Chapter 6 - Inference for Categorical Data

Bryan Persaud

2010 Healthcare Law. (6.48, p. 248) On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.
(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.
(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.
(d) The margin of error at a 90% confidence level would be higher than 3%.

(a)

This is false because the confidence interval is used for the population and we know that 46% of the sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b)

This is true because the margin of error is 3% so we can assume with 95% confidence that 46% + 3% and 46% - 3% support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c)

This is false because the 95% confidence interval would show the true percentage of Americans who support the U.S. Supreme Court.

(d)

This is false because the margin of error would decrease along side the confidence interval. So if we lower the confidence interval from 95% to 90%, we would be lowering the margin of error as well.

Legalization of marijuana, Part I. (6.10, p. 216) The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.

(a) Is 48% a sample statistic or a population parameter? Explain.
(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.
(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.
(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

(a)

48% is a sample statistic because it found from the 1259 US residents.

(b)

```
SE <- (((0.48) * (0.52)) / (1259))^ (0.5)
upper_tail <- 0.48 + 1.96 * SE
lower_tail <- 0.48 - 1.96 * SE
c(lower_tail, upper_tail)
```

```
## [1] 0.4524028 0.5075972
```

The 95% confidence interval for the proportion of US residents who think marijuana should be made legal is (0.4524, 0.5076).

(c)

This is true because we have a sample size that is 1259, so it large enough to assume a normal distribution. Also (1259 * 0.48) and (1259 * (1 - 0.48)) are both grater than 10.

(d)

This news piece's statement is not justified because the confidence interval is between 0.4524 and 0.5076, so that shows us that almost 50% of Americans want marijuana legalized. But for the true population this could be less if it falls out of range. Either way it does not show that majority of Americans want marijuana legalized, at most it only shows 50%.

Legalize Marijuana, Part II. (6.16, p. 216) As discussed in Exercise above, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

```
(0.48 * 0.52) / (0.02 / 1.96)^ 2
```

```
## [1] 2397.158
```

You would have to survey 2398 Americans if you want to limit the margin of error of a 95% confidence interval to 2%.

Sleep deprivation, CA vs. OR, Part I. (6.22, p. 226) According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insuffient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

```
SE_2 <- sqrt(((0.08 * (1 - 0.08)) / 11545) + ((0.088 * (1 - 0.088) / 4691)))
ME <- 1.96 * SE_2
lower_tail_2 <- 0.088 - 0.08 - ME
upper_tail_2 <- 0.088 - 0.08 + ME
c(lower_tail_2, upper_tail_2)
```

```
## [1] -0.001498128  0.017498128
```

The 95% confidence interval for the difference between the proportions of Californains and Oregonians who are sleep deprived is (-0.0014, 0.0175). Since the 95% confidence interval includes 0, we can assume that the proportions between Californains and Oregonians are not statistically different.

Barking deer. (6.34, p. 239) Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7% and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	67	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.
(b) What type of test can we use to answer this research question?
(c) Check if the assumptions and conditions required for this test are satisfied.
(d) Do these data provide convincing evidence that barking deer pre- fer to forage in certain habitats over others? Conduct an appro- priate hypothesis test to answer this research question.

(a)

H0: There is no preference for barking deer to forage in certain habitats over others. H1: There is a preferecne for barking deer to forage in certain habitats over others.

(b)

The type of test we can use to answer this research question is a chi-squared test.

(c)

The assumptions and conditions for this test are satisfied. We can assume each observation is independent. Also each scenario is expected to have at least 5 cases. We can see this for all scenarios.

```
0.048 * 426 # Woods
```

```
## [1] 20.448
```

```
0.147 * 426 # Grassplot
```

```
## [1] 62.622
```

```
0.396 * 426 # Forests
```

```
## [1] 168.696
```

```
(1 - (0.048 + 0.147 + 0.396)) * 426 # Other
```

```
## [1] 174.234
```

(d)

```r
chisq.test(x = c(4, 16, 67, 345), p = c(0.048, 0.147, 0.396, 0.409))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(4, 16, 67, 345)
## X-squared = 272.69, df = 3, p-value < 2.2e-16
```

These data do provide convincing evidence that barking deer prefer to forage in certain habitats over others since the value of p is less than 0.05. This means that we reject H0 and accept H1.

Coffee and Depression. (6.50, p. 248) Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

{

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

}

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?
(b) Write the hypotheses for the test you identified in part (a).
(c) Calculate the overall proportion of women who do and do not suffer from depression.
(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$).
(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?
(f) What is the conclusion of the hypothesis test?
(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.64 Do you agree with this statement? Explain your reasoning.

(a)

The type of test that is appropriate for evaluating if there is an association between coffee intake and depression is a chi-squared test.

(b)

H0: There is no an association between coffee intake and depression in women. H1: There is an association between coffee intake and depression in women.

(c)

```
2607 / 50739 # depressed women
```

```
## [1] 0.05138059
```

```
48132 / 50739 # nondepressed women
```

```
## [1] 0.9486194
```

The proportion of women who do suffer from depression is 5.14%. The proporion of women who do not suffer from depression is 94.86%.

(d)

```
expected_count <- 5.138 / 100 * 6617
expected_count
```

```
## [1] 339.9815
```

```
(373 - expected_count)^ 2 / expected_count
```

```
## [1] 3.206716
```

The highest cell is 373. The expected count for this cell is 339.98 and the contribution to the test statistic is 3.21.

(e)

```
chisq <- 20.93
df <- (2 - 1) * (5 - 1)
1 - pchisq(chisq, df)
```

```
## [1] 0.0003269507
```

The p-value is 0.00033.

(f)

The conclusion of the hypothesis test is there is an association between coffee intake and depression in women. Since the value of p is less than 0.05, we reject H0 and accept H1.

(g)

Yes, I agree with this statement because this test was done by observation and not by experiments. There could be other factors that could lead to depression in the women in the survey. More studies have to be done to show the benefits and side effects of coffee. An experimental test would have to be done to conclude that more coffee intake can reduce depression.