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, CI is for infernece about the population, 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, CI infers about population.

(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%.

True, while the CI is meant for population inference if we did many sample so as to simulate the populations we would see (with some variation) around 95% of them would be fall within our CI for the population mean.

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

False, a 90% CI encapsulates a narrower band of data, thus a lower 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.

Sample. This is based on a random sample of 1259 US residents. We can infer about he population from this sample, but it is a sample.

(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.

```
SE<-sqrt((0.48*(1-0.48))/1259)
lower<-round(0.48- 1.96*SE,3)
upper<- round(0.48 + 1.96*SE,3)
paste("95% CI",lower,"-",upper)
```

```
## [1] "95% CI 0.452 - 0.508"
```

We can be 95% confident that 45-51% the population agrees with this statement.

(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.

This critic is wrong for our understanding of normality assumption. Success/Fail condition holds both ($/geq10$)

```
support<-1259*(0.48)
not<-1259*(1-0.48)
paste("support:", support, "Not:", not)
```

```
## [1] "support: 604.32 Not: 654.68"
```

(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?

Not entirely true. A partial truth. The range of values in the CI spans 50% agreement, but the CI also contains values below that to 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?

If ME must be 0.02 and Z = 1.95, then we get the following for N.

```
(1.96)^2*((0.48)*(1-0.48))/(.02^2)
```

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

We would need to double the sample size.

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.

Difference Test

H0: Change in sleep quality is 0, HA: change in sleep quality is not 0

```r
n_cal<-11545
n_or<-4691
pct_cal<-0.08
pct_or<-0.088
pct_diff<- pct_cal - pct_or
SE <- sqrt(((pct_cal*(1-pct_cal))/n_cal)+((pct_or*(1-pct_or))/n_or))
upper<-round(pct_diff + (1.96*SE),4)
lower<-round(pct_diff - (1.96*SE),4)
paste("CI for sleep Cal vs Org:", lower, "-", upper)
```

```
## [1] "CI for sleep Cal vs Org: -0.0175 - 0.0015"
```

Since the 95% CI contains 0%, we cannot reject the Null Hypo - the difference between Cal and Org is due to chance.

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: barking deer forage in microhabitats that are the same as the natural habitat distribution. HA: barking deer forage in microhabitats that are NOT the same as the natural habitat distribution.

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

Chi-squared

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

This seems to be independent and there are more than 5 counts of each type. We are good to use Chi-squared

(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.

```
n_habitat<- c(4,16,67,345)
other_pct<- (1-.048-.147-.396)
pct<- c(.048,.147,.396,other_pct)
prop=426*pct

chi_2 <- sum((n_habitat - prop )^2 / prop)
chi_2
```

```
## [1] 276.6135
```

This is a huge value when considered similar to Z-score. Reasonable that we will reject the Null.

Check with R for p-value, Degree of Freedom = 3.

```
pchisq(chi_2, 3, lower.tail = F)
```

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

Calling that zero, we can reject the null hypothesis; barking deer prefer a microhabitat some microhabitats to 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?

two-way Chi squared

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

H0: There is no association between these vairables, coffee and depression. HA: There is an association between coffee and depression.

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

```
depress<-2607/50739
not_depress<-48132/50739
paste(depress, "Have depression", not_depress, "Do have depression")
```

```
## [1] "0.051380594808727 Have depression 0.948619405191273 Do have depression"
```

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

Expected count

```
expect<-depress*6617
expect
```

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

Contribution

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

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

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

Degrees of Freedom $(5-1) \times (2-1) = 4$

```
pchisq(20.93, 4, lower.tail = F)
```

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

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

Given a critical value of p =0.05 we can reject the null. There is some association between coffee and depression.

(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.

We do not know the association, we only there is *some* association. Also, this is only one study that is not an experiment and we don't know the nature of the smapling so we don't know if we can infer about the population yet.