Chapter 6 - Inference for Categorical Data

Leticia Salazar

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.

False, the sample has a 46% approval rating.

(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.

True, with the confidence interval of 95%, there is a 3% margin of error meaning, 46% - 3% = 43% and 46% + 3% = 49%.

(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%.

False, not all of the random samples of 1,012 Americans will fall under this, but we are 95% confident that 95% of the population will fall within this range.

(d) The margin of error at a 90% confidence level would be higher than 3%.

False, if the confidence level went down to 90% so would the margin of error since the z-score would also change from 1.96 to 1.65.

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.

48% is only a sample statistic since it was taken from asking 1,259 US residents and not the US population.

(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.

Based on the results below, the margin of error is 2.76%. We are 95% confident that the approval rate of US residents for marijuana to be legal is between 43.24% and 50.76%.

```
#95% interval

n <- 1259
p <- .48
z <- 1.96
me <- z * sqrt(p * (1 - p) / n)
me
```

```
## [1] 0.02759723
```

```
lower <- (p - me) * 100
lower
```

```
## [1] 45.24028
```

```
upper <- (p + me) * 100
upper
```

```
## [1] 50.75972
```

(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.

Being that the sample size is large enough and if the observations are independent, then this statement will be true for this data.

(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?

This statement is not justified because basing on the confidence interval range almost 50% of the sample size think that marijuana should be legalized. For the true population, the the approval can be less since it will be of a larger size.

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?

We'd need to survey 2,398 Americans if we wanted to limit the margin of error to 2%.

```r
me <- 0.02
p <- .48
z <- qnorm(0.975)

se <- me / z
nu <- (p * (1 - p) / se^2)
nu
```

```
## [1] 2397.07
```

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 insufficient 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.

Ho: There is no difference in sleep deprivation Ha: There is a difference in sleep deprivation

Based on the calculation below at a 95% confidence interval we cannot reject our null hypothesis. There is not so much of a difference in the confidence intervals of the two states.

```
#California
cal_n <- 11545
cal_p <- 0.08
z <- 1.96

#Standard Error
se_cal <- sqrt(cal_p * (1 - cal_p) / cal_n)

#Margin of Error
me_cal <- z * se_cal
me_cal
```

```
## [1] 0.004948778
```

```
#Confidence interval
lower <- (cal_p - me_cal) * 100
lower
```

```
## [1] 7.505122
```

```
upper <- (cal_p + me_cal) * 100
upper
```

```
## [1] 8.494878
```

```
#Oregon
ore_n <- 4691
ore_p <- 0.088
z <- 1.96

#Standard Error
se_ore <- sqrt(ore_p * (1 - ore_p) / ore_n)

#Margin of Error
me_ore <- z * se_ore
me_ore
```

```
## [1] 0.008107036
```

```r
#Confidence interval
lower <- (ore_p - me_ore) * 100
lower
```

```
## [1] 7.989296
```

```r
upper <- (ore_p + me_ore) * 100
upper
```

```
## [1] 9.610704
```

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	61	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

Ho: The barking deer doesn't have a preference of a certain habitat to forage. Ha: The barking deer does have a preference of a certain habitat to forage.

(b) What type of test can we use to answer this research question?

A Chi-squared test can be used to answer this research question

(c) Check if the assumptions and conditions required for this test are satisfied.

Based on the calculations below, we can assume the observations are independent. The assumptions are met and test are satisfied.

```
#Woods
0.048 * 426
```

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

```
#Grassplot
0.147 * 426
```

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

```
#Forests
0.396 * 426
```

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

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

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

(d) Do these data provide convincing evidence that barking deer pre- fer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

The p-value is < 2.2e-16 so we can reject the null hypothesis and barking deer does have a preference of a certain habitat to forage.

```r
bd <- c(4, 16, 61, 345)
percentags <- c(0.048, 0.147, 0.396, 0.409)

chisq.test(x = bd, p = percentags)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  bd
## X-squared = 284.06, df = 3, p-value < 2.2e-16
```

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	2-6	1	2-3	≥ 4	
		cup/week	cups/week	cup/day	cups/day	cups/day	Total
Clinical	Yes	670	373	905	564	95	2,607
depression	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?

Chi-square test can be used to evaluate if there is an association between coffee intake and depression

(b) Write the hypotheses for the test you identified in part (a).

Ho: There is no association between coffee intake and depression in women. Ha: There is an association between coffee intake and depression in women..

(c) Calculate the overall proportion of women who do and do not suffer from depression.

5.14 of women suffer from depression and 94.86% of women do not.

```
#Depressed women
dep_wo <- 2607 / 50739
dep_wo
```

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

```
#Not depressed women
non_dep_wo <- 48132 / 50739
non_dep_wo
```

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

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$).

The highlighted cell is of 373 and the expected count for this cell is of 340.11. The contribution of this cell to the test statistic is 3.18.

```
expected_count <- 5.14 / 100 * 6617
expected_count
```

```
## [1] 340.1138
```

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

```
## [1] 3.179824
```

(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?

The p-value is 0.000327.

```r
# Find p-value
chi_squ <- 20.93

df <- (2 - 1) * (5 - 1)

1 - pchisq(chi_squ, df)
```

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

(f) What is the conclusion of the hypothesis test?

We reject the null hypothesis, there is an association between the coffee ntake and depression.___

(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. Do you agree with this statement? Explain your reasoning.

I agree with this statement made from the authors of this study. Being that this test was just an observation it would be best if an experiment was done to have more information on the relationship between coffee and depression..