Module 1: Comparing expected and observed frequencies within populations. Categorical Data and the G test

Introduction

In this module we will explore statistical approaches for evaluating whether or not categorical data is consistent with our a priori expectations about the frequencies or probability with which various outcomes occur. For instance, we might be interested in evaluating support for the hypothesis that the sex ratio of a population is ½ given a sample of individuals. Or, perhaps we would like to verify our expectation that the genotype frequencies within a population are in Hardy-Weinberg proportions. In such cases, it is possible to evaluate the statistical support for our *a priori* hypothesis using what is known as a *G-test*.

Conceptual/Mathematical Background

Our focus in this module is on categorical data of the form shown in Table 1. The categories under study could be anything, really, as long as it is possible to logically divide the data into *m* discrete bins. Examples of data that naturally lend itself to categorization include counts of genotypes, habitat occupancy, sex, ploidy, or species.

Table 1. Data					
	Category				
Category name	1	2	3	…	m
Observed counts	x_1	x_2	x_3	…	x_m
Observed frequencies	p_1	p_2	p_3	…	p_m

After collecting the data shown in Table 1, we might wish to ask if the counts we observe are consistent with our a priori expectations about the probability with which individuals belong to the various categories. A classic example would be to ask whether or not a random sample containing x_1 males and x_2 females provides evidence that the sex ratio deviates significantly from our a priori expectation of ½.

Before we can proceed to ask whether the data is consistent with our a priori expectations we need to define what, exactly, our expectations are! The place to start is to create a table much like Table 1, but this time with entries corresponding to our expectations rather than observations (Table 2).

Table 2. *A priori* expectations					
	Category				
Category name	1	2	3	…	m
Expected frequencies	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	…	$\hat{p}_m$
Expected counts	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	…	$\hat{x}_m$

So what does it really mean to have "a priori" expectations? Here an example might be most helpful, so we return to the case of investigating sex ratio using counts of males and females. Given a random sample of 50 humans, how many do you expect

to be male and how many female? Most of us would predict 25 male and 25 female, based on our knowledge of sex determination in humans, corresponding to the values $\hat{x}_1 = 25$, $\hat{x}_2 = 25$, $\hat{p}_1 = 0.5$ and $\hat{p}_2 = 0.5$. The next thing we need to do is ask whether the observed counts are consistent with these expectations.

A straightforward method for evaluating whether the data is consistent with our expectations is to employ a G-test. Fortunately, the calculations involved in this test are quite simple and require only that we calculate the test statistic, G:

$$G = 2\sum_{i=1}^{m} x_i ln\left[\frac{x_i}{\hat{x}_i}\right] \qquad (1)$$

where once again, x_i is the number of individuals observed in category i and $\hat{x}_i$ is the number of individuals expected to be in category i. This test statistic follows a χ^2 distribution with m-1 degrees of freedom. Thus, in order to test whether the data we observe are consistent with the frequencies we expect, we compare G to the appropriate value in the χ^2 table; if G exceeds the critical value from the χ^2 table, the data is not consistent with our expectations.

In most cases, *a priori* expectations for frequencies of categories will be known and independent of the data. In some special cases, however, developing *a priori* expectations may require that we first estimate a quantity from the data. For instance, if we are given counts of diploid genotypes (e.g., 25 AA, 50 Aa, 25 aa), and our goal is to evaluate whether the genotype differ significantly from Hardy-Weinberg expectations, we must first use the data to estimate the frequency of the A allele. In cases like these, one degree of freedom is lost because we have used it to estimate the allele frequency. Thus, for Hardy-Weinberg problems, the correct degrees of freedom is (m-1)-1 = 1.

Assumptions and Limitations

- You must have *a priori* expectations for the frequencies of categories
- No category should have less than five observations

WORKED EXAMPLE

Problem:

You are studying a population of wild pronghorn on the National Bison Range as part of your work with the USFWS. Your supervisor insists that the sex ratio is becoming biased toward males and argues that the best course of action is to relocate some of the males to reduce pressure for food. Being somewhat skeptical of this claim, you counted the number of male and female pronghorn within a 10km^2 study plot. Your survey revealed 14 male pronghorn and 11 female pronghorn. Does

this data support your supervisor's claim that the sex ratio deviates significantly from ½?

Solution:

If we wish to use the data described above to establish that the sex ratio of pronghorn deviates significantly from ½, we can do so using formula (1). The first step in using this formula is to identify and calculate all of the quantities appearing in the formula. Let's start with *m*, which is the number of categories to which the data can possibly belong. In this case, the only possible categories are male and female so *m*=2. Next, let's expand formula (1) appropriately given that we now know that *m*=2, and so know how many times to iterate the summation:

$$G = 2\left(x_1 ln\left[\frac{x_1}{\hat{x}_1}\right] + x_2 ln\left[\frac{x_2}{\hat{x}_2}\right]\right)$$

With the formula now expanded, we see that we need to decide which values to assign to four different symbols: x_1, x_2, $\hat{x}_1$, and $\hat{x}_2$. The easiest of these to figure out are x_1 and x_2 which are simply the number of males and females, respectively, observed in our sample. Thus, we set $x_1 = 14$ and $x_2 = 11$. Note that these assignments are arbitrary and it would make no difference if we used x_1 to represent the number of females and x_2 to represent the number of males, as long as we were consistent throughout. Next, we need to find the values for $\hat{x}_1$ and $\hat{x}_2$which is a bit trickier. These values are the number of males and females we would expect to observe **IF** the null hypotheses of an equal sex ratio were true. Because in this case we have a total of 25 pronghorn, we would expect to see $\hat{x}_1 = 12.5$ males and $\hat{x}_2 = 12.5$ females if the sex ratio were truly equal. Now, obviously there can't actually be half pronghorns out there wandering around, but this is not a problem for the formula itself and our calculations will turn out just fine so don't worry too much about fractional pronghorn. Since we now know what all the values are, we can plug them into the expanded equation and evaluate the expression as follows:

$$G = 2\left(14 \times ln\left[\frac{14}{12.5}\right] + 11 \times ln\left[\frac{11}{12.5}\right]\right) = 0.360869$$

The final step is to ask whether the value of the test statistic, G, we just calculated corresponds to a p value of less than 0.05 using the table of χ^2 values and (m-1) = 1 degrees of freedom. Inspecting the table of χ^2 values reveals that the critical value for the test statistic G equals 3.841, for our specified p value of 0.05 and our (m-1) = 1 degrees of freedom. Because the value of G we calculated is less than this critical value, we conclude the data does not support the supervisors claim that the sex ratio in this population of pronghorn deviates significantly from ½. There does not appear to be any reason to begin relocating males.

Appendix Table. Parameters and their definitions	
Parameter	Definition
m	The number of categories
n	Sample size
x_i	The number of individuals observed in category i
p_i	The frequency of individuals observed in category i
$\hat{x}_i$	The number of individuals expected to be in category i
$\hat{p}_i$	The frequency of individuals expected to be in category i