The goodness of fit test

So far we've been looking at continuous data that were arranged in one or two groups. Each of our "groups" had more than one observation, or measurement, on something. For example, we might have had 10 data values for height in men, or 13 values for systolic blood pressure in women. Everything we've done so far dealt with some kind of measurement. Now we are changing our data type, and looking at data that can be sorted into categories.

Let's suppose we're interested in blood types. We go out and collect the following data from 263 people:

Blood type	Number of people
A	115
B	20
AB	17
O	111

Note that the data are *counts*, not *measurements* (we didn't measure anything).

Now suppose we had some idea about the distribution of blood types. For example, in the U.S., people have the following proportions for blood type:

Blood type	Percent
A	42
B	10
AB	4
O	44

Are our data compatible with the above idea? To answer this question we will need to do a little bit of math first. For instance, how can we compare our sample of $n = 263$ with the above percentages?

First we need to figure out how many people we *expect* for each blood type if we have 263 people:

For example, if we have 263 people, how many people do we expect with blood type A? If 42% of people have blood type A, this is a simple percentage problem:

$$42\% \text{ of } 263: \ 0.42 \times 263 = 110.46$$

(Fractional or decimal values are fine for expected values, even if it seems a little silly to talk of 110.46 people).

We can do the other expected values the same way:

$$0.10 \times 263 = 26.30 \text{ for blood type B}$$
$$0.04 \times 263 = 10.52 \text{ for blood type AB}$$
$$0.44 \times 263 = 115.72 \text{ for blood type O}$$

And now we can directly compare our observed values with our expected values:

Blood type	Observed	Expected
A	115	110.46
B	20	26.30
AB	17	10.52
O	111	115.72

Now we need to put this into some sort of testable hypothesis. We'd like to know if the observed results are consistent with the expected results. This is *backwards* from what we usually do - if the data are consisted with the expected values we will be going with the null hypothesis!

So let's write down a hypothesis. Our null hypothesis will be a list of expected probabilities or proportions. Let's do the following:

$H_0 : Pr\{\text{blood type A}\} = 0.42,$

$Pr\{\text{blood type B}\} = 0.10,$

$Pr\{\text{blood type AB}\} = 0.04,$

$Pr\{\text{blood type O}\} = 0.44.$

And our alternative hypothesis becomes:

H_1 : at least one of these proportions is incorrect.

Our H_0 consists of multiple parts which is different from what we've done up until now. You need to list *each* expected outcome (using probability or proportion).

Let's pick $\alpha = 0.01$. Now we need to figure out how to actually do our test - what is our test statistic and what do we calculate? The test statistic is given as follows:

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

Let's figure out what it all means:

c = the number of categories.

i = the index for each category (goes from 1 to c).

O_i = the observed value for each category.

E_i = the expected value for each category.

So for our blood type example, we have:

$$\chi^{2^*} = \frac{(115 - 110.46)^2}{110.46} + \frac{(20 - 26.30)^2}{26.30} + \frac{(17 - 10.52)^2}{10.52} + \frac{(111 - 115.72)^2}{115.72} = 5.88$$

Now that we've calculated our test statistic we need to compare it to the value we get from the χ^2 tables.

How many degrees of freedom do we have? $D.f. = \nu = c - 1$, so we have $4 - 1 = 3$ $d.f.$ in our example. So we have:

