Chi-Squared Tests

Ryan Miller

Outline

1. Chi-Squared testing for goodness of fit
2. Chi-Squared testing for association
3. Sample size conditions and exact tests

Introduction

So far we've discussed two hypothesis tests for *categorical data*:

1) The *one-sample Z-test* evaluates a hypothesis about a single proportion, ie: $H_0 : p = 0.5$
2) The *two-sample Z-test* evaluates a hypothesis about a difference in proportions, ie: $H_0 : p_1 - p_2 = 0$

Both of these tests implicitly treat the outcome variable as *binary* (ie: a variable with only two possible categories)

Nominal categorical variables

- A *nominal categorical variable* has many different categories without a natural ordering
  - Examples include: geographic region, race/ethnicity, blood type, etc.

Nominal categorical variables

- A *nominal categorical variable* has many different categories without a natural ordering
    - Examples include: geographic region, race/ethnicity, blood type, etc.
- Knowing something about one category of a nominal variable doesn't adequately describe the variable as a whole
    - For example, there are four blood types: A, B, AB, and O. If you know that 45% of the US is type O, you don't have enough information to determine prevalence of the other types

Nominal categorical variables

- A *nominal categorical variable* has many different categories without a natural ordering
  - Examples include: geographic region, race/ethnicity, blood type, etc.
- Knowing something about one category of a nominal variable doesn't adequately describe the variable as a whole
  - For example, there are four blood types: A, B, AB, and O. If you know that 45% of the US is type O, you don't have enough information to determine prevalence of the other types
- Contrast this with a binary variable like "survival"
  - If 85% of study participants survived, then exactly 15% must have died

Example - AP Exam answers

Below is the distribution of correct answers to 400 randomly selected AP Exam questions:

A	B	C	D	E
85	90	79	78	68

Example - AP Exam answers

Below is the distribution of correct answers to 400 randomly selected AP Exam questions:

A	B	C	D	E
85	90	79	78	68

1. If AP Exam answers are truly random, what proportion of answers do you expect to be “A’s”?
2. Why won’t a one-sample Z-test on the proportion of “A” answers give you enough information to determine if AP Exam’s answers are randomly distributed?

Example - Expected counts

- A one-sample Z-test only compares a *single observed outcome* with a *single expected outcome*
  - We need to simultaneously compare an *entire set of observed outcomes* with an *entire set of expected outcomes*
  - That is, we want to evaluate:
    $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$

Example - Expected counts

- A one-sample Z-test only compares a *single observed outcome* with a *single expected outcome*
  - We need to simultaneously compare an *entire set of observed outcomes* with an *entire set of expected outcomes*
  - That is, we want to evaluate: $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$

If this null hypothesis were true, we'd *expect* to observe $400 * 0.2 = 80$ correct answers in each category:

A	B	C	D	E
80	80	80	80	80

Example - Observed vs. expected counts

We can then compare the **observed counts** with the **expected counts** (if H_0 were true):

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

- To find a *p*-value describing the discrepancy, we should focus on the question: "If H_0 is true, do the observed counts deviate from the expected counts by more than we'd reasonably expect due to random chance?"
  - Can we come up with a test statistic to summarize these deviations as a single number?

Example - Calculating a test statistic

For a one-sample or two-sample Z-test, we've used the *test statistic*:

$$Z = \frac{\text{observed}-\text{null}}{SE}$$

For a **Chi-squared test**, we'll use the *test statistic*:

$$X^2 = \sum_{i=1}^{k} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

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

Example - Calculating a test statistic

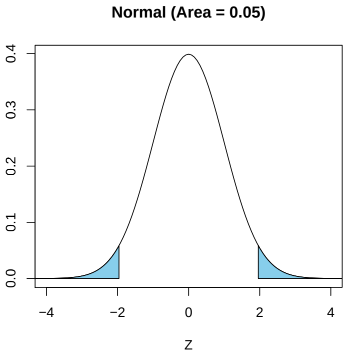
For the AP Exam example:

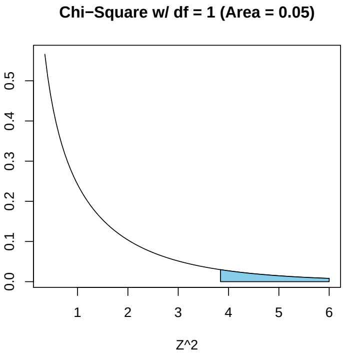
$$\begin{aligned}
X^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\
&= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\
&= 3.425
\end{aligned}$$

Each expected count was found via $e_i = n * p_i$, which was $e_i = 400 * 0.2 = 80$ for every category in this example. In general, p_i can differ for each category.

Example - The Chi-Squared distribution

The Chi-squared distribution is a squared version of the Standard Normal curve:





Example - The Chi-Squared distribution

The relationship between the χ^2 distribution and the Normal distribution is clear when comparing test statistics:

$$Z = \frac{\text{observed} - \text{null}}{SE} \implies Z^2 = \frac{(\text{observed} - \text{null})^2}{SE^2}$$

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

Example - The Chi-Squared distribution

The relationship between the χ^2 distribution and the Normal distribution is clear when comparing test statistics:

$$Z = \frac{\text{observed} - \text{null}}{SE} \implies Z^2 = \frac{(\text{observed} - \text{null})^2}{SE^2}$$

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

- Essentially, the χ^2 test is just a squared version of the Z-test
  - This makes the χ^2 test *naturally two-sided* when we calculate p-values using only the right tail of the χ^2 curve
  - Under H_0, the SE of each category count is approximately the square root of that category's expected count

