Chapter 6 - Inference for Categorical Data

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.

This is false. The confidence interval would apply to the entire population, not the sample taken.

(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.

This is true. Since we are speaking out the population at large and *not* the sample of 1,012, this aligns with our confidence interval given that the proportion of 46% has a ME of +/- 3% (which is a confidence interval of 43% to 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%.

This is true (most of the time). 95% of the sample proportions will most likely fall between the 43% and 49% confidence interval.

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

This is false. Since we have more certainty and a smaller margin of error as we reduce the confidence level. The z-score used to calculate the margin of error will be lower at 90% than at 95%, which would then result in a smaller ME.

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.

The 48% measure is a sample statistic since it is based on a proportion from the 1,259 US residents that were surveyed. The population parameter would be the actual proportion of all U.S. residents that support marijuana legalization legislation.

(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.

In order to establish the confidence interval, we'll need to do the following:

```r
# calculate the z-score of a 95% confidence interval
z_score_1 <- qnorm(0.975)

n <- 1259
p_hat <- 0.48

# calculate the margin of error
se <- sqrt(p_hat * (1 - p_hat) / n)

# calculate the confidence interval
(p_hat) + z_score_1 * (se)
```

```
## [1] 0.5075967
```

```r
(p_hat) - z_score_1 * (se)
```

```
## [1] 0.4524033
```

Therefore, the 95% confidence interval would be from 45.2% to 50.8%. This would state that we are 95% confident that the proportion of all U.S. residents that support marijuana legalization legislation would fall between 45.2% and 50.8%

(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.

The critic is correct, the 95% confidence interval is only valid if the observations are independent and representative of the population of interest – it has to be a random sample. Additionally, there must be at least 10 successes and 10 failures. We do know from the proportion of the sample that there are at least 10 successes and 10 failures:

```r
# successes
paste0('number of successes: ', n * p_hat)
```

```
## [1] "number of successes: 604.32"
```

```r
# failures
paste0('number of failures: ', n * (1 - p_hat))
```

```
## [1] "number of failures: 654.68"
```

However, there isn't mention of this being a random sample, although we may be able to assume that the observations are independent.

(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?

Based on the confidence interval that I calculated earlier, we actually wouldn't be able to decipher either way in this case since the upper bound of the confidence interval is just slightly past 50%. Additionally, a majority of the confidence interval falls below 50% in support of the legalization legislation, so I'd be more inclined to say that this is not justified, but when accounting for a 95% confidence interval, it could go either way.

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 ?

To perform this calculation, we'd have to do the following:

We can use the formula $1.96^2 \times \frac{p(1-p)}{ME^2}$

```r
sample <- (1.96^2) * ((0.48*(1-0.48))/(0.02^2))

paste0('We would need to survey at least ', round(sample, digits = 0), ' Americans.')
```

```
## [1] "We would need to survey at least 2397 Americans."
```

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.

To perform this calculation, we'd have to work through the following using this formula:

$\hat{p}_1 - \hat{p}_2 \pm z^\star \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

```r
p_hat_1 <- 0.080
p_hat_2 <- 0.088
z_score <- 1.96
n_1 <- 11545
n_2 <- 4691

standard_error <- (sqrt(((p_hat_1*(1-p_hat_1))/n_1)+(p_hat_2*(1 - p_hat_2)/n_2)))

upper_bound_cal <- abs(p_hat_1-p_hat_2)+(z_score*standard_error)
lower_bound_cal <- abs(p_hat_1-p_hat_2)-(z_score*standard_error)

upper_bound_org <- p_hat_1-p_hat_2+(z_score*standard_error)
lower_bound_org <- p_hat_1-p_hat_2-(z_score*standard_error)

paste0('California perspective (CI): ',
       round(lower_bound_cal, digits = 4), ' to ', round(upper_bound_cal, digits = 4),
       '.')
```

```
## [1] "California perspective (CI): -0.0015 to 0.0175."
```

```r
paste0('Oregon perspective (CI): ',
       round(lower_bound_org, digits = 4), ' to ', round(upper_bound_org, digits = 4),
       '.')
```

```
## [1] "Oregon perspective (CI): -0.0175 to 0.0015."
```

We are 95% confident that the proportion of California residents that are sleep deprived is from 0.15% less than Oregon residents to 1.75% more than Oregon residents. Stated differently, we are 95% confident that the proportion of Oregon residents that are sleep deprived is from 1.75% less than California residents to 0.15% more than California residents.