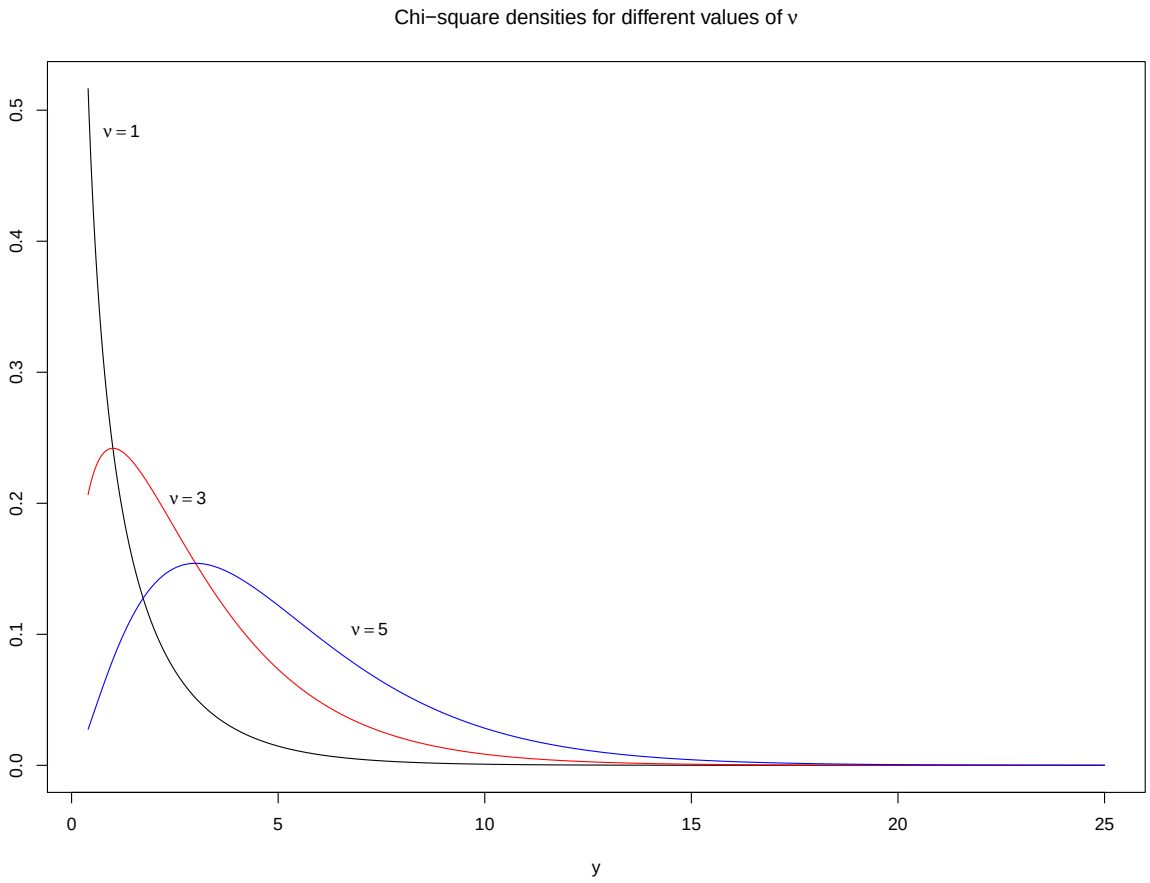
$$\chi^2_{table} = \chi^2_{3,0.01} = 11.34$$

Our comparison is the same as always: If $\chi^{2^*} \geq \chi^2_{table}$ we reject H_0.

In this case our χ^{2^*} is less than the table value so we *fail to reject* H_0 and conclude that our data are consistent with the null hypothesis.

Notice that in this case we're actually trying to go with the null hypothesis, which is a bit unusual. Usually we are trying to reject the null hypothesis. We still can't "accept" the null hypothesis, but we can try to say something a bit better than "fail to reject" by saying that the data are consistent with the null hypothesis (which is what we did).

So what about the χ^2 distribution? Let's take a quick look at this distribution:



The important thing is that the degrees of freedom can have a strong effect on the appearance of the χ^2 distribution. This is one reason why you really need to be careful to use the correct degrees of freedom. Fortunately, as you saw above, this is usually not difficult to calculate.

Let's do another example, this time from genetics where the goodness of fit test is used often. Mendelian theory predicts that for corn (assuming heterozygous parents) we should get three purple kernels for every one yellow kernel in the offspring. In other words, when we look at an ear of corn we should have:

3 purple : 1 yellow

We count the kernels in an ear of corn and get the following result:

purple: 157

yellow: 110

Which gives us a total of 267 corn kernels.