Example - 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

Example - 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
    - For the AP Exam example, $X^2 = 3.425$ and $k = 5$ (so $df = 4$), the corresponding p-value is 0.49

Practice

Prospective jurors are supposed to be randomly chosen from the eligible adults in a community. The American Civil Liberties Union (ACLU) studied the racial composition of the jury pools in 10 trials in Alameda County, California. Displayed below is the racial and ethnic composition of the $n = 1453$ individuals included in these jury pools, along with the distribution of eligible jurors (according to the US Census):

Race/Ethnicity	White	Black	Hispanic	Asian	Other	Total
Number in jury pools	780	117	114	384	58	1453
Census percentage	54%	18%	12%	15%	1%	100%

1) Based upon the US Census, create a table of expected counts
2) Use these expected counts to perform a *Chi-squared goodness of fit test*

Practice (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

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

- 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 Other are over-represented in the jury pools.

Testing for Association

- Both examples so far (AP exam questions and Alameda jury composition) involved only a single categorical variable
  - A χ^2 test on a single variable is called "Goodness of Fit Testing"

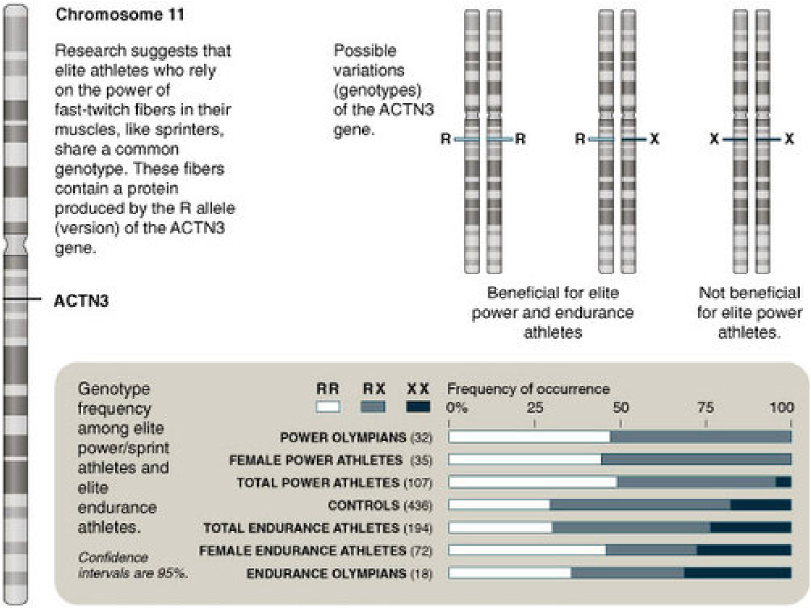
Testing for Association

- Both examples so far (AP exam questions and Alameda jury composition) involved only a single categorical variable
  - 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*
  - This is called "Testing for Association"
  - The only difference is that expected counts must be calculated for a *two-way frequency table* for a χ^2 test for association

Example - Introduction

- The ACTN3 gene encodes a protein that affects muscle fiber composition
    - Everyone has one of three genotypes: XX, RR, or RX
- People with the XX genotype are unable to produce ACTN3 proteins, which is believed to lead to *decreased muscle power*
    - However, the protein that the XX genotype produces is believed to lead to *increased muscle endurance*

Example - Introduction



Chromosome 11

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.

Genotype frequency among elite power/sprint athletes and elite endurance athletes.

Confidence intervals are 95%.

ACTN3
R
R
R
X
X
X
RR
RX
XX
Frequency of occurrence
0%
25
50
75
100
POWER OLYMPIANS (32)
FEMALE POWER ATHLETES (35)
TOTAL POWER ATHLETES (107)
CONTROLS (436)
TOTAL ENDURANCE ATHLETES (194)
FEMALE ENDURANCE ATHLETES (72)
ENDURANCE OLYMPIANS (18)

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

Example - Testing for association

Researchers collected the genotypes of 107 sprint/power athletes and 194 endurance athletes:

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

To determine whether there is an association between “sport” and genotype, our null hypothesis must be “no association”

Example - Expected counts

- If there is *no association* between sport and ACTN3 genotype, we'd the same distribution of genotypes within each sport
  - This would imply that the *row-proportions* of each sport are *equal*

	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

Example - 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 $\hat{p}_{xx} = 52/301 = 0.17$

Example - 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	$194{*}0.38 = 72.83$	$194{*}0.45 = 87.65$	$194{*}0.17 = 33.51$

Example - Calculating a test statistic

- 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 an R by C two-way table, the degrees of freedom of the test statistic are $(R-1)(C-1)$, 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

Mathematical shortcuts

The pooled proportion approach is mathematically equivalent to:

$$\text{Expected Count} = \frac{\text{Row Total} * \text{Column Total}}{\text{Sample Size}}$$

It's often more efficient to use the formula above to fill out a table of expected counts.

Practice

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

A) 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.

B) 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

Comments on sample size

- Chi-squared tests can be inaccurate when some cells have *small expected counts*
  - In general, each cell should have an *expected count of at least 5* in order for the test to be considered reliable
- If this condition is not met, StatKey provides a randomization testing approach that is an appropriate alternative
  - If you're using statistical software, Fisher's Exact test is the approach you should use in these circumstances

Summary

This presentation introduced two different types of Chi-squared test:

1) Chi-squared goodness of fit tests (a single categorical variable)
2) Chi-squared tests of association (relating two categorical variables)

Both tests are based upon comparing observed vs. expected counts. Goodness fit tests compare the sample distribution in a one-way frequency table against a set of hypothesized proportions, while tests of association evaluate whether two variables used to create a two-way frequency table are related.