Chi-Squared Tests

Ryan Miller

Categorical Variables with Many Categories

- Up until now we've analyzed categorical data in two ways:
  - Single proportions (one-sample data) via the z-test and exact binomial test
  - Differences in proportions (two-sample data) via the z-test or Fisher's exact test
- These are foundational statistical approaches; however, they tend to be poorly suited for non-binary categorical variables
  - Put differently, many categorical variables aren't appropriately summarized using a single proportion

AP Exam Answers

- Today we'll explore 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 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 determine if AP Exam's answers are randomly distributed?

AP Exam Answers

- Below are the proportions 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 this table, we'd need at least 4 of its 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 fully characterize these data?

AP Exam Answers

- We could analyze these data using *four different* single proportion tests, but that is a convoluted approach to analyzing a single variable
- A more efficient test would evaluate the hypotheses:

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

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

- To see if we can come up with a test of this hypothesis, we'll begin by assuming the null hypothesis is true:
  - So, had we randomly sampled 400 AP questions under the null hypothesis, what is the most likely distribution of the 400 answers?

AP Exam Answers - Expected Counts

- The most likely frequencies under the null hypothesis are called the **expected counts**
- 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, but that won't always be the case

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, do the observed counts deviate from the expected counts by more than we'd reasonably expect due to 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 previously defined) using the **Chi-Square Test**, the test statistic is given below:

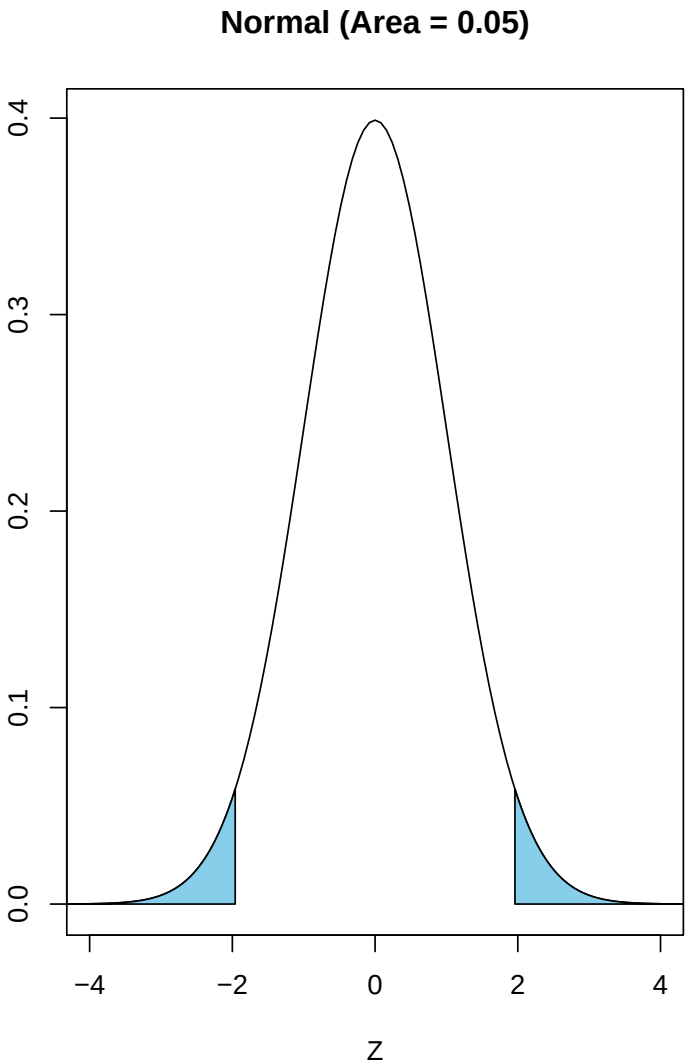
$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

