Homework Assigment 6

CUNY MSDA DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

April 2, 2017

Chapter 6 - Inference for Categorical Data

```
Practice: 6.5, 6.11, 6.27, 6.43, 6.47
Graded: 6.6, 6.12, 6.20, 6.28, 6.44, 6.48
```

6.6 2010 Healthcare Law.

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.

Answer

(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: For this sample, we are certain that 46% of this sample supports the 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: This sample allows us to make an inference about the 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%.

False: Our confidence interval gives us a range of possible values for the true population proportions.

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

False: It would be lower, since we are lowering our confidence.

6.12 Legalization of marijuana, Part I.

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.

Answer

(a) Is 48% a sample statistic or a population parameter? Explain.

The 48% is a sample statistic. That is, it was calculated based on a 1,259 sample of the total US 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.

```
n <- 1259
p <- 0.48
cip <- 0.95 # Defining our confidence interval percentage
SE <- ((p * (1 - p)) / n) ^ 0.5
t <- qt(cip + (1 - cip)/2, n - 1)
ME <- t * SE
ci <- c(p - ME, p + ME)
```

The 95% confidence interval for the proportion of US residents who think marijuana should be made legal is from 45.24% to 50.76%.

(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: That is because we are looking at a proportion $\hat{p}$.

As long as the sample observations are independent and the sample size is large enough such that $n \cdot p \geq 10$ and $n \cdot (1-p) \geq 10$, $\hat{p}$ will be normally distributed.

(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: We cannot reject the hypothesis that the proportion of Americans who think marijuana should be legalized is above 50%, however we also cannot reject the hypothesis that the proportion is below 50%.

6.20 Legalize Marijuana, Part II.

As discussed in Exercise 6.12, 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?

Answer

Since $ME = t \cdot SE$

$$ME = t \cdot \sqrt{\frac{p(1-p)}{n}}$$

Solving for **n** we obtain

$$n = \frac{p \cdot (1-p) \cdot t^2}{ME^2}$$

```
# By carrying the results from the previous problem
ME <- 0.02
n <- (p * (1 - p ) * t ^ 2) / ME ^ 2
```

The survey should require 2403 Americans.

6.28 Sleep deprivation, CA vs. OR, Part I.

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.

Answer

```
pCA <- 0.08
nCA <- 11545
pOR <- 0.088
nOR <- 4691
cip <- 0.95 # Defining confidence interval

pDiff <- pOR - pCA


# Compute standard error and margin of error for the proportion difference.
SE <- ( ((pCA * (1 - pCA)) / nCA) +  ((pOR * (1 - pOR)) / nOR)) ^ 0.5
z <- qnorm(cip + (1 - cip) / 2)

me <-z * SE

# Construct the 95% confidence interval.
ci <- c(pDiff - me, pDiff + me )
```

The 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived is from -0.0015 to 0.0175.

This interval overlaps 0, therefore we can conclude with a 95% confidence level that the proportions are not statistically different. In other words, CA and OR population proportion might be equal given the results from this sample.

6.44 Barking deer.

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

Figure 1:

Answer

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

H_0: The sites where barking deer forage were distributed according to the portion of land in each habitat. That is:

Woods	Cultivated grassplot	Deciduous Forests	Other	Total
20.45	62.62	168.70	174.23	426

H_A: The sites where barking deer forage is not every distributed across the habitats of the region. That is, the above table results are different.

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

We can use a Chi-square test for one-way table.

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

Independence: We will have to assume it since is not given in the description.

Sample size / distribution: In our expected cases scenario, all habitats have at least 5 expected cases, therefore this condition is satisfied since it has at least 5 expected cases.

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

```
habitats <- c(4, 16, 67, 345)
region <- c(20.45, 62.62, 168.70, 174.23)
k <- length(habitats)
df <- k - 1
# Compute the chi2 test statistic
chi <- (habitats - region ) ^ 2 / region
chi <- sum(chi)

# check the chi2 test statistic and find p-val
p_Val <- 1 - pchisq(chi, df = df)
```

The chi-Square value is large enough that the p-value is 0. Hence, we conclude that there is convincing evidence the barking deer forage in certain habitats over others.

6.48 Coffee and Depression.

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.

Answer

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

The Chi-squared test for two-way tables is appropriate for evaluating if there is an association between coffee intake and depression.

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

		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

Figure 2:

The hypotheses for the Chi-squared two-way table test are as follows.

H_0: There is no association between caffeinated coffee consumption and depression.

H_A: There is an association between caffeinated coffee consumption and depression.

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

```
yes_dep <- 2607
no_dep <- 48132
total_dep <- yes_dep + no_dep
```

The overall proportion of women who do suffer from depression is 5.14%. The overall proportion of women who do not suffer from depression is 94.86%

(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.

$$\frac{(O_k - E_k)^2}{E_k}$$

```
ntot2cup <- 6617
Ek <- (yes_dep * ntot2cup) / total_dep
chipart <- ((373 - Ek) ^ 2) / Ek
```

The expected count for the highlighted value is 3.2059144

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

```
n <- 5
k <- 2

df <- (n-1)*(k-1)
chi2 <- 20.93

p_value <- 1 - pchisq(chi2, df)
```

The p-value is 0.0003269507

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

Since the p-value is below 0.05, we cannot reject the null hypothesis that coffee doesn't cause 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.

Based on the above statement, I agree with it. That is, according to the chi-square test, we found no link between coffee consumption and depression.