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:

Number of people with blood type A: y_1
Number of people with blood type B: y_2
Number of people with blood type AB: y_3
Number of people with blood type O: y_4

- 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 (χ^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?

3/4 x 267 = 200.25

ii) yellow?

1/4 x 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 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{purple} = .75, Pr{yellow} = .25

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

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

6) Decide on α (let's pick .05)

7) Calculate your test statistic. This is now χ^{2*} (or χ^2_s, as your text calls it):

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

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

- for our example we have:

$$X^{2*} = \frac{(157-200.25)^2}{200.25} + \frac{(110-66.75)^2}{66.75} = 37.36$$

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

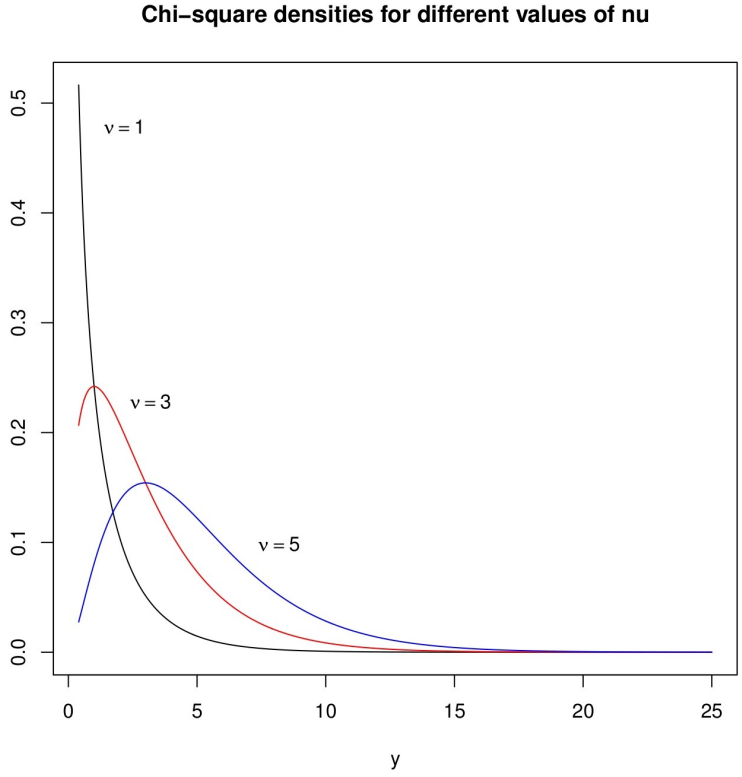
a) the χ^2 distribution.

- the value you calculate, χ^{2*}, will follow a χ^2 distribution if n is moderately large.

(notice that the χ^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 χ^2 distribution depends on d.f. (or ν).

- here is what it looks like for a couple of different values of ν:



- 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:

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

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: Pr{Male} = .6 (and therefore Pr{Female} = .4)

H_1: Pr{Male} > .6 (so what are females?)

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

- if you really want to do a one sided test, you might want to use the binomial test (it does a bit better, but we don't have time to talk about it).

C. Two examples of the χ^2 test:

1) Exercise 10.1 from p. 392 **[10.1, p. 399]** *{9.4.1, p. 357}*:

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 (12+3+1=16, so 12/16 = .75)
Pr{yellow} = .1875
Pr{green} = .0625 (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 x 205 = 153.75 (our total sample
.1875 x 205 = 38.4375 size is 205)
.0625 x 205 = 12.8125

e) Calculate χ^{2*}:

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

f) Our χ^2_{table}:

$$X^2_{2,.1} = 4.61$$

g) Because χ^{2*} is less than our χ^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}*. 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{red} = 1/3 (so Pr{white} = 2/3)

H_1: Pr{red} ≠ 1/3 Incidentally, a squirrel is going to do the best it can to get food, so the alternative probably should be one sided here (H_1: Pr{red} > 1/3).

choose α = .02 (book says to use this).

calculate our expected:

for red, 75 x 1/3 = 25
for white, 75 x 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 χ^{2*}:

$$X^{2*} = \frac{(45-25)^2}{25} + \frac{(30-50)^2}{50} = 24$$

our χ^2_{table} is:

$$X^2_{1,0.02} = 5.41$$

Obviously, since $\chi^{2*} >> \chi^2_{table}$, we get to reject and conclude that squirrels can see the color red.

- We're pretty much done. But if we wanted to get an approximate p-value, here's how:

  - look in table 8 **[9]** *{9}* until we get to the closest number that is less than our critical value.

  - In the table, for one *d.f.*, this gives us 15.14, and an associated $p < .0001$.

Bottom line: we are very confident that squirrels can see the color red.

- remember, you're really just interested in the minimum possible *p*-value, so don't bother with all the bracketing stuff you see in your text.

Incidentally, if you wanted to do a one sided test, you'd just use the appropriate column in your χ^2 tables, but divided by 2 (i.e., $\alpha/2$). For our squirrel example, you'd get:

$X^2_{1,0.025} = 3.84$ (the tables don't give us $\alpha = .04$, so we used the column for .05)

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.

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

  - there are other techniques (e.g., Fisher's exact test) for dealing with smaller sample sizes, but we won't learn them here.