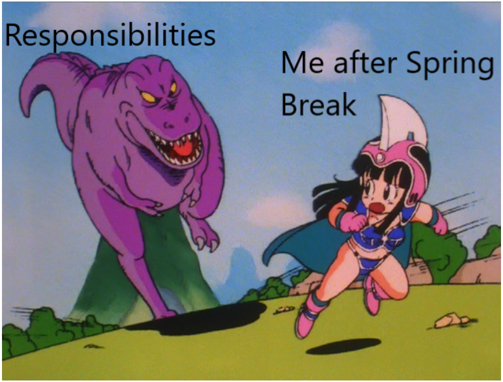
Chi-Square Tests for Categorical Variables

Javier E. Flores

April 3, 2019

Welcome Back...

I'm sure most of you all can relate...



(If you're a fan of Dragonball, you'll get the second layer to this meme)

Recap

So far, we've learned of a few ways of analyzing categorical data. Some of the approaches were implemented in our previous lab and include both one and two-sample methods.

However, for each of these learned methods, we have been limited to testing either a single proportion or single difference in proportions.

As we have seen in some of the datasets we've analyzed in the past, categorical variables aren't always binary. Given this fact, a single proportion or difference in proportions isn't necessarily sufficient to answer a particular research question.

As an example, part of our previous lab asked about the proportion of black victims to police killings. What if we were instead interested in comparing the proportions across all races/ethnicities?

Chi-Square Tests

While we could still technically analyze these data, doing so would involve multiple testing (i.e. we would have to test each pairwise comparison) which would increase our chances of making a type I error.

The methods we'll learn during this lecture, **Chi-Square tests**, allow us to analyze non-binary data and circumvent the multiple testing problem.

We'll introduce Chi-Square tests using data on the correct answer choices among 400 randomly sampled Advanced Placement (AP) Exam questions.

A	**B**	**C**	**D**	**E**
85	90	79	78	68

Chi-Square Tests

Converting these counts to proportions, we obtain the following table:

A	B	C	D	E
0.2125	0.225	0.1975	0.195	0.17

Using these data, we would like to determine whether the distribution of correct answers is truly random.

Question: If the distribution were truly random, what proportion of answers would you expect to be "A"?

Question: Couldn't we just test to see whether the proportion of "A" answers is the same as what we'd expect under a random distribution? Why or why not?

Chi-Square Tests

If the correct answer choices were truly random, we would expect an equal amount of "A's" as we would each other choice. This means that we would expect a proportion of 20%.

Even if we determined that the proportion of "A's" is 20%, we still would not be able to make any claims about the randomness of the distribution of answers as a whole (for all we know, the proportion of "B's" is 80%)!

To fully characterize the data using our currently learned methods, we would need to test at least four different proportions: p_A, p_B, p_C, p_D.

Question: Why don't we need to worry about p_E?

Chi-Square Tests

Instead of performing four different tests for each proportion, Chi-Square tests allow us to perform a single test with hypotheses:

$$H_0 : p_A = p_B = p_C = p_D = p_E = 0.20$$

$$H_A : p_{choice} \neq 0.20 \quad \text{for at least one choice.}$$

Prior to constructing our test statistic, which I will soon introduce, it is essential that we determine what our distribution is expected to look like if we assume H_0 above.

In other words, assuming H_0, exactly how many "A's", "B's", "C's", "D's", and "E's" do we expect to see in a sample of size 400?

Expected Counts

The **expected counts** are the frequencies we would expect to see when assuming that H_0 is true. For our AP dataset, these counts are:

A	**B**	**C**	**D**	**E**
80	80	80	80	80

Generally speaking, we can compute the expected counts for each of the i categories using the following formula:

$$e_i = np_i,$$

where e_i represents the expected count for category i.

Comparing Counts

Intuitively, with the expected counts in hand, the most sensible way to assess the validity of H_0 would be to compare the **observed counts** in our data (denoted o_i) to those expected under the null.

	A	**B**	**C**	**D**	**E**
Expected	80	80	80	80	80
Observed	85	90	79	78	68

If the differences between the observed and expected counts are large enough, we may reasonably suspect the validity of H_0.

