2.3: Chi-Square Test of Goodness-of-Fit

Learning Objectives

- Study the use of chi-square test of goodness-of-fit when you have one nominal variable
- To see if the number of observations in each category fits a theoretical expectation, and the sample size is large

When to use it

Use the chi-square test of goodness-of-fit when you have one nominal variable with two or more values (such as red, pink and white flowers). You compare the observed counts of observations in each category with the expected counts, which you calculate 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 you should use an exact test instead. See the web page on small sample sizes for discussion of what "small" means.

The chi-square test of goodness-of-fit is an alternative to the *G*–test of goodness-of-fit; each of these tests has some advantages and some disadvantages, and the results of the two tests are usually very similar. 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. Much of the information and examples on this page are the same as on the *G*–test page, so once you've decided which test is better for you, you only need to read one.

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, where you knew the expected proportions 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 you were investigating where seagulls like to stand, your null hypothesis would be that 59% of the seagulls were standing on sand, 28% on mud and 13% on rocks.

In some situations, you have an intrinsic hypothesis. This is a null hypothesis where you calculate the expected proportions after you do the experiment, 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 you estimate p and q from the data after you collect the data, you can't predict p and q before the experiment.

How the test works

Unlike the exact test of goodness-of-fit, the chi-square test does not directly calculate the probability of obtaining the observed results or something more extreme. Instead, like almost all statistical tests, the chi-square test has an intermediate step; it uses the data to calculate a test statistic that measures how far the observed data are from the null expectation. You then use a mathematical relationship, in this case the chi-square distribution, to estimate the probability of obtaining that value of the test statistic.

You calculate the test statistic 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 is between observed and expected. Squaring the differences makes them all positive. You then divide each difference by the expected number, and you add up these standardized differences. The test statistic is approximately equal to the log-likelihood ratio used in the *G*–test. It is conventionally called a "chi-square" statistic, although this is somewhat confusing because it's just one of many test statistics that follows the theoretical chi-square distribution. The equation is:

$$\text{chi}^2 = \sum \frac{(O-E)^2}{E} \tag{2.3.1}$$

As with most test statistics, the larger the difference between observed and expected, the larger the test statistic becomes. To give an example, let's say your null hypothesis is a $3:1$ ratio of smooth wings to wrinkled wings in offspring from a bunch of *Drosophila* crosses. You observe 770 flies with smooth wings and 230 flies with wrinkled wings; the expected values are 750 smooth-winged and 250 wrinkled-winged flies. Entering these numbers into the equation, the chi-square value is 2.13. If you had

observed 760 smooth-winged flies and 240 wrinkled-wing flies, which is closer to the null hypothesis, your chi-square value would have been smaller, at 0.53; if you'd observed 800 smooth-winged and 200 wrinkled-wing flies, which is further from the null hypothesis, your chi-square value would have been 13.33.

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 value and the number of degrees of freedom, you can calculate the probability of getting that value of chi-square using the chi-square distribution. The number of degrees of freedom is the number of categories minus one, so for our example there is one degree of freedom. Using the CHIDIST function in a spreadsheet, you enter **=CHIDIST(2.13, 1)** and calculate that the probability of getting a chi-square value of 2.13 with one degree of freedom is $P = 0.144$.

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 where 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. There are other statistical issues involved in testing fit to Hardy-Weinberg expectations, so if you need to do this, see Engels (2009) and the older references he cites.

Post-hoc test

If there are more than two categories and you want to find out which ones are significantly different from their null expectation, you can use the same method of testing each category vs. the sum of all other categories, with the Bonferroni correction, as I describe for the exact test. You use chi-square tests for each category, of course.

Assumptions

The chi-square of goodness-of-fit assumes independence, as described for the exact test

Examples

Extrinsic Hypothesis examples

Example

European crossbills (*Loxia curvirostra*) have the tip of the upper bill either right or left of the lower bill, which helps them extract seeds from pine cones. Some have hypothesized that frequency-dependent selection would keep the number of right and left-billed birds at a $1 : 1$ ratio. Groth (1992) observed 1752 right-billed and 1895 left-billed crossbills.

Fig. 2.3.1 Male red crossbills, Loxia curvirostra, showing the two bill types.

Calculate the expected frequency of right-billed birds by multiplying the total sample size (3647) by the expected proportion (0.5) to yield 1823.5. Do the same for left-billed birds. The number of degrees of freedom when an for an extrinsic hypothesis is the number of classes minus one. In this case, there are two classes (right and left), so there is one degree of freedom.

The result is chi-square=5.61, $1d.f.$, $P = 0.018$, indicating that you can reject the null hypothesis; there are significantly more left-billed crossbills than right-billed.

Example

Shivrain et al. (2006) crossed clearfield rice, which are resistant to the herbicide imazethapyr, with red rice, which are susceptible to imazethapyr. They then crossed the hybrid offspring and examined the F_2 generation, where they found 772 resistant plants, 1611 moderately resistant plants, and 737 susceptible plants. If resistance is controlled by a single gene with two co-dominant alleles, you would expect a $1:2:1$ ratio. Comparing the observed numbers with the $1:2:1$ ratio, the chi-square value is 4.12. There are two degrees of freedom (the three categories, minus one), so the P value is 0.127; there is no significant difference from a $1:2:1$ ratio.

Example

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, 28% 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=13.59, $3d.f.$, $P = 0.0035$).

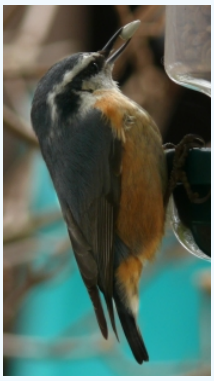

Fig. 2.3.2 Female red-breasted nuthatch, Sitta canadensis.

