Inference for Categorical Variables

We've seen that you can turn a qualitative variable into a quantitative one (by counting the number of successes and failures), but that's a compromise—it forces us into a very binary existence. Life is so much more varied than that! If only there was a way to keep all of those categories…

Maybe there is—but not with confidence intervals. Confidence intervals attempt to estimate a single parameter; a number. We don't have that with a categorical variable. So; no confidence intervals here.

A hypothesis test tries to determine if an observed result could reasonably be expected; to determine if there is a significant difference between what was observed and what was expected. Perhaps that might work. We'd just need to find some way to measure the difference between an observed distribution and an expected distribution…

The Problem

Measuring the Difference in Distributions

Why don't we try? Let's take a simple categorical variable, like favorite (primary) color. Perhaps the observed data might look like this:

Color	*Red*	*Yellow*	*Blue*
Observed	15	10	20

The expected distribution might look like anything; probably the most popular version might be of no preference—all values are equally likely (all colors are equally preferred). In that case, the expected distribution can be found by dividing the sample size by the number of categories.

Color	*Red*	*Yellow*	*Blue*
Expected	15	15	15

Now—how can we distill the difference in these distributions into a single number? Perhaps we could start by finding the differences for each category.

Color	*Red*	*Yellow*	*Blue*
Difference	0	-5	5

What now? We've got a bunch of deviations (differences from the expectation) that we want to condense into a single number. Have we done that before?

Yes! Standard deviation! There, we added the deviations. So, let's sum up our deviations.

Sum of Deviations = 0.

Well, that's perhaps a bit counter-intuitive. Like standard deviation, it would make sense that a value of zero would mean **no** difference…what did we do to avoid this "meaningless zero" problem when we did standard deviation?

We squared then before adding! Let's do that.

Sum of Squared Deviations = 50.

There is another problem—not all differences are equally significant. Consider the following observed and expected distributions:

Color	*Red*	*Yellow*	*Blue*
Observed	150	95	205

Color	*Red*	*Yellow*	*Blue*
Expected	150	100	200

Our measure of difference is the same; but are these differences really as "big" as those in the original example? 5 out of 15 is a much bigger difference than 5 out of 200! So how can we take *that* into consideration?

How about dividing by the expected value, to make some kind of proportion?

Sounds good! We have now created a new statistic, called Chi Square.

Equation 1 - Chi-Square

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Well, actually, that's a *parameter*—if we had all the data, then we could calculate the value of this Chi-Square parameter (and no, that doesn't mean that we're going to construct a confidence interval—the parameter isn't a "real" value that has intrinsic value or meaning. It's an abstract measure that we're going to use to measure a difference.).

When we use data from a sample, then we've got an actual statistic—which we usually denote X^2. Often, though, we don't make a distinction.

The Chi-Square Distribution

So, we've got a new variable. I wonder what its distribution looks like?

Well, it depends. The size of the sample determines how unusual a particular difference (value of χ^2) is—for small samples, it doesn't take much of a difference to be quite unusual; for large samples, it takes a really big difference. Thus, the chi-square distribution has degrees of freedom (just like *t*). So—here are some chi-square distributions for various degrees of freedom (there isn't a single formula for calculating *df*—more on that later).

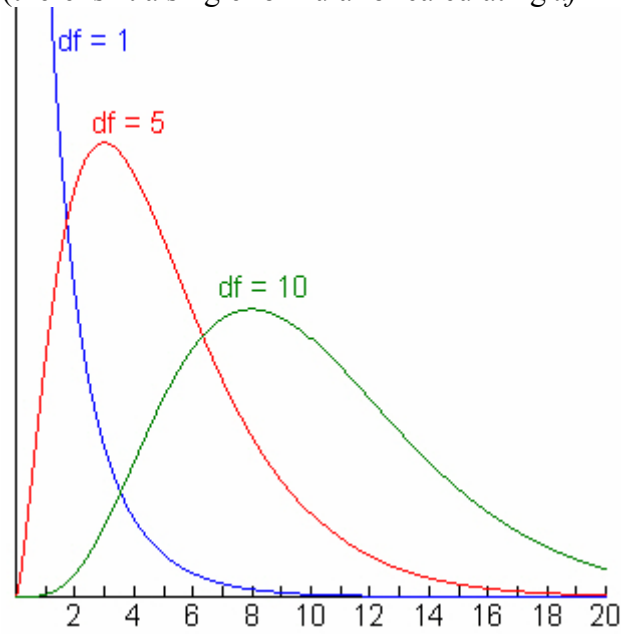

Figure 1 - Various Chi-Square Distributions

As you can see, the distribution is skew right—although it becomes less so as the degrees of freedom get larger.

To calculate the area under the chi-square curve, you can use a table (which works exactly like the *t* table!), or you can use the CDF function on your calculator (which *also* works exactly like the tcdf function!). It is almost always the case that you will want to calculate the right-hand area for a chi-square (*the probability of observing a difference as large, or larger…*).