Test Statistic

We formalize this intuition through the Chi-Square Test and compute the following test statistic:

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

Looking at the form of this statistic, we see some similarities to previous test statistics.

Similar to previous statistics, the Chi-Square statistic involves a kind of standardized difference.

The difference here is that we square the numerator (so that positive and negative differences don't cancel out) and sum over all the categories.

χ^2 Distribution

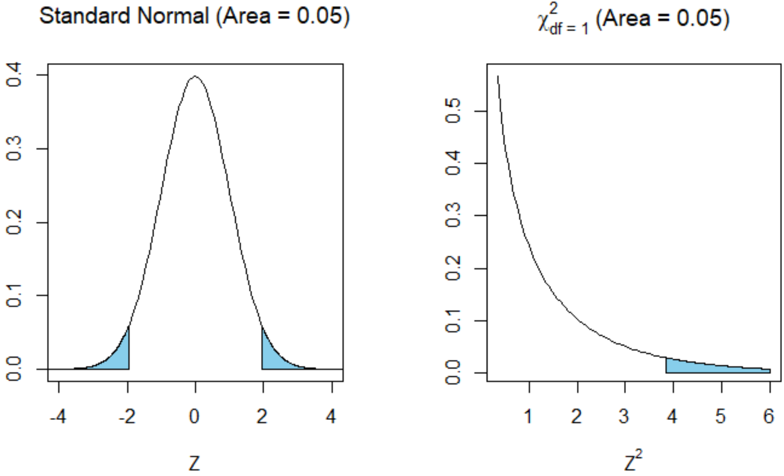
With previous test statistics, we would turn to some reference distribution in order to obtain a p-value. These included randomization distributions, normal distributions, and t-distributions.

The appropriate reference distribution for the Chi-Square statistic is the Chi-Square (χ^2) distribution, which is parameterized by the degrees of freedom.

As it turns out, the χ^2 distribution is related to the normal distribution.

> Suppose we generate 100k observations from the standard normal distribution.
>
> If we were to square each of these observations, the resulting distribution would be χ^2_1, a Chi-Square distribution with $df = 1$.

Standard Normal vs. $\chi^2_{df=1}$



Standard Normal vs. $\chi^2_{df=1}$

To further demonstrate the relationship between the normal and χ^2 distributions, we look to the z-test test statistic:

$$z_{test} = \frac{\text{observed statistic} - \text{null value}}{\text{SE}}$$

$$z^2_{test} = \frac{(\text{observed statistic} - \text{null value})^2}{\text{SE}^2}$$

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

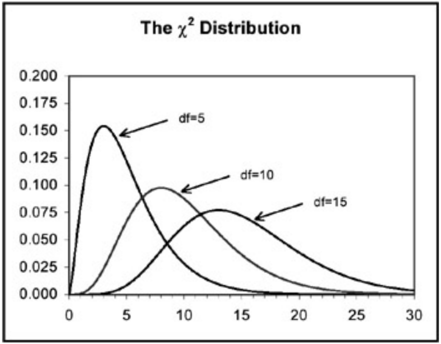
Based on the above relationship, we can think of the χ^2 test as a squared z-test.

This immediately implies that the χ^2 test is always two-sided, even though we only use the right tail of the χ^2 curve to compute p-values.

χ^2_{df}

Earlier I mentioned that the χ^2 distribution was *parameterized* by the degrees of freedom.

Very much like what we see with the t-distribution, this means that the shape of the χ^2 distribution shifts depending on its degrees of freedom.



χ^2_{df}

When performing a χ^2 test, the specific χ^2 distribution we use will depend on the number of categories of our categorical variable.

If we let k represent the number of categories our variable of interest has, then we would want to use the $\chi^2_{df=k-1}$ distribution when computing our p-value.

We use $k-1$ degrees of freedom since all of our category proportions are constrained to sum to 1. Not all proportions are free to vary as they please, so we lose one degree of freedom.

Statkey may be used to find the areas under the different χ^2 curves.