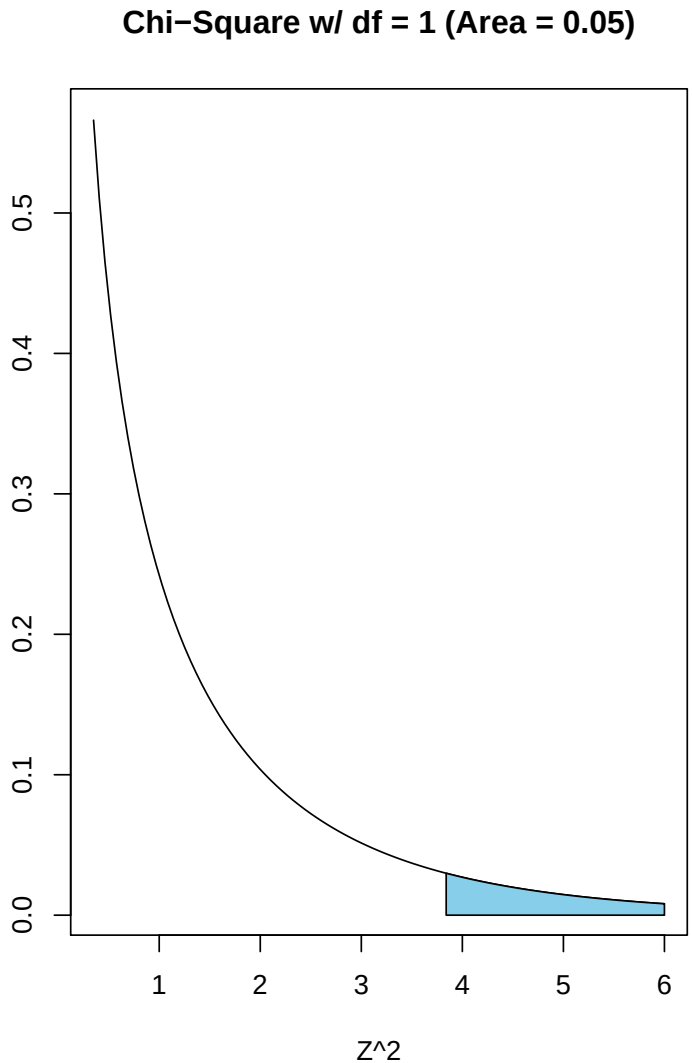
- Like other test statistics, it compares the observed data to what we'd expect under the null hypothesis, while standardizing the differences
  - Different is that we must sum over the variable's i categories
  - Also different is that the numerator is squared so that positive and negative deviations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

The Chi-Square Distribution

- The 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 standard 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:

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

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

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

- 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 tail of the χ^2 curve
  - Under H_0, the SE of each category 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 unique categories 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 is because the category proportions are 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 areas under the various different χ^2 curves

Performing the 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: $p = 0.49$

Example - Jury Composition

- 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

Example - 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.

