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 - confidence interval can only be used on a population proportion and not a 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.

True - if we take the mean and add and subtract the SE we will get a confidence interval of (43, 49)

(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 - By definition we are confident that 95% of the sample proportions will be between 43% and 49%

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

False - As the confidence level decreases so does 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 statstistics because it is just a point estimate of the population

(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 = function(p, n){sqrt((p * (1-p)) / n)}

ME = se(.48, 1259) * 1.96

lower = .48 - ME
upper = .48 + ME

lower
```

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

```
upper
```

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

We can see that the confidence interval is (0.45, 0.50)

(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 - As long as the all the observations are independent and the sample size must be sufficiently large

(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?

False - While the upper confidence interval does pass 50% i would only state that a majority of people would agree only if the lower confidence interval passes 50%. I believe this is based on how one wants to manipulate the data

---

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 = 1.96
ME = 0.02
p = 0.48
q = 1-p

n = (Z^2 * p * q)/(ME^2)

n
```

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

We would need to interview 2398 people in order to get a confidence interval of 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.

```
p_cal = 0.08
n_cal = 11545

p_org = 0.088
n_org = 4691

SE = sqrt(
  (p_cal * (1-p_cal)/n_cal) + (p_org * (1 - p_org)/n_org)
)

lower = (p_cal - p_org) - 1.96 * SE
upper = (p_cal - p_org) + 1.96 * SE

lower
```

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

```
upper
```

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

> We can see that the confidence interval is (-0.0175, 0.0015) which mean that the difference between California and Oregan is -0.0175 and 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 = Barking Deer prefer all forage habitats equally H_A = Barking Deer prefers a certain habitat over another

(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.

Assume that it is indpendent, the sample size is at least 5 for each observation, and Degree of Freedom > 1.

```
n = 426

Wood = n * (0.048)

Grass = n * (0.147)

Forest = n * (0.396)

Other = n * (1 - 0.048 - 0.147 - 0.396)

Wood
```

```
## [1] 20.448
```

```
Grass
```

```
## [1] 62.622
```

```
Forest
```

```
## [1] 168.696
```

```
Other
```

```
## [1] 174.234
```

(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.

```r
X = (4-Wood)^2/Wood + (16-Grass)^2/Grass + (61-Forest)^2/Forest + (345-Other)^2/Other

pchisq(q = X, df = 3, lower.tail = FALSE)
```

```
## [1] 2.799724e-61
```

As we can see the P-Value is so small that i needed to use Lower.tail = False. This means that we reject the null hypothesis and accept the alternative 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	$\boxed{373}$	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 will be using a Chi Square test

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

H_0: Depression it the same regardless of coffee intake H_A: Depression decreases if one consumes coffee

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

```
yes_depression = 2607/50739
no_depression = 48132/50739

yes_depression
```

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

```
no_depression
```

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

```
expected = (2607*6617) / 50739

contribution = (373 - expected)^2 / expected

contribution
```

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

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

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

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

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

we can see that the P value is very small therefore we will reject the null hypothesis and instead accept the alternative 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.

There needs to be more studies and more experiments to verify that this experiment results are indeed correct. This is the reason why experiments need to be reproducable in order for results to be verifed