Back to the Example

Returning to our original AP question example, we'll perform a χ^2 from start to finish:

1) We first state our null and alternative hypotheses:

$$H_0 : p_A = p_B = p_C = p_D = p_E = 0.20$$
$$H_A : p_{choice} \neq 0.20 \quad \text{for at least one choice}$$

2) Next, we calculate the expected counts under the null:

$$e_A = e_B = e_C = e_D = e_E = 0.20 * 400 = 80$$

3) We then compute the χ^2 test statistic:

$$\chi^2 = \sum_{i=A}^{E} \frac{(o_i - e_i)^2}{e_i} =$$
$$\frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} = 3.425$$

4) Finally, we use the test statistic to compute the p-value by finding the right-tail area under the curve of a $\chi^2_{df=5-1}$ distribution:

Area under the $\chi^2_{df=5-1}$ curve and to the right of $3.425 =$ $0.429 =$ p-value.

Practice

The American Civil Liberties Union (ACLU) studied the racial composition of jury pools for a sample of 10 trials in Alameda County, California.

The ACLU was interested in determining whether the racial composition of these jury pools was the same as the racial distribution of Alameda County, as determined by the census.

Race/Ethnicity	White	Black	Hispanic	Asian	Other
Number in Jury Pools	780	117	114	384	58
Census %	54	18	12	15	1

Perform a χ^2 test to determine whether the racial composition of jury pools in Alameda County differs from what is expected based upon the census.

Solution

1) State the null and alternative hypotheses:

$$H_0 : p_w = 0.54, p_b = 0.18, p_h = 0.12, p_a = 0.15, p_0 = 0.01$$

H_A: At least one $p_{raceeth}$ differs from what was specified in H_0

2) Calculate the expected counts under the null:

$$e_w = 1453(0.54) = 784.6, e_b = 1453(0.18) = 261.5,$$
$$e_h = 1453(0.12) = 174.4, e_a = 1453(0.15) = 218,$$
$$e_0 = 1453(0.01) = 14.5$$

3) Compute the χ^2 test statistic:

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \frac{(780-784.6)^2}{784.6} + \frac{(117-261.5)^2}{261.5} + \frac{(114-174.4)^2}{174.4} + \frac{(384-218)^2}{218} + \frac{(58-14.5)^2}{14.5} = 357$$

4) Since there are five categories, we compute the p-value using the $\chi^2_{df=5-1}$ distribution and obtain a value of (essentially) 0.

We reject the null hypothesis. The racial composition of the juror pools is not consistent with that of the county. Comparing the observed and expected counts, "Blacks" and "Hispanics" seem to be underrepresented while "Asians" and "Others" are overrepresented.

Goodness of Fit and Association

Broadly speaking, in both of the previous examples we performed a χ^2 test.

However, both of these examples used only a single categorical variable. Our primary interest was in determining how well the distribution of this variable matched some hypothesized distribution.

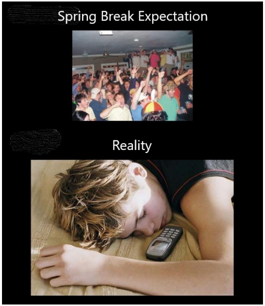
These χ^2 tests involving a single variable are often called **Goodness of Fit** tests.

Throughout the remainder of this lecture, we'll learn how χ^2 tests may also be used to test for associations between two categorical variables - each of which may have several categories.

Spring Break 2019?

To introduce χ^2 tests for association, we'll be using data obtained from a 1976 study investigating factors believed to be correlated with marijuana use among middle-class youths.

In theme with our recent return from spring break, we'll use these data to answer the question: "Is there an association between partying and marijuana use?"



Testing for Association

Party Frequency by Marijuana Use

	Marijuana Use			
Party Frequency	Never	$< 1/\text{month}$	$> 1/\text{month}$	$> 1/\text{day}$
Not at All	40	3	1	0
Somewhat	213	55	44	17
A Great Deal	118	40	54	32

With tests for association, our null hypothesis is that the two variables of interest - party frequency and marijuana use in this example - are not associated.

This differs from the null hypothesis of a goodness of fit test.

Given this difference in H_0, the way that we compute the expected counts for tests for association also differs.

Expected Counts

In our example if H_0 were true (i.e. amount of partying a marijuana use were not associated), we would expect the same distribution of marijuana use across each of the partying categories.

In other words, we would expect the same *proportion* of people who never used marijuana among those who don't party as we would those who party somewhat and those who party a great deal.

Generally speaking, we compute the expected counts (e_{rc}) using the following formula:

$$e_{rc} = n_r * p_c = n_r \frac{n_c}{n},$$

where r indexes the row and c indexes the column of your table.

Expected Counts

(Observed)

Marijuana Use

Party Frequency	Never	$< 1/\text{month}$	$> 1/\text{month}$	$> 1/\text{day}$	
Not at All	40	3	1	0	44
Somewhat	213	55	44	17	329
A Great Deal	118	40	54	32	244
	371	98	99	49	617

(Expected)

Marijuana Use

Party Frequency	Never	$< 1/\text{month}$	$> 1/\text{month}$	$> 1/\text{day}$
Not at All	$44\frac{371}{617} = 26.46$	$44\frac{98}{617} = 6.99$	$44\frac{99}{617} = 7.06$	$44\frac{49}{617} = 3.49$
Somewhat	$329\frac{371}{617} = 197.83$	$329\frac{98}{617} = 52.26$	$329\frac{99}{617} = 52.79$	$329\frac{49}{617} = 26.13$
A Great Deal	$244\frac{371}{617} = 146.72$	$244\frac{98}{617} = 38.76$	$244\frac{99}{617} = 39.15$	$244\frac{49}{617} = 19.34$

Testing for Association

After computing the expected count for each cell in your table, we then can compute the same test statistic used for the goodness of fit test:

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

Using our data and their expected counts, we obtain $\chi^2 = 43.38$.

To find the p-value, we refer to a χ^2 distribution with $df = (r-1)(c-1)$.

In our case, we use a $\chi^2_{df=6}$ distribution to obtain a p-value of (essentially) 0.

Practice

Shown below are data from Joseph Lister's sterile surgery experiment.

	Died	**Survived**
Control	16	19
Sterile	6	34

1) With your groups, perform a χ^2 test for association.
2) Next, perform a test for the difference in proportion who survived.
3) How do the results of these tests compare? Do your conclusions differ?

Solution

1) χ^2 test for association:

	Observed		Expected	
	Died	**Survived**	**Died**	**Survived**
Control	16	19	10.27	24.73
Sterile	6	34	11.73	28.27

$$\chi^2 = \frac{(16-10.27)^2}{10.27} + \frac{(19-24.73)^2}{24.73} + \frac{(6-11.73)^2}{11.73} + \frac{(34-28.27)^2}{28.27} = 8.5$$

Using a $\chi^2_{df=(2-1)*(2-1)}$, we obtain a p-value of 0.0036.

2) Test for difference in proportions:

$$\hat{p}_{con} = 0.54,\ \hat{p}_{ster} = 0.85,\ \hat{p}_{pooled} = 0.71$$

$$z_{test} = \frac{0.53 - 0.85}{\sqrt{\frac{0.71(1-0.71)}{35} + \frac{0.71(1-0.71)}{40}}} = -2.91$$

Using the standard normal distribution, we obtain a p-value of 0.0036.

Same Results!

Comparing these two tests, we see that they yield exactly the same p-value!

Furthermore, you may have also noticed that if we squared our test for proportions test statistic, $z_{test} = -2.91$, we obtain the same value as our χ^2 test statistic, 8.5.

The similarity between these results is no coincidence. Subject to rounding error, the test for a difference in proportions and the χ^2 test for association are equivalent for 2x2 frequency tables.

Despite this equivalence, χ^2 tests are used more since they are more widely applicable (i.e. they can be used for any *rxc* frequency table).

χ^2 Testing in Minitab

As you've probably realized (after these past practices), performing χ^2 tests manually can be a bit tedious. This is particularly true for larger frequency tables.

Fortunately, we can use Minitab to perform χ^2 tests for us. The goodness of fit and association tests are found under the "Stat" -> "Tables" menu.

To perform a χ^2 test in Minitab, your data may be left in its raw form (i.e. two columns each representing a different categorical variable of interest), or may be entered as a summarized *rxc* frequency table.

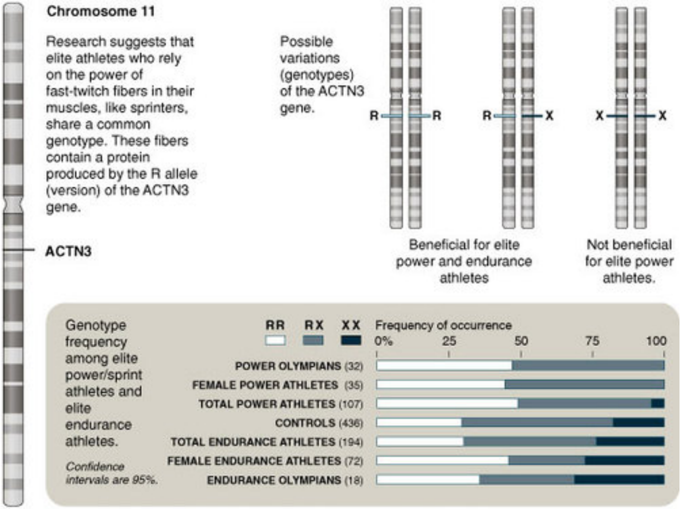
Fast-Twitch Muscle Example

The gene ACTN3 encodes a protein that affects muscle fiber composition and has three genotypes: XX, RR, and RX.

People with the XX genotype are unable to produce any ACTN3 protein but are able to produce a different protein, ACTN2.

It is thought that the ACTN3 protein is associated with increased muscular power whereas the ACTN2 protein is associated with increased muscular endurance capacity.

Fast-Twitch Muscle Example



Research suggests that elite athletes who rely on the power of fast-twitch fibers in their muscles, like sprinters, share a common genotype. These fibers contain a protein produced by the R allele (version) of the ACTN3 gene.

Possible variations (genotypes) of the ACTN3 gene.

Beneficial for elite power and endurance athletes

Not beneficial for elite power athletes.

Sources: Stephen M. Roth, Ph.D., University of Maryland; American Journal of Human Genetics

Fast-Twitch Muscle Practice

The table below contains data from a study on ACTN3 comparing the genotypes of elite sprint/power athletes and elite endurance athletes.

	RR	RX	XX
Sprint/power	53	48	6
Endurance	60	88	46

1) With your groups, use Minitab to perform a χ^2 test for association.

2) State your conclusion. Based on a comparison between the observed and expected counts, what can you say about ACTN3 genotypes and muscular power? What about muscular endurance?

Solution

Observed Counts:

	RR	RX	XX
Sprint/power	53	48	6
Endurance	60	88	46

Expected Counts:

	RR	RX	XX
Sprint/power	40.17	48.35	18.49
Endurance	72.83	87.65	33.51

Using the computed test statistic of $\chi^2 = 19.4$, we obtain a p-value of essentially 0 (note that Minitab is using a $\chi^2_{df=(2-1)(3-1)}$ distribution to get this p-value).

Comparing observed and expected counts, there are far more power athletes with the RR genotype than expected. There are also far more endurance athletes with the XX genotype than expected.

Limitations

While χ^2 tests can be used widely applied in the analysis of two-way frequency tables, they can yield inaccurate results when some cells have small expected counts.

One common rule of thumb is that each cell in your table of expected counts should have a value of five or greater.

This rule is not absolute, but when some cells have expected counts of one or less, the χ^2 test does become extremely inaccurate.

In these instances, where you have one or more expected cell counts that are very low, a randomization test may be performed.

Wrap-Up

Right now, you should...

- Understand the similarities and differences between a χ^2 goodness of fit and association test.
- Be able to appropriately apply each of these testing methods.
- Recognize the relationship between the χ^2 test and the z-test for a difference in proportions for 2x2 tables.
- Know when χ^2 tests may provide inaccurate results, and what alternative option may be available in such a situation.

These notes cover sections 7.1 and 7.2 of the textbook. Please read through the section and its examples along with any links provided in this lecture.