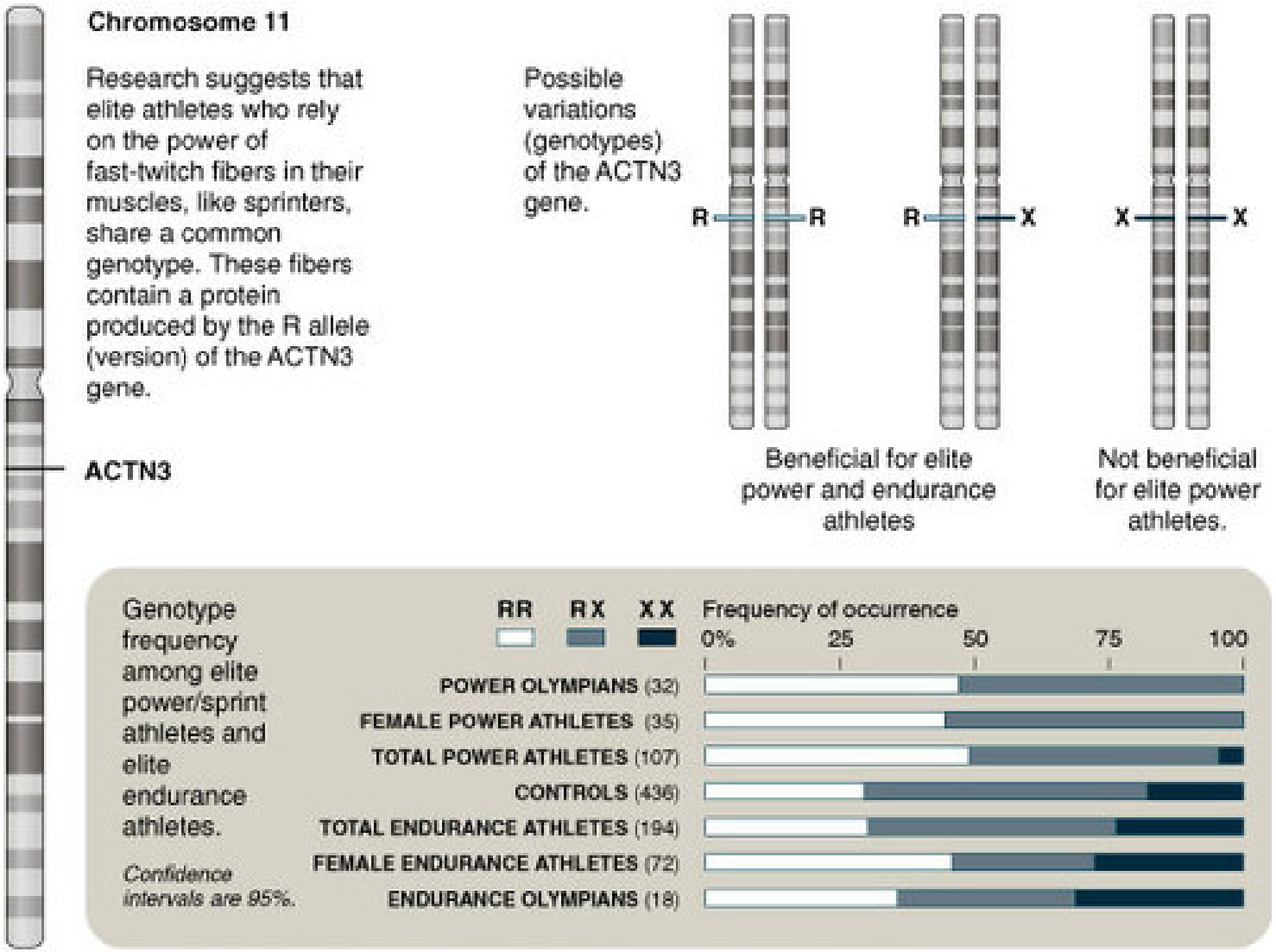
Testing for Association

- Both examples so far (AP exam questions and Alameda jury composition) involved only a single categorical variable (one-sample data)
  - A χ^2 test on a single variable is called "Goodness of Fit Testing"
- The χ^2 test can also be used to evaluate the relationship between two categorical variables (two-sample data)
  - This is often called "Testing for Association"
  - The χ^2 test for association is quite similar to the examples we've seen, but it uses the two-way frequency table of the two variables

Fast-twitch Muscles and ACTN3 - Introduction

- 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 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 and ACTN3 - Introduction



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

Fast-twitch Muscles and ACTN3 - Data Table

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

- We might hypothesize that "sport" is associated with ACTN3 genotype
- To evaluate this claim, we must determine expected counts under a null hypothesis of *no association*

Fast-twitch Muscles and ACTN3 - The Null Hypothesis

- If there is *no association* between sport and ACTN3 genotype, we'd expect genotypes to be distributed identically within each sport
  - This would imply that the *row-proportions* of each sport are *identical*

	RR	RX	XX	Total
Sprint/power	p_{rr}	p_{rx}	p_{xx}	1
Endurance	p_{rr}	p_{rx}	p_{xx}	1

