Categorical Data Analysis

So far we’ve been looking at continuous data arranged into one or two, where each “group” has more than one observation.

- e.g., a series of measurements on one or two “things”.

Now we’re changing things: we’re interested in data that is categorical. The data is often just counts that these different categories have.

- for example:

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of people with blood type A:</td>
<td>$y_1$</td>
</tr>
<tr>
<td>Number of people with blood type B:</td>
<td>$y_2$</td>
</tr>
<tr>
<td>Number of people with blood type AB:</td>
<td>$y_3$</td>
</tr>
<tr>
<td>Number of people with blood type O:</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

- if we had some theory about the distribution of blood types, how would we be able to test it?

- For example, we have reason (somehow) to believe that 34% of people have blood type A, 15% blood type B, 23% blood type AB, and 28% blood type O.

- we go out and collect a sample of 100 people, and find the following:

- A: 12    B: 56    AB: 2    O: 30

- is our result compatible with our hypothesis??

III. The Chi-square ($\chi^2$) goodness of fit test.

- Essentially, we have a number of categories for our data, and we have some idea as to the “proportion” or frequency of “things” we expect in each of our categories.

A) Here’s another example (from genetics):

1) We have a simple dominant-recessive relationship. Say, yellow and purple corn kernel color. Purple is dominant, and yellow is recessive.

   Without going into the details, if we have two heterozygous parents, for our offspring we would expect 3/4 of our kernels to be purple and 1/4 yellow

   - (if you don't understand the genetic terminology here, you'll get the details in genetics - it's not necessary to understand the details here).

2) You go out and collect a sample of 267 corn kernels.
i) how many do you expect to be purple?

\[ \frac{3}{4} \times 267 = 200.25 \]

ii) yellow?

\[ \frac{1}{4} \times 267 = 66.75 \]

*(notice - take the proportion you expect, and multiply this by the total number in your sample to get what you expect in your sample)*

3) Now you can compare this to what you actually got. Suppose you took your sample and got:

157 purple \hspace{1cm} 110 yellow

4) Looking at this, you would think you should have gotten more yellows and less purples. But is this due to random chance?

5) Set up your hypotheses:

\[ H_0: \Pr\{\text{purple}\} = .75, \Pr\{\text{yellow}\} = .25 \]

(your \( H_0 \) is different from what you’re used to - we’ll learn more soon)

\[ H_1: \text{At least one probability listed in } H_0 \text{ is incorrect (see your text for another way of phrasing this).} \]

6) Decide on \( \alpha \) (let’s pick .05)

7) Calculate your test statistic. This is now \( \chi^2 \) (or \( \chi^2_\text{s} \), as your text calls it):

\[ \chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} \]

- \( i \) goes from 1 to \( c \), where \( c \) is the number of categories (2 in our example).

- for our example we have:

\[ \chi^2 = \frac{(157-200.25)^2}{200.25} + \frac{(110-66.75)^2}{66.75} = 37.36 \]

8) Compare this to the tabulated \( \chi^2 \) with \( c-1 \) degrees of freedom and the appropriate level of alpha.

a) the \( \chi^2 \) distribution.

- the value you calculate, \( \chi^2 \), will follow a \( \chi^2 \) distribution if \( n \) is moderately large.
(notice that the $\chi^2$ test is based on an approximation, just like the KW test. You can get exact values, but they’re a bit of a pain.)

- Just like many other distributions, the value of the $\chi^2$ distribution depends on d.f. (or $\nu$).

- here is what it looks like for a couple of different values of $\nu$:

![Chi-square densities for different values of $\nu$](image)

- just like before, we reject for values that wind up in the tails (usually only in the upper tail).

b) If $\chi^2 \geq \chi^2_{\text{table}}$, we reject our $H_0$.

- here’s our comparison:

$$\chi^2 = 37.36 \geq \chi^2_{\nu-1} = 3.84$$

so we reject our $H_0$ and conclude that at least one of our proportions is not as specified in $H_0$.

- important point - notice that with just two categories, if one of our proportions is wrong, that immediately implies that the other one is wrong as well (why??).
B) Some comments:

a) except in the case of two categories, the alternative hypothesis is non-directional.

b) in the two category case you could specify a directional alternative as follows:

\[ H_0: \text{Pr\{Male\}} = .6 \quad \text{(and therefore Pr\{Female\} = .4)} \]

\[ H_1: \text{Pr\{Male\}} > .6 \quad \text{(so what are females?)} \]

- you should be fairly comfortable with this concept by now.

