Chapter 6 - Inference for Categorical Data

Shana Green

DATA 606 - Homework 6

Due Date: 10/11/2020

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, because the sample is 46%.The statement mentions population, not 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

(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, because the confidence interval does not say anything about the proportion of samples, just about the population.

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

False, because as the confidence level decreases, so does z, which means that the margin of error decreases.

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.

Since 48% is the sample proportion for a sample of 1259 US residents, it is a characteristic of a sample hence it is a statistic.

(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.

```
# 95% confidence interval
z_score <-1.96

n <-1259

p_value <- 0.48

# Standard error
se <- round(sqrt(p_value*(1-p_value)/n),4)
se
```

```
## [1] 0.0141
```

```
upper <- p_value + z_score*se
lower <- p_value - z_score*se
ci <- round(c(upper, lower),4)
ci
```

```
## [1] 0.5076 0.4524
```

(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 normal model would make a good approximation, since the p-value is near half at 0.48 and 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 results in (b), the news piece's statement is not justified because 95% confidence interval from 0.5 to 0.45.

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?

```
z_score <-round(qnorm(0.975),2) #1.96
n <-1259
error<-0.02 #margin of error
p_value <- 0.48
se <- error/z_score
se
```

```
## [1] 0.01020408
```

```
n<- round((z_score^2*p_value*(1-p_value) / error^2),0)
paste0("We would need to survey ", round(n,0), " Americans.")
```

```
## [1] "We would need to survey 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.

```
# 95% confidence interval
z_score <- round(qnorm(0.975),2)
p_o_c <- 0.08
cali_res <- 11545
p_o_o <- 0.088
oreg_res <-4691

# SE for California
se_cali <- round(sqrt(p_o_c*(1-p_o_c)/cali_res),6)
se_cali
```

```
## [1] 0.002525
```

```
# SE for Oregon
se_oregon <- round(sqrt(p_o_o*(1-p_o_o)/oreg_res),6)
se_oregon
```

```
## [1] 0.004136
```

```
cali_oregon <- round(sqrt(se_cali^2+se_oregon^2),6)
cali_oregon
```

```
## [1] 0.004846
```

```
lower<-(p_o_o-p_o_c)-z_score*cali_oregon
upper<-(p_o_o-p_o_c)+z_score*cali_oregon
ci <- round(c(upper, lower),4)
ci
```

```
## [1]  0.0175 -0.0015
```

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.

H_0: Distribution of foraging preference does follow distribution of available land type.

H_A: At least one of the values are different.

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

We can use a chi-squared goodness of fit test.

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

```
sites <-426
wood<-0.048
grass<-0.147
forest<-0.396
other<-1-(wood+grass+forest)

exp_w<-round(sites*wood,0)
exp_g<-round(sites*grass,0)
exp_f<-round(sites*forest,0)
exp_o<-round(sites*other,0)

exp_w
```

```
## [1] 20
```

```
exp_g
```

```
## [1] 63
```

```
exp_f
```

```
## [1] 169
```

```
exp_o
```

```
## [1] 174
```

These are all above 5.

(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.

```r
obser<-c(4,16,67,345)
expected<-c(sites*wood,sites*grass,sites*forest,sites*other)

chi_sq<-sum((obser-expected)^2/expected)
df<-4-1

pchisq(chi_sq,df,lower.tail=FALSE)
```

```
## [1] 1.144396e-59
```

Since the p-value <0.0001, we have to reject the null hypothesis.

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 the chi square test.

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

H_0 - The variables are independent.

H_1 - The variables are dependent.

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

```
cd <- 2607
no_cd<- 48132
total <- cd + no_cd
prop_cd <-round(cd/total,3)
prop_no_cd <- round(no_cd/total,3)

prop_cd
```

```
## [1] 0.051
```

```
prop_no_cd
```

```
## [1] 0.949
```

5.138% of women suffer from depression and 94.9% 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. $(Observed - Expected)^2/Expected$).

```
total_two_six<-6617
exp_e<-(cd*total_two_six)/total

exp_e
```

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

```
obs_o <-373

chisq <- round(((obs_o-exp_e)^2)/exp_e,4)
paste0("The contribution of this cell is ", chisq,".")
```

```
## [1] "The contribution of this cell is 3.2059."
```

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

```
chi_test <- 20.93
df <- (5-1)*(2-1)
#df=4

pchisq(chi_test,df,lower.tail = FALSE)
```

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

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

Since the p-value <0.05, we reject H_0. This means that there is sufficient evidence to support the claim that clinical depressions are related to the caffeinated coffee consumption.

(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 because the effects in this study could be due to another variable that was not included in the story. So in this case, we cannot recommend that women load up on extra coffee.