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.

False: The confidence interval estimates the population proportion, not the sample.

(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: As explained above, we are certain of this being correct.

(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: The 95% is how certain we are that the true proportion of the population exists within our interval, not how certain we are for all confidence intervals. The values we get for different samples will vary.

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

False: As we decrease confidence levels, the margin of error likewise decreases as our critical value is also decreasing.

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% is a sample statistic. There was no easy way to survey all of Americans, and no mention of this survey being an exception to that.
(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.

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

c( p - (z * SE),  p + (z * SE))
```

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

(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.

We can infer that the sampling distribution is approximately normal given that our observations are independent, and if the success-failure condition is met. In this case, both conditions appear to be met, so we can assume that our distribution is normal.

(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?

It's a bit of a stretch to say this is true. It's possibly true, given that the upper limit of our confidence interval is above 50 percent, with a margin of error that could push it over that. However, it's also possible that our true value is less than 45 percent as well.

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?

```
p = 0.48

ME = 0.02

z = 1.96

ze = ME/z

round(((p*(1-p))/ze^2),0)
```

```
## [1] 2397
```

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.

```
z <- 1.96

n_c = 11545
p_c = .08



n_o = 4691
p_o = .088

p_d =  p_o - p_c
SE <- sqrt((p_o*(1-p_o))/n_o)  + sqrt((p_c*(1-p_c))/n_c)

moe = qnorm(0.95)*SE

c(p_d - moe, p_d + moe)
```

```
## [1] -0.002956584  0.018956584
```

Since our interval includes zero within it, we can conclude that there is significant proof that the sleep deprivation for both groups is the same.

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.

H0: Barking deer prefer to forage certain in certian habitats over others.

HA: There is no distinguishable difference in how much the barking deer forage in each area.

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

We could test this with a Chi-Square Test for one way tables.

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

We can assume that observations in our data are independent as there is no apparent relationship between them and they are also not paired in any way, one type of land is separate from another. Our observations are all categorical. And if we look at the overall values for each region, the smallest 4.8% piece of woods makes up a 20.45 portion of the land, so we have at least 5 of each of our variables as well.

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

```
habitats <- c(4, 16, 67, 345)

region <- c(.048*42, .147*426, .396*426, .409*426)

chiSq<-sum((habitats-region)^2/region)


pchisq(chiSq,3,lower.tail=FALSE)
```

```
## [1] 3.151971e-57
```

Our value from the test is so small that we can conclude that there is convincing evidence for our null hypothesis. Barking deer forage in certian habitats 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 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?

We can use a two-way table chi squared test to test if an association exists or not.

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

H0: There is no association between coffee intake and depression.

HA: There is an association between coffee intake and depression.

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

```
yes = 2607
no = 48132

c(yes/(yes+no), no/(yes+no) )
```

```
## [1] 0.05138059 0.94861941
```

5.14 percent of women do suffer and 94.86 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. $(Observed - Expected)^2/Expected$).

```
Expect = (yes * 6617) / (yes+no)

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

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

We expect a vakye of around 3.2.

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

```
chi = 20.94

freedom = (5-1) * (2-1)

 pchisq(chi, freedom,lower.tail =F)
```

```
## [1] 0.0003254619
```

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

Since our p value is way less than .05, we can not reject our null hypothesis, we conclude that a significant difference does exist.

(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 would agree. While the study points to a significant difference existing, it does not prove that coffee is the cause for this. There are likely other factors at play that are leading the coffee group to be different than the no coffee group.