So let's write down our hypotheses:

H_0 : $Pr\{\text{purple}\} = 0.75$

H_1 : $Pr\{\text{purple}\} \neq 0.75$

We only have two categories, so we can actually write our hypotheses this way. We *can't* do this if we have more than two categories. We can also do a one sided test if we have only two categories; more on this below. Let's continue:

$\alpha = 0.05$

Let's calculated our expected values:

$\text{purple} = 0.75 \times 267 = 200.25$

$\text{yellow} = 0.25 \times 267 = 66.75$

Now we can calculate our value for χ^{2^*}:

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

And we compare this to our table value using $d.f. = \nu = c - 1 = 2 - 1 = 1$:

Since $\chi^{2^*} = 37.36 \geq \chi^2_{table} = 3.841$ we *reject* the H_0.

We conclude that our data indicate that the 3:1 ratio is not correct.
Let's think about this problem a bit more. Suppose we had noticed a bunch of purple kernels lying in the bottom of the box that had the ear of corn in it. Does that change anything (assuming we didn't want to pick them all up and try to count them)? What about our alternative hypothesis?

We would expect that our ear of corn would have *less* purple kernels than expected since a lot of them fell off. This implies a *one sided* alternative hypothesis:

H_1 : $Pr\{\text{purple}\} < 0.75$

So how do we do a one sided goodness of fit test? Pretty much as you would expect. All the math stays the same, but now (in addition to modifying H_1) you need to do two things:

Verify that your data agree with H_1: In this case we expected 200.25 purple kernels but only got 157, so yes, the proportion of purple is less than 0.75, which agrees with H_1.

Use the one sided χ^2_{table} value: χ^2_{table} (one sided) = 2.706.

And again we can conclude that since $\chi^{2^*} = 37.36 \geq \chi^2_{table} = 2.706$ we *reject* the H_0. The fact that the value of χ^2_{table} decreased from 3.841 to 2.706 should tell you that we do have more power with a one sided test, which shouldn't be a surprise.

So what about the assumptions of the χ^2 goodness of fit test? There are only two that we need to worry about:

1. The data are random.
2. The smallest expected value is greater than or equal to 5 ($Exp.\ Val. \geq 5$).

Let's talk about the second assumption a bit. First, notice that this applies to the *expected values* not the observed values. We don't care what the observed values are.

Why is this assumption important? Because the test statistic, χ^{2^*} will have a χ^2 distribution if our expected values are large enough. If they are not, then our calculated value won't have a χ^2 distribution and we can't use the χ^2 tables.

Sometimes we can figure out what sample size we need ahead of time to get our smallest expected value to be ≥ 5. Suppose for example that you were trying to establish if the kernels on an ear of corn follow a 9:3:3:1 ratio. You count kernels, but since you're in a rush, you only count 16 kernels (incidentally, note that $9+3+3+1=16$). What are your expected values?

For the category with 9, we expect: $9/16 \times 16 = 9$

For the category with 3, we expect: $3/16 \times 16 = 3$

For the category with 3, we expect: $3/16 \times 16 = 3$

For the category with 1, we expect: $1/16 \times 16 = 1$

Three of our categories have expected values less than 5, so we're obviously violating our assumption. How can we fix it? We can calculate the smallest sample size we need. The expected value for the smallest category is 1. To get 5 times as many kernels (so our smallest expected value is 5) we can multiply our sample size by 5. If you counted 80 kernels ($16 \times 5 = 80$) we get:

For the category with 9, we expect: $9/16 \times 80 = 45$

For the category with 3, we expect: $3/16 \times 80 = 15$

For the category with 3, we expect: $3/16 \times 80 = 15$

For the category with 1, we expect: $1/16 \times 80 = \ \ 5$

This is much better. Of course, the other thing to remember is that larger sample sizes always do better.