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.

- **H0: Deer do not have preference to forage in a certain habitat over others. Deer forage at equal rates across different habitat types.**
- **HA: Deer prefer certain habitats over others for foraging and will forage at higher rates in some habitats over others.**

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

I will use a chi-square to answer this research question.

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

The assumptions and conditions that need to be checked and required for this test to accurate are:

- **Independence: Each case that contributes a count to the table must be independent of all other cases in the table – this seems to be satisfied for our particular research question.**
- **Sample size / distribution: Each particular scenario must have at least 5 expected cases – we can see that the woods scenario only has 4 observed cases, but when checking the number of expected cases (426 * 0.048), we see that there are 20 expected cases, so this satisfies this requirement as well.**

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

We will first do the chi-square analysis:

Chi-Square Formula: $X^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + ... + \frac{(O_k-E_k)^2}{E_k}$

```
other_o <- 345
other_e <- 426 * (1 - (0.048 + 0.147 + 0.396))
decid_o <- 67
decid_e <- 426 * 0.396
grass_o <- 16
grass_e <- 426 * 0.147
woods_o <- 4
woods_e <- 426 * 0.048

chi_square <- (((woods_o-woods_e)^2)/woods_e) +
  (((grass_o-grass_e)^2)/grass_e) +
  (((decid_o-decid_e)^2)/decid_e) +
  (((other_o-other_e)^2)/other_e)

df <- 4-1

p_deer <- 1-pchisq(chi_square, df=df)

p_deer
```

```
## [1] 0
```

From the chi-square test, we have arrived at a p-value of less than 0.0001. Therefore, we can reject the null hypothesis that deer do not have a preference to forage in certain habitat types over others.

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?

We can also use a chi-square to evaluate if there is an association between coffee intake and depression.

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

- **H0: There is not an association between coffee intake and depression – there is no difference in the frequency of depression among women across different levels of coffee consumption.**
- **HA: There is an association between coffee intake and depression – there is a difference in the frequency of depression among women across different levels of coffee consumption.**

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

The overall proportion of women who do a do not suffer from depression can be calculated this way:

```
women_depression <- 2607 / 50739
women_nodepression <- 48132 / 50739

paste0('Overall proportion of women with depression = ',
       round(women_depression, digits = 4))
```

```
## [1] "Overall proportion of women with depression = 0.0514"
```

```
paste0('Overall proportion of women that do not have depression = ',
       round(women_nodepression, digits = 4))
```

```
## [1] "Overall proportion of women that do not have depression = 0.9486"
```

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

To calculate the expected count of the highlighted cell:

```
two_six_e <- 6617 * women_depression
two_six_e
```

```
## [1] 339.9854
```

The expected count for women who drink 2-6 cups per week with clinical depression is roughly 340.

To calculate the contribution of this cell to the test statistic:

```r
two_six_o <- 373

t_contrib <- ((two_six_o-two_six_e)^2)/two_six_e
t_contrib
```

```
## [1] 3.205914
```

The highlighted cell contributes about 3.206 to the test statistic.

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

Given that we have the test statistic = 20.93, and the degrees of freedom are $(5-1columns)\times(2-1rows) = 4$**, we have all the information we need to find the p-value:**

```r
p_dep_coffee <- pchisq(20.93, 4, lower.tail = FALSE)

paste0('The p-value is equal to ', round(p_dep_coffee, digits = 5), '.')
```

```
## [1] "The p-value is equal to 0.00033."
```

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

The conclusion of the test is that we can reject the null hypothesis that there is no association between coffee intake and women with depression. This is the case because the p-value is less than 0.05

(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 that it is too early to recommend that women load up on extra coffee. Since we used chi-square tests, we were able to find a weak relationship between the two variables, but we'd need to look at effect sizes and directionality across the different levels of the categorical data before drawing firmer conclusions about coffees effect on depression in women.