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, 46% of the Americans in this sample agree with this decision.

(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, margin of error is the supposed percentage difference between the results and the real population value.

46 + 3 = 49 46-3 = 43

(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% confdidence interval relates to the true value of the population not the sample proportions.

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

False when teh confidence level decreases the margins outside of our range decrease which in turn decreases the margin of error.

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.

Sample statistic. It is in reference to the proportion of the survey which is the sample taken in this case.

(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.

```
stE = sqrt((.48*(1-.48))/1259)
ME = 1.96*stE
lowerbound <- .48-ME
lowerbound
```

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

```
upperbound <- .48+ME
upperbound
```

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

We are 95% confident that between 45.2% and 50.8% of US residents think marijuana should be made legal

(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 following conditions are met which makes the normal model a good approximation.

-Independent observations from random sampling?

Yes

-The sample size is large enough (success-failiure condition)

np=1259(.48)= 604.32 n(1-p) = 1259(1-.48)=654.68

Yes both numbers are above 10

-Sample is less than 10% of the toal poulation

Yes

(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?

No, you can not definitively say that a majority of americans thing it should be legal. The CI ranges from 45.2-50.8 Making only a small bit of this range would suport that headline.

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?

solve ME = .02

~~~
```r
#ME = 1.96*sqrt((.48*(1-.48))/p)
#.0004 =3.84*((.48*.52)/p)
p = (3.84*(.48*.52))/.0004
p
```
~~~

~~~
```
## [1] 2396.16
```
~~~

You would need to survey 2397 Americans to limit the margin of error to 2%.

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.

$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

$(.088 - .08) \pm 1.96\sqrt{\dfrac{.088(1-.088)}{4691} + \dfrac{.08(1-.08)}{11545}}$

$(.008) \pm 1.96\sqrt{.0000171 + .00000638}$

$(.008) \pm 1.96\sqrt{.0000235}$

$(.008) \pm 1.96 * .004845$

$(.008) \pm 1.96 * .009496$

$(.008) \pm .009496$

(-.001496,.017496) We are 95% confident that the difference between oregonians are -.1496% less sleep deprived then californians and 1.7496% more sleep deprived then californians.

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: There is no difference in the preference for forage habitats of barking dear HA: There is a difference in the preference for forage habitats of barking dear

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

chi-square test

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

This data meets the following assumptions

randomly sampled, categorical data

This data set fails to meet the assumption that all observations are at least 5 cases (there are 4 cases of woods).

(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.

```
counts <- c(4,16,67,345)
expected_counts <- c(.048,.147,.396,(1-.396-.147-.048))
chisq.test(counts,p = expected_counts)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  counts
## X-squared = 272.69, df = 3, p-value < 2.2e-16
```

The p-value is nearly 0 (less than .05) so we will reject the null hypothesis in favor of the alternate 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	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

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

H0: The rate of coffee consumption has no impact on the rates depression HA: TThe rate of coffee consumption has impact on the rates depression

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

```
nd<-48132/50739
d<-  2607/50739
```

The proporiton of women who do not suffer from depression in this study is .05138.

The proporiton of women who suffer from depression in this study is .94861.

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

```
d<-  2607/50739
expected<- d*6617
expected
```

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

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

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

We expected 339.99 in this cell.

Teh contribution of this cell to the test statistic is 3.21

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

```
df<-(2-1)*(5-1)
1-pchisq(20.93, df)
```

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

p-value = .000327

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

The p-value is below .05 so we reject the null hypothesis in favor of the alternate 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.

I agree due to the fact this study was observational which makes it unwise to suggest coffee consumption causes the rate of depression to change.