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. I believe this statement to be false. This is because we are 100% certain of the results by the poll. By definition the confidence interval is constructed to estimate the population proportion. Not to estimate sample proportion

(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. I think this is true because we are 95% confident(confidence for the given sample proportions). The margin of error is 3%.[46%+3% and 46%-3%, (43% and 49%) as confidence interval.]

(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%. From the context this seems to be False. The 95% confidence level is for the given sample proportions. That is to say not for the random sample proportions. 95% confidence intervals is defined[population proportions]

(d) The margin of error at a 90% confidence level would be higher than 3%. This appears to be false. We know decreasing the confidence level would make th confidence interval narrower. The lower the confidence level, the lower the margin of error will be.

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 for a given sample, then it's a sample 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.

```
zs <- 1.96
n <- 1259
p <- .48
SE <- sqrt((p*(1-p))/n)
CI.LOWER <- p - (zs * SE)
CI.UPPER <- p + (zs * SE)
CI <- c(CI.LOWER, CI.UPPER)
CI
```

```
## [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. True. The central Limit Theorem for Proportions apply the conditions. Observations are independent. The sample is no more than 10% of population. Success-failure condition. The success and failure are both greater than 10.

(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?

The confidence interval is (0.45, 0.51). THE CI is in between 45% ad 51%, which is 50% of the population.the lower limit of confidence interval is less than 50% population who think that marijuana should be made legal. This is not justified

---

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
SE <- ME / 1.96
n <- (p * (1-p)) / (SE^2)

n
```

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

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.

```
p_cali <- 0.08
p_ore <- 0.088

p <- p_cali - p_ore

n_cali <- 11545
n_ore <- 4691

SE_cali <- (p_cali * (1-p_cali)) / n_cali
SE_ore <- (p_ore * (1-p_ore)) / n_ore

SEprop <- sqrt(SE_cali + SE_ore)
ME <- 1.96 * SEprop

p-ME
```

```
## [1] -0.01749813
```

```
p+ME
```

```
## [1] 0.001498128
```

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: Barking deer has no preference of certain habitats for foraging. HA: Barking deer prefers some habitats over others for foraging.

(b) What type of test can we use to answer this research question? chi-square test can be used to answer this question

(c) Check if the assumptions and conditions required for this test are satisfied. Independence. Each case that contributes a count to the table must be independent of all the other cases in the table. The behavior of the barking deer are likely independent and each expected value is above 5.

(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.

we reject the null hypothesis that deer have no preference

```
obs <- c(4, 16, 61, 345, 426)
exp_p <- c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396, 1)
expected <- exp_p * 426
deer <- rbind(obs, expected)
colnames(deer) <- c("Woods", "Grassplot", "Forests", "Other", "ALL")
deer
```

```
##           Woods Grassplot Forests   Other ALL
## obs       4.000    16.000  61.000 345.000 426
## expected 20.448    62.622 168.696 174.234 426
```

```
k <- 4
df <- k-1
chisquaretest <- sum(((obs - expected)^2)/expected)
( p_value <- 1 - pchisq(chisquaretest, df) )
```

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

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? chi-square. this is to test to evaluate if there is independence between depression and coffee intake.

(b) Write the hypotheses for the test you identified in part (a). (H0) : There is no difference in rates of depression in women based on caffeine consumption. (H1) : There is difference in rates of depression in women based on caffeine consumption.

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

```
dep <- 2607
no_dep <- 48132
t <- dep + no_dep

prop_dep <- dep/t
p_nodep <- no_dep/t

print("proportion suffering")
```

```
## [1] "proportion suffering"
```

```
prop_dep
```

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

```
print("proportion not suffering")
```

```
## [1] "proportion not suffering"
```

```
p_nodep
```

```
## [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. $(Observed - Expected)^2/Expected$).

```
e <- prop_dep*6617
e
```

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

```r
obs <- 373
stat <- ((obs-e)^2)/e
stat
```

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

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

```r
c <- 20.93

d <- (2 - 1) * (5 - 1)
pchisq(c, d,lower.tail=FALSE)
```

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

(f) What is the conclusion of the hypothesis test? Since the p values is very small 0.0003, we reject the null hypothesis

(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.

Since this was all an observational study and we cannot infer causation then we can agree with the author. However,radomized experiments will need to be created to prove if there is depression as a result of coffee intake.