis is one in which you estimate one or more parameters from the data in order to get the numbers for your null hypothesis. As described above, one example is Hardy-Weinberg proportions. For an intrinsic null hypothesis, the number of degrees of freedom is calculated by taking the number of values of the variable, subtracting 1 for each parameter estimated from the data, then subtracting 1 more. Thus for Hardy-Weinberg proportions with two alleles and three genotypes, there are three values of the variable (the three genotypes); you subtract one for the parameter estimated from the data (the allele frequency, *p*); and then you subtract one more, yielding one degree of freedom.

Examples: extrinsic hypothesis

Mendel crossed peas that were heterozygotes for Smooth/wrinkled, where Smooth is dominant. The expected ratio in the offspring is 3 Smooth: 1 wrinkled. He observed 423 Smooth and 133 wrinkled.

The expected frequency of Smooth is calculated by multiplying the sample size (556) by the expected proportion (0.75) to yield 417. The same is done for green to yield 139. The number of degrees of freedom when an extrinsic hypothesis is used is the number of values of the nominal variable minus one. In this case, there are two values (Smooth and wrinkled), so there is one degree of freedom.

The result is chi-square=0.35, 1 d.f., P=0.557, indicating that the null hypothesis cannot be rejected; there is no significant difference between the observed and expected frequencies.

Mannan and Meslow (1984) studied bird foraging behavior in a forest in Oregon. In a managed forest, 54% of the canopy volume was Douglas fir, 40% was ponderosa pine, 5% was grand fir, and 1% was western larch. They made 156 observations of foraging by red-breasted nuthatches; 70 observations (45% of the total) in Douglas fir, 79 (51%) in ponderosa pine, 3 (2%) in grand fir, and 4 (3%) in western larch. The biological null hypothesis is that the birds forage randomly, without regard to what species of tree they're in; the statistical null hypothesis is that the proportions of foraging events are equal to the proportions of canopy volume. The difference in proportions is significant (chi-square=13.593, 3 d.f., P=0.0035).

The expected numbers in this example are pretty small, so it would be better to analyze it with an exact test or a randomization test. I'm leaving it here because it's a good example of an extrinsic hypothesis that comes from measuring something (canopy volume, in this case), not a mathematical theory.

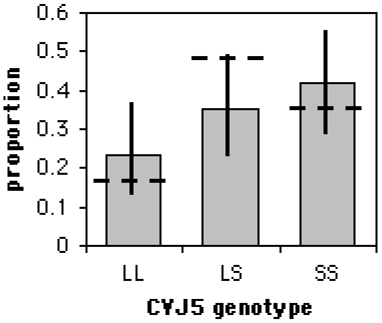
Example: intrinsic hypothesis

McDonald et al. (1996) examined variation at the CVJ5 locus in the American oyster, *Crassostrea virginica*. There were two alleles, L and S, and the genotype frequencies in Panacea, Florida were 14 LL, 21 LS, and 25 SS. The estimate of the L allele proportion from the data is 49/120=0.408. Using the Hardy-Weinberg formula and this estimated allele proportion, the expected genotype proportions are 0.167 LL, 0.483 LS, and 0.350 SS. There are three classes (LL, LS and SS) and one parameter estimated from the data (the L allele proportion), so there is one degree of freedom. The result is chi-square=4.54, 1 d.f., P=0.033, which is significant at the 0.05 level. We can reject the null hypothesis that the data fit the expected Hardy-Weinberg proportions.

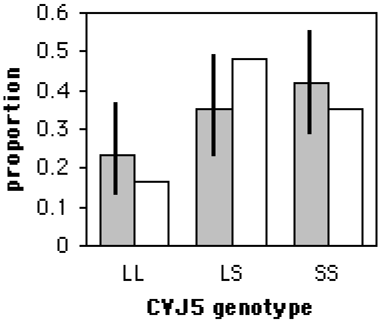
Graphing the results

If there are just two values of the nominal variable, you wouldn't display the result in a graph, as that would be a bar graph with just one bar. Instead, you just report the proportion; for example, Mendel found 23.9% wrinkled peas in his cross.

With more than two values of the nominal variable, you'd usually present the results of a goodness-of-fit test in a table of observed and expected proportions. If the expected values are obvious (such as 50%) or easily calculated from the data (such as Hardy–Weinberg proportions), you can omit the expected numbers from your table. For a presentation you'll probably want a graph showing both the observed and expected proportions, to give a visual impression of how far apart they are. You should use a bar graph for the observed proportions; the expected can be shown with a horizontal dashed line, or with bars of a different pattern.



Genotype proportions at the CVJ5 locus in the American oyster. Horizontal dashed lines indicate the expected proportions under Hardy–Weinberg equilibrium; error bars indicate 95% confidence intervals.



Genotype proportions at the CVJ5 locus in the American oyster. Gray bars are observed proportions, with 95% confidence intervals; white bars are expected proportions under Hardy–Weinberg equilibrium.

One way to get the horizontal lines on the graph is to set up the graph with the observed proportions and error bars, set the scale for the Y-axis to be fixed for the minimum and maximum you want, and get everything formatted (fonts, patterns, etc.). Then replace the observed proportions with the expected proportions in the spreadsheet; this should make the columns change to represent the expected values. Using the spreadsheet drawing tools, draw horizontal lines at the top of the columns. Then put the observed proportions back into the spreadsheet. Of course, if the expected proportion is something simple like 25%, you can just draw the horizontal line all the way across the graph.