The expected numbers in this example are pretty small, so it would be better to analyze it with an exact 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; I've had a hard time finding good examples of this.

Intrinsic Hypothesis examples

Example

McDonald (1989) examined variation at the *Mpi* locus in the amphipod crustacean *Platorchestia platensis* collected from a single location on Long Island, New York. There were two alleles, Mpi^{90} and Mpi^{100} and the genotype frequencies in samples from multiple dates pooled together were 1203 $Mpi^{90/90}$, 2919 $Mpi^{90/100}$, and 1678 $Mpi^{100/100}$. The estimate of the Mpi^{90} allele proportion from the data is $5325/11600 = 0.459$ Using the Hardy-Weinberg formula and this estimated allele proportion, the expected genotype proportions are 0.211 $Mpi^{90/90}$, 0.497 $Mpi^{90/100}$, and 0.293 $Mpi^{100/100}$. There are three categories (the three genotypes) and one parameter estimated from the data (the Mpi^{90} allele proportion), so there is one degree of freedom. The result is chi-square=1.08, $1d.f.$, $P = 0.299$, which is not significant. You cannot 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 shouldn't display the result in a graph, as that would be a bar graph with just one bar. Instead, just report the proportion; for example, Groth (1992) found 52.0% left-billed crossbills.

With more than two values of the nominal variable, you should 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.

If you want to add error bars to the graph, you should use confidence intervals for a proportion. Note that the confidence intervals will not be symmetrical, and this will be particularly obvious if the proportion is near 0 or 1.

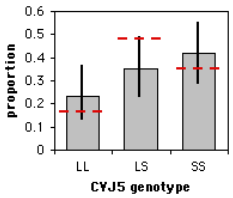

Fig. 2.3.3 Habitat use in the red-breasted nuthatch.. Gray bars are observed percentages of foraging events in each tree species, with 95% confidence intervals; black bars are the expected percentages.

Some people use a "stacked bar graph" to show proportions, especially if there are more than two categories. However, it can make it difficult to compare the sizes of the observed and expected values for the middle categories, since both their tops and bottoms are at different levels, so I don't recommend it.

Similar tests

You use the chi-square test of independence 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 three goodness-of-fit tests: the exact test of goodness-of-fit, the *G*–test of goodness-of-fit,, or the chi-square test of goodness-of-fit. For small values of the expected numbers, the chi-square and *G*–tests are inaccurate, because the distributions 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 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 people had to do statistical calculations 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 as the computationally simpler chi-square or *G*–test, unless the sample size is so large that even computers can't handle it. I recommend that you use the exact test when the total sample size is less than 1000. With sample sizes between 50 and 1000 and expected values greater than 5, it generally doesn't make a big difference which test you use, so you shouldn't criticize someone for using the chi-square or *G*–test for experiments where I recommend the exact test. 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, the G-values are additive; you can conduct an elaborate experiment in which the G-values of different parts of the experiment add up to an overall G-value for the whole experiment. Chi-square values come close to this, but the chi-square values of subparts of an experiment don't add up exactly to the chi-square value for the whole experiment. *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 ability to do more elaborate statistical analyses is one reason some people prefer the *G*–test, 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 use whichever is more commonly used.

Of course, you should *not* analyze your data with both the *G*–test and the chi-square test, then pick whichever gives you the most interesting result; that would be cheating. Any time you try more than one statistical technique and just use the one that give the lowest *P* value, you're increasing your chance of a false positive.

How to do the test

Spreadsheet

I have set up a spreadsheet for the chi-square test of goodness-of-fit chigof.xls . 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 web pages that will perform the chi-square test here and here. None of these web pages lets you set the degrees of freedom to the appropriate value for testing an intrinsic null hypothesis.

R

Salvatore Mangiafico's *R Companion* has a sample R program for the chi-square test of goodness-of-fit.

SAS

Here is a SAS program that uses PROC FREQ for a chi-square test. It uses the Mendel pea data from above. The "WEIGHT count" 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. You must give the expected percentages in alphabetical order: because "smooth" comes before "wrinkled," you give the expected frequencies for 75% smooth, 25% wrinkled.

```
DATA peas;
INPUT texture $ count;
DATALINES;
smooth 423
wrinkled 133
;
PROC FREQ DATA=peas;
WEIGHT count / ZEROS;
TABLES texture / CHISQ TESTP=(75 25);
RUN;
```

Here's a SAS program that uses PROC FREQ for a chi-square test on raw data, where you've listed each individual observation instead of counting them up yourself. I've used three dots to indicate that I haven't shown the complete data set.

```
DATA peas;
INPUT texture $;
DATALINES;
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

To do a power analysis using the 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% 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. If you've done this correctly, your result should be a total sample size of 964.

References

1. Picture of nuthatch from kendunn.smugmug.com.
2. Engels, W.R. 2009. Exact tests for Hardy-Weinberg proportions. Genetics 183: 1431-1441.
3. Groth, J.G. 1992. Further information on the genetics of bill crossing in crossbills. Auk 109:383–385.
4. Mannan, R.W., and E.C. Meslow. 1984. Bird populations and vegetation characteristics in managed and old-growth forests, northeastern Oregon. Journal of Wildlife Management 48: 1219-1238.
5. McDonald, J.H. 1989. Selection component analysis of the *Mpi* locus in the amphipod *Platorchestia platensis*. Heredity 62: 243-249.
6. Shivrain, V.K., N.R. Burgos, K.A.K. Moldenhauer, R.W. McNew, and T.L. Baldwin. 2006. Characterization of spontaneous crosses between Clearfield rice (*Oryza sativa*) and red rice (*Oryza sativa*). Weed Technology 20: 576-584.