Examples

[1.] $P(\chi^2 > 4.219)$ when $df = 3$?
0.2392 (from the table: between 0.20 and 0.25)

[2.] $P(\chi^2 > 17.214)$ when $df = 9$?
0.0455 (from the table: between 0.025 and 0.050)

[3.] $P(\chi^2 > 13.238)$ when $df = 14$?
0.5079 (from the table: greater than 0.25)

The Chi-Square Tests

Goodness of Fit

Indications

This is what we used to develop the idea. A Goodness of Fit test is used when you are measuring a single categorical variable in a single population.

Requirements

First of all, we need a random sample from the population (technically, you can also use this for populations—but that shouldn't be foremost in your mind).

Second, we need for all of the expected values to be at least one. Think about it—in the formula for chi-square, we divide by the expected value. If you divide by a number smaller than one, what happens?

The result is large—larger than the dividend. So very small expected values skew the chi-square that we calculate, and make probability statements (p-value) worthless.

Finally, we need most (at least 80%) of the expected values to be at least 5. This test expands our work with proportions—what was the requirement for z tests for proportions?

np and $n(1-p)$ each had to be at least 10 (or 5). Well, np is the expected number of successes, and $n(1-p)$ is the expected number of failures. It only makes sense, then, that we want our expected values to be at least 5—the only new part is that *most* must be at least 5.

Expected Values

For the Goodness of Fit test, an expected distribution must be given. The most common is that there is no preference / no difference in the categories. In this case, divide the sample size by the number of categories to obtain the expected values.

Other expected distributions are possible—for example, from the work with primary colors above, perhaps we expect 40% to prefer red, 20% to prefer yellow, and 40% to prefer blue. Simply apply these percentages to the sample size to obtain the expected values.

DO NOT ROUND! These are expected values—it is OK for them to have decimal values! Only the observed values must be integers (since they are counts).

Degrees of Freedom

For the Goodness of Fit test, the degrees of freedom are $n - 1$.

Example

[4.] A professor, concerned about the number of blue M-Ms in his hand, decided to collect some data. Here are the observed counts:

Table 1 - Obeserved M-M Counts

Color	Brown	Yellow	Red	Orange	Green	Blue
Number	152	114	106	51	43	43

Here is the expected distribution (at that time):

Table 2 - Expected Color Distribution

Color	Brown	Yellow	Red	Orange	Green	Blue
Percent	30	20	20	10	10	10

Is the distribution as the company claims, or is it off?

H_0: the distribution of color is as the company claims

H_a: the distribution is not as the company claims

This calls for a Chi-Square test for Goodness of Fit. There are a few requirements that must be met to conduct this test.

(a) the sample must be a random sample from the population. This is not given, and will be assumed.

(b) all expected values must be at least 1, and most (at least 80%) must be at least 5. Here is a table of the expected values:

Table 3 - Expected Counts for Example 4

Color	Brown	Yellow	Red	Orange	Green	Blue
Number	152.7	101.8	101.8	50.9	50.9	50.9

All are at least 5; the requirement is met.

I'll choose a 1% level of significance (I want a lot of evidence before I go and accuse the company of lying!).

$X^2 =$

$$\frac{(152-152.7)^2}{152.7}+\frac{(114-101.8)^2}{101.8}+\frac{(106-101.8)^2}{101.8}+\frac{(51-50.9)^2}{50.9}+\frac{(43-50.9)^2}{50.9}+\frac{(43-50.9)^2}{50.9}=$$

4.091. $df = 6 - 1 = 5$.

$P(X^2 > 4.091) = 0.5364$.

If the distribution is as the company claims, then I can expect to find a sample distribution this far (or farther) from the expected distribution in about 53.64% of samples. This happens often

enough at the 1% level of significance to attribute to chance; it is not significant, and I fail to reject the null hypothesis.

It appears that the company is correct.

Homogeneity of Proportions

Indications

The Homogeneity of Proportions test is used if a single categorical variable is measured in two or more populations. Typically, the variable will be listed horizontally, and each population will occupy one row (giving a two-dimensional table for the data).

Requirements

The requirements for this test are the same as for the Goodness of Fit test.

Expected Values

No expected distribution will be given for this test. The expected values are calculated by applying the percentage of the sample in each row to each column total—e.g., if 20% of the sample is contained in row one, then 20% of the number in column one are in row one, column one; 20% of the number in column two are in row one, column two; etc.

This can be summarized in the following formula:

Equation 2 - Expected Values

$$\text{expected value for cell (i, j)} = \frac{(\text{total in row i})(\text{total in column j})}{\text{total sample size}}$$

Degrees of Freedom

The degrees of freedom are (rows – 1)(columns – 1).

Example

[5.] Do Republicans, Democrats and Independents think alike about the primary system in the U.S.? Each group was asked if all primaries should be held in June of presidential election years. 718 Republicans were asked; 734 Democrats were asked; and 76 Independents were asked. The results are shown below:

Table 4 - Political Party and Primary Preference

	Good Idea	Poor Idea	No Opinion
Republican	266	266	186
Democrat	308	250	176
Independent	28	27	21

Is there evidence that these groups think alike on this issue?

