Chi-Squared Tests

Ryan Miller

Outline

1. Goodness of fit tests
2. Tests of association
3. Measures of effect size

Motivating example - AP Exam answers

Summarized below are the correct answers to 400 randomly selected AP Exam questions:

A	B	C	D	E
85	90	79	78	68

Motivating example - AP Exam answers

Summarized below are the 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 hypothesis test involving the proportion of "A" answers give you enough information to determine if AP Exam's answers are randomly distributed?

Expected counts

- The exact binomial test, as well as the one-sample Z-test, compare 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:
    $H_0 : p_A = p_B = p_C = p_D = p_E = 0.2$

Expected counts

- The exact binomial test, as well as the one-sample Z-test, compare 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* the sample data to have $400 * 0.2 = 80$ correct answers in each category:

A	B	C	D	E
80	80	80	80	80

Observed vs. expected counts

We can 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

- The goal is to find p-value describing this discrepancy:
  - "If H_0 were, what is the probability of deviations at least this large?"
  - Can you come up with a *test statistic* (ie: a Z-value)?

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

Calculating a test statistic

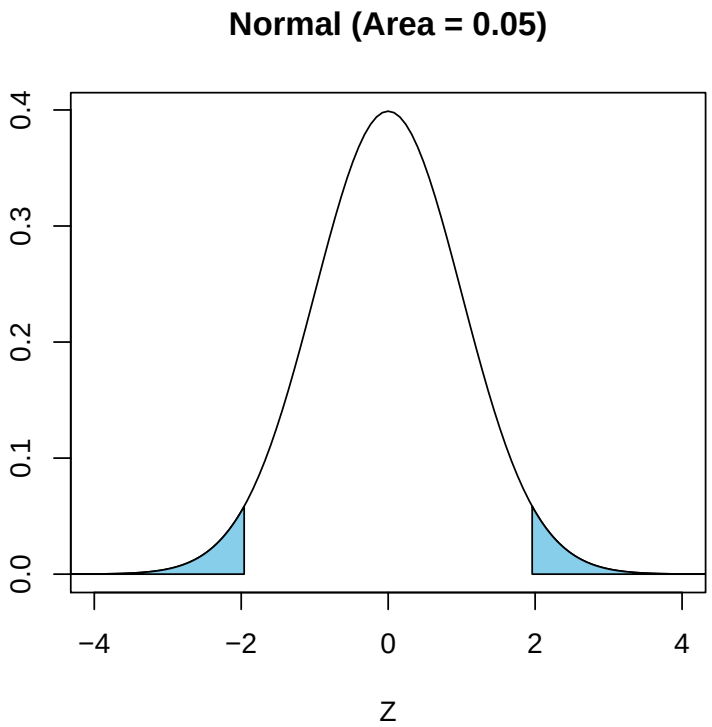
For the AP Exam example:

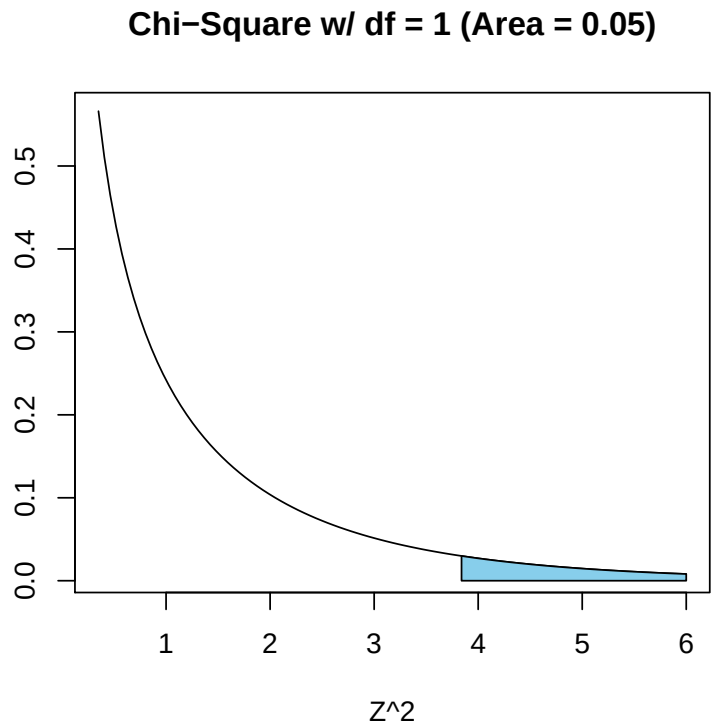
$$
\begin{aligned}
X^2 &= \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \\
&= \frac{(85-80)^2}{80} + \frac{(90-80)^2}{80} + \frac{(79-80)^2}{80} + \frac{(78-80)^2}{80} + \frac{(68-80)^2}{80} \\
&= 3.425
\end{aligned}
$$