- As we did with differences in proportions, we must use **pooled proportions** to satisfy the null hypothesis while being consistent with the data

Fast-twitch Muscles and ACTN3 - Expected Counts

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

- The pooled proportions are $\hat{p}_{rr} = 113/301 = 0.38$, $\hat{p}_{rx} = 136/301 = 0.45$, and $\hat{p}_{xx} = 52/301 = 0.17$
- We can then determine the expected counts (had the null hypothesis been true) by multiplying the number of athletes in each sport by these pooled proportions:

	RR	RX	XX
SP	107*0.38 = 40.17	107*0.45 = 48.35	107*0.17 = 18.49
EN	194*0.38 = 72.83	194*0.45 = 87.65	194*0.17 = 33.51

Fast-twitch Muscles and ACTN3 - χ^2 Test

- Once we've determined the expected counts, the χ^2 test statistic is calculated in the usual manner:

$$\begin{aligned}\chi^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\ &= \frac{(53-40.2)^2}{40.2} + \frac{(48-48.4)^2}{48.4} + \frac{(6-18.5)^2}{18.5} \\ &+ \frac{(60-72.8)^2}{72.8} + \frac{(88-87.7)^2}{87.7} + \frac{(46-33.5)^2}{33.5} \\ &= 19.4\end{aligned}$$

- For a I by J two-way table, the degrees of freedom of the test statistic are $(I-1)(J-1)$, so $df = 2$ for these data
- The p-value of this test is nearly zero, so we conclude that there is strong evidence that sport is associated with ACTN3 genotype

Chi-Squared Testing in Minitab

- In Minitab, Chi-Squared tests are found under the "Stat -> Tables" menu, you can input data in two ways:
  - The raw data (2 columns of categorical values for each case)
  - A summarized two-way frequency table
- For practice, repeat the Chi-Square test on the ACTN3 data using Minitab:

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

Practice - Chi-Squared Testing

- Chase and Dummer (1992) asked 478 children (grades 4 to 6) from three school districts in Michigan to choose whether good grades, athletic ability, or popularity was most important to them. The table below displays the results of the study broken by gender:

	Grades	Sports	Popularity	Total
Boys	117	60	50	227
Girls	130	30	91	251
Total	247	90	141	478

1. Do these data support the hypothesis that Grades, Sports, and Popularity are equally valued among children in these districts? Answer this question using an appropriate χ^2 test.
2. Is there evidence that boys and girls in this district have different priorities? Answer this question using an appropriate χ^2 test.

Practice - Solution

A):

- $H_0 : p_{grades} = p_{sports} = p_{popular} = 1/3$ versus H_A : at least one proportion is different
- Under H_0, we expect $478 * 0.333 = 159.3$ children to prioritize each category
- Then, $X^2 = \frac{(247-159.3)^2}{159.3} + \frac{(90-159.3)^2}{159.3} + \frac{(141-159.3)^2}{159.3} = 80.5$
- Comparing X^2 with a Chi-Squared distribution with $df = 2$, the p-value is nearly zero

B):

- H_0 : Gender and priority aren't associated
- Under H_0 the expected counts are 117.3, 42.7, and 67.0 for boys, and 129.7, 47.3, 74.0 for girls
- Then, $X^2 = \frac{(117-117.3)^2}{117.3} + \frac{(60-42.7)^2}{42.7} + \frac{(50-67.0)^2}{67.0} + \frac{(130-129.7)^2}{129.7} + \frac{(30-47.3)^2}{47.3} + \frac{(91-74.0)^2}{74.0} = 21.56$
- Next, $df = (3-1) * (2-1) = 2$, so the p-value is nearly zero

Limitations of Chi-Squared Testing

- χ^2 tests are very widely used, 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
- One alternative approach, Fisher's Exact Test, is an exact method that is suitable for these situations
- Another alternative would be to resort to randomization tests (which 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 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