C. Two examples of the \( \chi^2 \) test:

1) Exercise 10.1 from p. 392 \( \{10.1, \text{p. 399}\} \) \( \{9.4.1, \text{p. 357}\} \) \( \{9.4.1, \text{p. 377}\} \):

a) Geneticists propose that the color of summer squash should follow a 12:3:1 ratio. Researchers collected the following data:

   white: 155  yellow: 40  green: 10

b) \( H_0: \) Pr\{white\} = .75 \quad \text{(12+3+1=16, so 12/16 = .75)}

   Pr\{yellow\} = .1875

   Pr\{green\} = .0625 \quad \text{(we didn’t have to specify green - why not?)}

\[ H_1: \] at least one of these proportions is not as specified.

c) \( \alpha = .10 \)

d) Calculate our expected values:

\[ .75 \times 205 = 153.75 \quad \text{(our total sample size is 205)} \]
\[ .1875 \times 205 = 38.4375 \]
\[ .0625 \times 205 = 12.8125 \]

e) Calculate \( \chi^2* \):

\[ X^2* = \left(\frac{155-153.75}{153.75}\right)^2 + \left(\frac{40-38.4375}{38.4375}\right)^2 + \left(\frac{10-12.8125}{12.8125}\right)^2 = 0.691 \]

f) Our \( \chi^2 \) table:

\[ X^2_{2.1} = 4.61 \]

g) Because \( \chi^2* \) is less than our \( \chi^2 \) table, we “fail to reject”, and conclude that our null hypothesis is consistent with the data:

\[ We \ have \ no \ evidence \ to \ show \ that \ summer \ squash \ does \ not \ follow \ a \ 12:3:1 \ ratio. \]
2) Color vision in squirrels [exercise 10.9, p. 401] [9.4.9, p. 358] [9.4.9, p. 378]. A squirrel was exposed to a red panel and two white panels. By pressing the red panel, the squirrel was rewarded; no reward was given for pressing the white panel. In 75 trials, the squirrel correctly pressed the red panel 45 times. Can the squirrel see color?

\[ H_0: \Pr\{\text{red}\} = 1/3 \quad (\text{so } \Pr\{\text{white}\} = 2/3) \]

\[ H_1: \Pr\{\text{red}\} \neq 1/3 \]

choose \( \alpha = .02 \) (book says to use this).

calculate our expected:

- for red, \( 75 \times 1/3 = 25 \)
- for white, \( 75 \times 2/3 = 50 \)

(also note that the proportion of red is .6, so if we had gone with our directional alternative, it would have made sense).

calculate our \( \chi^2^* \):

\[
\chi^2^* = \frac{(45-25)^2}{25} + \frac{(30-50)^2}{50} = 24
\]

our \( \chi^2^\text{table} \) is:

\[
\chi^2_{1,0.02} = 5.41
\]

Obviously, since \( \chi^2^* \gg \chi^2\text{table} \), we get to reject and conclude that squirrels can see the color red.

BUT, is a two sided test appropriate here? Don't squirrels like food? What kind of squirrel doesn't like food?

So really we should have done a one sided test. Our squirrel is going to do it's best to get at the food, so if it can see red, the proportion of red should be higher than 1/3. We really should have done:

\[ H_0: \Pr\{\text{red}\} = 1/3 \quad H_1: \Pr\{\text{red}\} > 1/3 \]

As usual, the math doesn't change. We need to do two things:

1) Make sure the data agree with \( H_1 \):

\[
1/3 \times 75 = 25, \text{ and observed for red is 45; that is bigger than 25, so we're okay here.}
\]

2) Use the one sided value for \( \chi^2\text{table} \):
\[ X^2_{1,0.025} = 3.84 \]  
\text{(the tables don't give us } \alpha = .02 \text{ for one sided values, so we used } \alpha = .025) 

Note that again, the one sided table value is lower than the two sided table value (even though the \( \alpha \)'s didn't quite match), so again, a one sided test has more power.

D. Some assumptions:

- the data are collected randomly (you just can’t get away from this one!)

- The smallest expected \# is at least 5.

- the chi-square test is an approximation, and approximations get better the bigger the sample size.

- Suppose we take an ear of corn and count kernels. We're very lazy, so we only count 16 kernels. Mendelian theory predicts that for this particular ear of corn we should have the following ratio:

16 purple smooth: 3 yellow smooth: 3 purple wrinkled: 1 yellow wrinkled

So if we only count 16 kernels, what are our expected values?

What would be the minimum sample size we need to make sure the smallest expected value is at least 5?

Sometimes we can easily adjust our sample size to meet our assumptions of the smallest expected value being at least 5.

- if each of your categories doesn’t have an expected value of at least 5, then your approximation can be pretty awful.

- there are other techniques for dealing with smaller sample sizes, but we won’t learn them here.