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.

The Chi-Squared distribution

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





The Chi-Squared distribution

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

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

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

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

Degrees of freedom

- There are many Chi-squared distributions, each is characterized by a *degrees of freedom* parameter
- For a goodness of fit test, $df = k - 1$, where k is the number of categories being tested
  - Intuitively, the reason for this is that knowing the proportions in $k - 1$ categories will completely describe the variable as a whole (using the complement rule)

Chi-squared goodness of fit testing (summary)

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 H_0:

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

3. Calculate the χ^2 test statistic:

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

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

Chi-Squared goodness of fit testing in R

Using the X^2 test statistic:

```
pchisq(3.425, df = 4, lower.tail = FALSE)
```

```
## [1] 0.4893735
```

Using the sample data directly:

```
observed <- c(85, 90, 79, 78, 68)
chisq.test(observed, p = c(.2, .2, .2, .2, .2))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 3.425, df = 4, p-value = 0.4894
```

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) Perform a *Chi-squared goodness of fit test* both "by hand" and using `chisq.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

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

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

$$= 357$$

- 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 Other are over-represented in the jury pools.

Chi-squared tests for association

Last week we discussed the results of an experiment performed by Joseph Lister involving a sterilization protocol that could be used prior to surgery:

	Died	Survived
Control	16	19
Sterile	6	34

We determined that if the sterilization protocol made no difference, we'd expect 29% in each group to have died.

Chi-squared tests for association

We can use this pooled proportion to create a table of *expected counts*:

	Died	Survived
Control	35*.29 = 10.2	35*.71 = 24.9
Sterile	40*.29 = 11.6	40*.71 = 28.4

Then, we can compare the observed and expected counts using a Chi-squared test:

$$X^2 = \frac{(16-10.2)^2}{10.2} + \frac{(19-24.9)^2}{24.9} + \frac{(6-11.6)^2}{11.6} + \frac{(34-28.4)^2}{28.4} = 8.5$$

For a two-way frequency table, the degrees of freedom are $df = (\text{N rows} - 1)(\text{N cols} - 1)$, or $df = 1$ in this example.

Chi-squared tests for association

In R:

```
## p-value from Chi-squared df =1
pchisq(8.5, df = 1, lower.tail = FALSE)
```

```
## [1] 0.003551465
```

```
## Using chisq.test for the entire test
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
chisq.test(tab, correct = FALSE)$p.value
```

```
## [1] 0.003560924
```

Example #2 - Fast twitch muscle fibers

- 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 #2 - Fast twitch muscle fibers

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". What would this hypothesis suggest in terms of *row proportions*?

Example #2 - Fast twitch muscle fibers

H_0 : "No association" suggests the row proportions are equal for both groups.

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

Thus, the pooled proportions are $\hat{p}_{rr} = 113/301 = 0.38$, $\hat{p}_{rx} = 136/301 = 0.45$, and $\hat{p}_{xx} = 52/301 = 0.17$, which can be used to determine expected counts:

	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 #2 - Fast twitch muscle fibers

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 $df = (2-1)*(3-1) = 2$, the p-value is nearly zero:

```
pchisq(19.4, df = 2, lower.tail = FALSE)
```

```
## [1] 6.12835e-05
```