Similar tests

The chi-square test of independence is used for two nominal variables, not one.

There are several tests that use chi-square statistics. The one described here is formally known as Pearson's chi-square. It is by far the most common chi-square test, so it is usually just called the chi-square test.

You have a choice of four goodness-of-fit tests: the exact binomial test or exact multinomial test, the G-test of goodness-of-fit,, the chi-square test of goodness-of-fit, or the randomization test. For small values of the expected numbers, the chi-square and G-tests are inaccurate, because the distribution of the test statistics do not fit the chi-square distribution very well.

The usual rule of thumb is that you should use the exact test or randomization test when the smallest expected value is less than 5, and the chi-square and G-tests are accurate enough for larger expected values. This rule of thumb dates from the olden days when statistics were done by hand, and the calculations for the exact test were very tedious and to be avoided if at all possible. Nowadays, computers make it just as easy to do the exact test or randomization test as the computationally simpler chi-square or G-test. I recommend that you use the exact test when the total sample size is less than 1000. With sample sizes between 50 and 1000, it generally doesn't make much difference which test you use, so you shouldn't criticize someone for using the chi-square or G-test (as I have in the examples above). See the web page on small sample sizes for further discussion.

Chi-square vs. G-test

The chi-square test gives approximately the same results as the G-test. Unlike the chi-square test, G-values are additive, which means they can be used for more elaborate statistical designs, such as repeated G-tests of goodness-of-fit. G-tests are a subclass of likelihood ratio tests, a general category of tests that have many uses for testing the fit of data to mathematical models; the more elaborate versions of likelihood ratio tests don't have equivalent tests using the Pearson chi-square statistic. The G-test is therefore preferred by many, even for simpler designs. On the other hand, the chi-square test is more familiar to more people, and it's always a good idea to use statistics that your readers are familiar with when possible. You may want to look at the literature in your field and see which is more commonly used.

How to do the test

Spreadsheet

I have set up a spreadsheet for the chi-square test of goodness-of-fit. It is largely self-explanatory. It will calculate the degrees of freedom for you if you're using an extrinsic null hypothesis; if you are using an intrinsic hypothesis, you must enter the degrees of freedom into the spreadsheet.

Web pages

There are also web pages that will perform this test here, here, or here. None of these web pages lets you set the degrees of freedom to the appropriate value for testing an intrinsic null hypothesis.

SAS

Here is a SAS program that uses PROC FREQ for a chi-square test. It uses the Mendel pea data from above, and it assumes you've already counted the number of smooth and wrinkled peas. The `weight count` command tells SAS that the 'count' variable is the number of times each value of 'texture' was observed. The `zeros` option tells it to include observations with counts of zero, for example if you had 20 smooth peas and 0 wrinkled peas; it doesn't hurt to always include the `zeros` option. `chisq` tells SAS to do a chi-square test, and `testp=(75 25);` tells it the expected percentages. The expected percentages must add up to 100. The expected percentages are given for the values of 'texture' in alphabetical order: 75 percent 'smooth', 25 percent 'wrinkled'.

```
data peas;
   input texture $ count / zeros;
   cards;
smooth 423
wrinkled 133
;
proc freq data=peas;
   weight count;
   tables texture / chisq testp=(75 25);
run;
```

Here's a SAS program that uses PROC FREQ for a chi-square test on raw data. I've used three dots to indicate that I haven't shown the complete data set.

```
data peas;
   input texture $;
   cards;
smooth
wrinkled
smooth
smooth
wrinkled
smooth
   .
```

```
      .
      .
   smooth
   smooth
   ;
   proc freq data=peas;
      tables texture / chisq testp=(75 25);
   run;
```

The output includes the following:

```
        Chi-Square Test
   for Specified Proportions
   -------------------------
   Chi-Square         0.3453
   DF                      1
   Pr > ChiSq         0.5568
```

You would report this as "chi-square=0.3453, 1 d.f., P=0.5568."

Power analysis

If your nominal variable has just two values, you can use the power calculator on the exact binomial page.

If your nominal variable has more than two values, use the free G*Power program. Choose "Goodness-of-fit tests: Contingency tables" from the Statistical Test menu, then choose "Chi-squared tests" from the Test Family menu. To calculate effect size, click on the Determine button and enter the null hypothesis proportions in the first column and the proportions you hope to see in the second column. Then click on the Calculate and Transfer to Main Window button. Set your alpha and power, and be sure to set the degrees of freedom (Df); for an extrinsic null hypothesis, that will be the number of rows minus one.

As an example, let's say you want to do a genetic cross of snapdragons with an expected 1:2:1 ratio, and you want to be able to detect a pattern with 5 percent more heterozygotes that expected. Enter 0.25, 0.50, and 0.25 in the first column, enter 0.225, 0.55, and 0.225 in the second column, click on Calculate and Transfer to Main Window, enter 0.05 for alpha, 0.80 for power, and 2 for degrees of freedom. The result is a total sample size of 964.

Further reading

Sokal and Rohlf, p. 701.

Zar, pp. 462-466.

References

Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. J. Wildl. Manage. 48: 1219-1238.

McDonald, J.H., B.C. Verrelli and L.B. Geyer. 1996. Lack of geographic variation in anonymous nuclear polymorphisms in the American oyster, *Crassostrea virginica*. Molecular Biology and Evolution 13: 1114-1118.

This page was last revised September 11, 2009. Its address is http://udel.edu/~mcdonald/statchigof.html. It may be cited as pp. 39-45 in: McDonald, J.H. 2009. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland.