Chi-Square Tests and Inference for Categorical Variables

Ryan Miller

Inference on Categorical Variables

- So far our analysis of categorical variables has been limited to a single proportion (or a single proportion for two groups)
  - What proportion of babies survive early gestation? (one-sample categorical data)
  - Does sterile surgery lead to a higher proportion of patients surviving? (two-sample categorical data)
- Most categorical variables aren't binary and trying to summarize them using a single proportion is unnatural

AP Exam Answers (One-sample data)

- Today we'll explore the topic of statistical inference for non-binary categorical variables
- Below is the distribution of correct answers for 400 randomly selected AP Exam questions:

A	B	C	D	E
85	90	79	78	68

- If AP Exam correct answers are truly random, what proportion of answers do you expect to be "A's"?
- Would a z-test on the proportion of "A" answers provide enough information to tell if the AP Exam's answers are randomly distributed?

AP Exam Answers

- Below are the proportion of AP Exam answers in each category:

A	B	C	D	E
0.2125	0.225	0.1975	0.195	0.17

- To fully characterize the table above we need consider at least 4 different proportions:

$$p_A = 85/400 = 0.213, \; p_B = 90/400 = 0.225$$

$$p_C = 79/400 = 0.198, \; p_D = 78/400 = 0.195$$

- Why don't we need to explicitly provide the fifth proportion, p_E, to summarize the data?

AP Exam Answers

- We analyze these data using 4 different single proportion tests, but that is a lot of effort to analyze a single categorical variable
- A more efficient test would evaluate the hypotheses:

$$H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$$

$$H_A : p_i \neq 0.2 \text{ for at least one } i \in \{A, B, C, D, E\}$$

- Like any statistical test, we begin by thinking about how to put ourselves into the world of the null hypothesis
  - Had we randomly sampled 400 AP questions under the null hypothesis, what is the most likely distribution of the 400 answers?

AP Exam Answers

- The most likely results under the null hypothesis are called the **expected counts**
  - They represent category frequencies we'd expect to observe if the null hypothesis is true
  - For the AP Exam data they are:

A	B	C	D	E
80	80	80	80	80

- In general, we calculate the expected counts for each of i possible categories as:

$$\text{expected}_i = n * p_i$$

- This is easy with the AP Exam data because under the null hypothesis p_i is the same for every category. However, it won't always be this simple.

AP Exam Answers - Chi-Square Testing

- To evaluate $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$ we can compare the **observed counts** with those we'd expect if the null hypothesis was true:

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

- In this framework, we seek to answer the question: "If the null hypothesis is true, are the observed counts in our sample farther from the expected counts than we'd reasonably expect to see by random chance"
- With your group, think about how you'd summarize the distance between the observed and expected counts?
  - Is the distance between 79 and 80 the same as the distance between 80 and 79?
  - Is it the same as the distance between 4 and 5?

The Chi-Square Statistic

- We evaluate H_0 as defined on the previous the **Chi-Square Test**, the test statistic is given below:

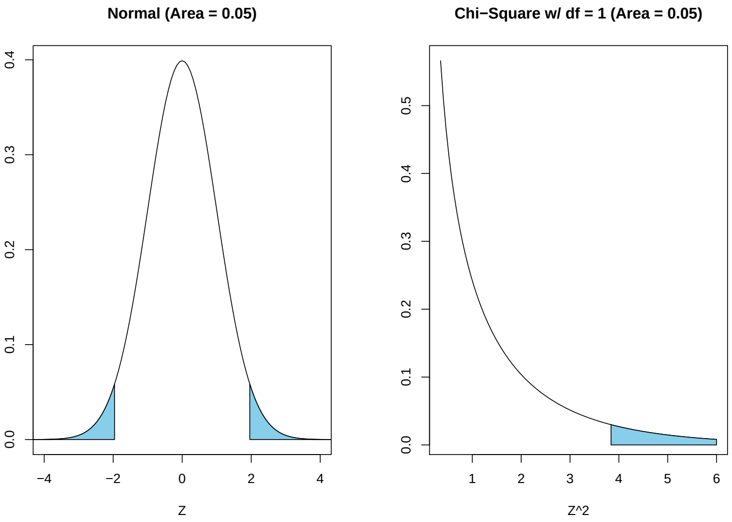
$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

- Like previous test statistics, it compares the observed data to what we'd expect under the null hypothesis, while standardizing the differences
  - Different is that we must sum over the variable's i categories
  - These differences must be squared so that positive and negative deviations from what is expected don't cancel each other out

The Chi-Square Distribution

- Conducting a Chi-Square test requires us to learn a new distribution, the χ^2 curve
- Fortunately, the χ^2 distribution is related to the standard normal distribution
  - Suppose we generated lots of data from the normal distribution, the histogram of these data would look like the normal curve
  - Now suppose we took these observations and squared them, this histogram looks like the χ^2 curve (with $df = 1$)

The Chi-Square Distribution



The Chi-Square Distribution

- The relationship between the χ^2 distribution and the normal distribution is clearly illustrated by looking at the test statistic for the z-test:

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

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

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

- Essentially, the χ^2 test is just a squared version of the z-test
  - This makes the test naturally two-sided, even though we only calculate p-values using the right hand tail of the χ^2 curve
  - Under H_0, the SE of each observed count is approximately the square root of the expected value of that count

Degrees of Freedom

- There are many different χ^2 distributions depending upon how many categorical levels we must sum over
- Letting k denote the number of categories of a categorical variable, the χ^2 test statistic for testing a single categorical variable has $k - 1$ degrees of freedom
  - This arises due to all of the category proportions being constrained to sum to 1
  - The mean and standard deviation of the χ^2 curve both depend upon its degrees of freedom
  - We can use StatKey to calculate the required areas under the various different χ^2 curves

Performing a Chi-Square Test (Quick Example)

1. State the Null Hypothesis:
   $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$
2. Calculate the expected counts under the null:

$$\mathrm{E}_A = 0.2 * 400 = 80,\ \mathrm{E}_B = 0.2 * 400 = 80,\ \ \ldots$$

3. Calculate the χ^2 test statistic:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\ &= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\ &= 3.425\end{aligned}$$

4. Locate the χ^2 test statistic on the χ^2 distribution with $k-1$ degrees of freedom to find the p-value: $p = 0.49$

Chi-Square Testing Example #1

- Pools of prospective jurors are supposed to be drawn at random from the eligible adults in that community
  - The American Civil Liberties Union (ACLU) studied the racial composition of the jury pools for a sample of 10 trials in Alameda County, California
  - The 1453 individuals included in these jury pools are summarized below. For comparison, census data describing the eligible jurors in the county is included

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

Directions: Use a Chi-Square test to determine whether the racial composition of jury pools in Alameda County differs from what is expected based upon the census

Chi-Square Testing Example #1 (solution)

$$H_0 : p_w = 0.54, p_b = 0.18, p_h = 0.12, p_a = 0.15, p_o = 0.01$$
$$H_A : \text{At least one } p_i \text{ differs from those specified in } H_0$$

Race/Ethnicity	White	Black	Hispanic	Asian	Other
Observed Count	780	117	114	384	58
Expected Count	1453*.54 = 784.6	1453*.18 = 261.5	1453*.12 = 174.4	1453*.15 = 218	1453*.01 = 14.5

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

- The *p*-value of this test is near zero and provides strong evidence that the jury pools don't match the racial proportions of the census
- Comparing the observed vs. expected counts, it appears that Blacks and Hispanics are underrepresented while Asians and Others are overrepresented in the jury pools.

Testing for Association

- Both examples so far (AP exam questions and Alameda jury composition) have used a single categorical variable (one-sample data)
  - Using a χ^2 test on a single variable is often called "Goodness of Fit Testing"
- Chi-Square testing is a very general approach that applies not only to categorical variable with two groups (two-sample data), but also data with many groups
  - Using a χ^2 test here is often called "Testing for Association"
  - The procedure is quite similar to the examples we've seen, except it uses the two-way frequency table between two categorical variables

Testing for Association

To illustrate the Chi-Square test for association, we will return to the data from Joseph Lister's sterile surgery experiment:

	Died	Survived
Control	16	19
Sterile	6	34

- The expected counts of the χ^2 test are computed assuming the null hypothesis is true
  - H_0 stipulates no association between surgery group and survival
  - If H_0 is actually true, we'd expect the same proportion of patients in each group to die
  - Overall, 29% (22/75) of patients died. So under H_0, we'd expect 29% of the 40 patients in the sterile group and 29% of the 35 patients in the control group to die, thereby providing the expected counts.

Testing for Association

- When performing a χ^2 test for association, it is useful to look at the observed vs. expected counts as side-by-side tables:

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

- Expected counts are calculated as: $E_{rc} = n_r * p_c = \frac{n_r * n_c}{n}$ where r indexes the row and c indexes the column
  - For example, the top left cell of the expected table is: $E_{11} = \frac{22*35}{75} = 10.267$

Testing for Association

- We use the same test statistic when using the χ^2 to determine association:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
$$= \frac{(16 - 10.27)^2}{10.27} + \frac{(19 - 24.73)^2}{24.73} + \frac{(6 - 11.73)^2}{11.73} + \frac{(34 - 28.27)^2}{28.27}$$
$$= 8.8$$

- The degrees of freedom we use when testing the association between two categorical variables is $(r - 1) * (c - 1)$
  - What *df* do we use when testing a 2x2 table? how might this relate the χ^2 test to the *z*-test?

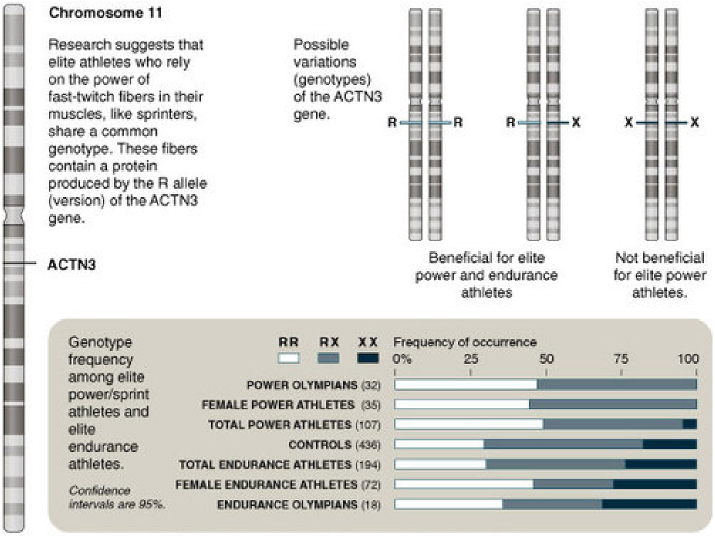
Testing for Association

- When we analyzed Lister's experiment using a difference in proportions test, we got a two-sided *p*-value of 0.003 (using the standard normal distribution)
- When analyzing it using a χ^2 test of association, we get a *p*-value of 0.0036 (using our test statistic of 8.5 and the χ^2_1 distribution)
- This not a coincidence, these tests are equivalent (subject to rounding error) for 2x2 two-way frequency tables
    - In practice, χ^2 tests are used far more frequently than difference in proportions tests due to their generalizability to larger two-way frequency tables

Fast-twitch Muscles - Example

- The gene ACTN3 encodes a protein that affects muscle fiber composition
- Everyone has one of three ACTN3 genotypes: XX, RR, or RX
  - People with the XX genotype can't produce any ACTN3 protein, which is thought to be related with increased muscular power
  - Instead they produce ACTN2, which is thought to relate to increased muscular endurance capacity

Fast-twitch Muscles - Example



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

Fast-twitch Muscles - Example

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	Total
Sprint/power	53	48	6	107
Endurance	60	88	46	194
Total	113	136	52	301

Directions: With your group, test whether there is an association between ACTN3 genotype and muscular power. You should include:

1. The table of expected counts
2. The χ^2 test statistic
3. Your *p*-value and conclusion

Fast-twitch Muscles - Example (solution)

1. Expected Counts:

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

2. Test Statistic: $\chi^2 = 19.4$
3. p-value = nearly 0 (using $\chi^2(df = 2)$ distribution). We conclude there is an association between ACTN3 genotype and muscular power. There more sprinters with RR genotype than expected and more endurance athletes with the XX genotype than expected.

Limitations of Chi-Square Testing

- χ^2 tests are very widely used in statistics, but they are inaccurate when some cells have small expected counts
  - A generally accepted rule is that each cell should have an expected count of at least 5
  - When some cells have expected counts of 1 or fewer, the test becomes wildly inaccurate
  - An alternative approach, Fisher's Exact Test, is an exact approach that suitable for these situations
  - Another alternative would be to resort to randomization tests (these are implemented in StatKey)

Conclusion

Right now you should...

1. Be able to use Chi-Square testing to assess the goodness of fit of a single categorical variable
2. Be able to use Chi-Square testing to assess the association between two categorical variables
3. Know the relationship between the Chi-Square test and the z-test for a difference in proportions for 2x2 tables
4. Know that the Chi-Square test can be inaccurate when cells have expected counts less than 5

These notes cover Sections 7.1 and 7.2 of the textbook, I encourage you to read through those sections and their examples