Fisher's exact test

- The Chi-squared test for association requires a large sample size (all table cells must have expected counts of at least 5)
  - If some cells have expected counts less than 5, Fisher's exact test can be used:

```
## Lister's experiment
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
chisq.test(tab, correct = FALSE)$p.value
```

```
## [1] 0.003560924
```

```
fisher.test(tab)$p.value
```

```
## [1] 0.005018047
```

```
## ACTN3 genotype study
tab <- data.frame(RR = c(53,60), RX = c(48,88), XX = c(6, 46))
chisq.test(tab, correct = FALSE)$p.value
```

```
## [1] 5.989183e-05
```

```
fisher.test(tab)$p.value
```

```
## [1] 2.503932e-05
```

Statistical vs. clinical significance

- The χ^2 test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the p-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)

Statistical vs. clinical significance

- The χ^2 test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the p-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)
- These methods do not tell us anything about the nature of the association
  - We could report the sample difference in proportions (accompanied by a confidence interval), but this summary measure has a major shortcoming

Statistical vs. clinical significance

- The χ^2 test for independence and Fisher's exact test can both be used to evaluate the strength of an association that exists between two categorical variables
  - The lower the p-value, the more strongly the variables are associated (That is, the more incompatible the sample data are with the variables being independent)
- These methods do not tell us anything about the nature of the association
  - We could report the sample difference in proportions (accompanied by a confidence interval), but this summary measure has a major shortcoming
- Consider the proportions of smokers and non-smokers that develop lung cancer in a 10-year period
  - These proportions are estimated at 0.00438 and 0.00045 respectively, or a difference of 0.0039 (far less than 1%)

Effect size and odds ratios

- The most commonly reported *measure of association* describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"

Effect size and odds ratios

- The most commonly reported *measure of association* describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"
- In our smoking example, the odds of a smoker developing lung cancer are $\frac{0.00438}{1-0.00438} = 0.00440$
  - Similarly, the odds of a non-smoker developing lung cancer are $\frac{0.00045}{1-0.00045} = 0.00045$

Effect size and odds ratios

- The most commonly reported *measure of association* describing the relationship between two categorical variables is the **odds ratio**
  - The *odds* of an event is the ratio of how often it happens to how often it doesn't happen
  - If a team has a 75% probability of winning a game, the odds of winning are 3, which is often spoken as "3 to 1"
- In our smoking example, the odds of a smoker developing lung cancer are $\frac{0.00438}{1-0.00438} = 0.00440$
  - Similarly, the odds of a non-smoker developing lung cancer are $\frac{0.00045}{1-0.00045} = 0.00045$
- Thus, the *odds ratio* is $\frac{0.00440}{0.00045} = 9.8$
  - We say that the odds of a smoker developing lung cancer are 9.8 times those of a non-smoker developing lung cancer

Confidence Interval for an Odds Ratio in R

In Lister's experiment, we could conclude with 95% confidence that the odds of death in the Control group are between 1.4 and 17.2 times higher than the odds of death in the Sterile group:

```
## 95% CI for an OR (Lister's Experiment)
tab <- data.frame(Died = c(16,6), Survived = c(19,34))
fisher.test(tab, conf.int = TRUE,
            conf.level = .95)$conf.int[1:2]
```

```
## [1]  1.437621 17.166416
```

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.

C) What is the odds ratio comparing the odds of a boy prioritizing sports relative to a girl prioritizing sports?

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

Practice (solution - continued)

C:

- The odds of a boy prioritizing sports are $\frac{60/227}{1-60/227} = 0.359$
- The odds of a girl prioritizing sports are $\frac{30/251}{1-30/251} = 0.136$
- The odds ratio (boy/girl) is $0.359/0.136 = 2.64$, indicating boys are 2.64 times a likely to prioritize sports a girls in these schools

```
## Can also be found (with 95% CI) via fisher.test
tab <- data.frame(Sports = c(60,30), Not = c(167,221))
fisher.test(tab, conf.int = TRUE,  conf.level = .95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 6.057e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.596333 4.443950
## sample estimates:
## odds ratio
##   2.641287
```

Summary

This presentation covered two types of Chi-squared tests:

1) **Goodness of fit** - used to analyze a single categorical variable
2) **Association** - used to find associations between two categorical variables

All of the fundamental concepts we've previously covered apply to these new situations, but we must be aware of when and how to implement these new statistical tests.