H_0: the opinions of the groups are alike

H_a: the opinions of the groups are different

This calls for a Chi-Square test for Homogeneity of Proportions. There are some requirements to conduct this test.

(a) the samples must be random samples from the respective populations. This is not give; it will be assumed.

(b) all expected values must be at least 1; and most (>80%) must be at least 5. The expected values are shown below:

Table 5 - Expected Counts for Example 5

	Good Idea	Poor Idea	No Opinion
Republican	282.8769634	255.1531414	179.9698953
Democrat	289.1806283	260.8390052	183.9803665
Independent	29.94240838	27.0078534	19.04973822

All are at least 5; the requirement is met.

I'll choose a 10% level of significance (I don't need much evidence to believe that these groups think differently).

$$X^2 = \frac{(266-282.9)^2}{282.9} + \frac{(266-255.2)^2}{255.2} + \frac{(186-180.0)^2}{180.0} + \frac{(308-289.2)^2}{289.2} + \frac{(250-260.8)^2}{260.8} + \frac{(176-184.0)^2}{184.0} + \frac{(28-29.9)^2}{29.9} + \frac{(27-27.0)^2}{27.0} + \frac{(21-19.0)^2}{19.0} = 4.017.\ df = (3-1)(3-1) = 4.$$

$P(X^2 > 4.017) = 0.4037$.

If these groups have the same opinions on primaries, then I can expect to find a sample distribution this far or farther from the expected distribution in about 40.37% of samples. This occurs frequently enough to attribute to chance at the 10% level; it is not significant, and I fail to reject the null hypothesis.

It appears that these groups do think alike on this issue.

Independence of Variables

Indications

A Chi-Square test for Independence of Variables is indicated when two categorical variables are measured in a single population. One variable will be labeled horizontally, and the other vertically, producing a table of values.

The difference between IOV and HOP is subtle; pay close attention! For HOP, I ask *one question* in *many groups*; in IOV, I ask *two questions* in *one group*.

The previous example was HOP because we asked one question (Opinion on One-Month of Primaries) in several populations (Reps, Dems and Inds).

Assume, for a moment, that the previous example had been conducted differently. In particular, assume that a random sample of voters (*one group*) had been asked "What is your party affiliation" and "What do you think about having all primaries in one month?" (*two questions*) We could obtain the exact same results, but the type of test would be different. *That's subtle*. **Pay close attention**.

Requirements

The requirements are (thankfully) the same.

Expected Values

The expected values are the same as the HOP case.

Degrees of Freedom

The degrees of freedom are the same as the HOP case.

Now, some of you are wondering, "If everything's the same except for the name, why have two tests?"

The answer is: I don't know, and it doesn't matter (right now). The AP Committee feels that the difference is significant (some textbook authors agree; others do not), so you need to know the difference. If you continue your studies in statistics, perhaps you'll find out why two are needed (or at least used).

Example

[6.] Gene Siskel was, and Roger Ebert is (as of 2004) a movie critic in Chicago. The pair became famous for their "Thumbs up / Thumbs down" reviews. Alan Agresti and Larry Winner wondered if the pair's opinions were related in some way. Here are the observed results for 160 movies:

Table 6 - Siskel and Ebert Reviews

		Ebert		
		Down	**Mixed**	**Up**
Siskel	**Down**	24	8	13
	Mixed	8	13	11
	Up	10	9	64

Does there appear to be any connection between their opinions?

H_0: Siskel's and Ebert's opinions are independent

H_a: Siskel's and Ebert's opinions are dependent

This calls for a Chi-Square test for Independence of Variables. There are some requirements to conduct this test.

(a) the sample must be a random sample from the population. This is not given; it will be assumed.

(b) all expected values must be at least 1 and most (>80%) must be at least 5. The expected values are shown below:

Table 7 - Expected Counts for Example 6

		Ebert		
		Down	**Mixed**	**Up**
Siskel	**Down**	11.8125	8.4375	24.75
	Mixed	8.4	6	17.6
	Up	21.7875	15.5625	45.65

All are at least 5; the requirement is met.

I'll choose a 5% level of significance (there is no reason to be concerned with how much evidence is needed).

$$X^2 = \frac{(24-11.8)^2}{11.8} + \frac{(8-8.4)^2}{8.4} + \frac{(13-24.8)^2}{24.8} + \frac{(8-8.4)^2}{8.4} + \frac{(13-6)^2}{6} + \frac{(11-17.6)^2}{17.6} + \frac{(10-21.8)^2}{21.8} + \frac{(9-15.6)^2}{15.6} + \frac{(64-45.7)^2}{45.7} = 45.36.$$ $df = (3-1)(3-1) = 4.$

$P(X^2 > 45.36) =$ (for all intents and purposes) 0.

If there is no relationship between Siskel's opinions and Ebert's opinions, then I can expect to find a sample distribution this far or farther from the expected distribution in almost no samples. This occurs too rarely to attribute to chance at the 5% level of significance; it is significant, and I reject the null hypothesis.

It appears that there is some relationship between the opinions of the two men.