Chapter 6 HW

Kishore Prasad

March 31, 2016

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.[39]

(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.

(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.

(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%.

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

(a) TRUE - Given a margin of error of 3% and a point estimate of 46%, the interval for a 95% confidence interval works out to 43% and 49%.

(b) TRUE - Given a margin of error of 3% and a point estimate of 46%, the interval works out to 43% - 49%. We can say that the interval of 43% to 49% would contain the true population mean 95% of the times.

(c) FALSE - 95% of the times, the "Population" proportion would be between 43% and 49% and not the "Sample" Proportion.

(d) FALSE - A lower confidence would mean a smaller Z score. A smaller Z score would lead to a smaller margin of error. The margin of error formula is :

margin_of_error = Z * standard_error

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.[44]

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

(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.

(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.

(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?

(a) It is stated "48% of the respondents". This clearly means that it is a sample statistic.

(b) Following are the calulations:

```
n <- 1259
p <- 0.48

# Assuming that the samples are independent
success <- p * n
failure <- (1-p) * n

success > 10
```

```
## [1] TRUE
```

```
failure > 10
```

```
## [1] TRUE
```

```
# Success / failure condition is met with at least 10 successes and at least 10 failures

#calculating standard error

se <- sqrt( (p * (1-p) )/ n)
se
```

```
## [1] 0.01408022
```

```
z <- qnorm(0.975)

# calculating Margin of Error

me <- z * (se)
me
```

```
## [1] 0.02759672
```

```
# Construct the 95% Confidence Interval

CI <- data.frame(lower=p - me, upper=p + me)
CI
```

```
##       lower     upper
## 1 0.4524033 0.5075967
```

** The confidence interval is 0.4524033, 0.5075967**

(c) The sample proportion of 0.48 is close to 0.5. So we can conclude that the distribution is nearly normal given that the data represents a binary state of 0 and 1.

(d) The confidence interval has a upper bound of 0.5075967. If we were to interpret based on this upper bound, we see that it is slightly above 50%. This can be used to justify the statement (even though it seems inappropriate)

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 ?

```
p <- 0.48
me <- 0.02
z <- qnorm(0.975)

# calculating standard error for new margin of error
se <- me / z

# working backward for  n in se formula

n <- (p * (1-p)) / se^2
n
```

```
## [1] 2397.07
```

A survey involving at least 2398 Americans would be needed to achieve a 2% margin of error for a 95% confidence interval

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.[53]

```
p_California <- 0.08
p_Oregon <- 0.088
n_California <- 11545
n_Oregon <- 4691

# calculating the difference

p_Diff <- p_Oregon - p_California
```

```
# calculating the standard error for the difference.
SE <- sqrt( ((p_California * (1 - p_California)) / n_California) +  ((p_Oregon * (1 - p_Oregon)) / n_Or

# calculating the margin of error for the difference.
me <- qnorm(0.975) * SE

# calculating the 95% confidence interval.
CI <- data.frame(lower=p_Diff - me, upper=p_Diff + me )
CI
```

```
##          lower      upper
## 1 -0.001497954 0.01749795
```

The 95% confidence interval of -0.001498 - 0.017498 contains 0. Hence we can conclude with a 95% confidence level that the proportions are not statistically different.

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.[62]

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.
(b) What type of test can we use to answer this research question?
(c) Check if the assumptions and conditions required for this test are satisfied.
(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.

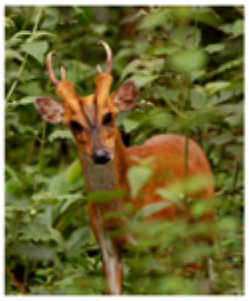

Photo by Shrikant Rao (http://flic.kr/p/4Xjdkk) CC BY 2.0 license

(a) The following is the hypothesis:

H_0: The number of sites where the deers forage is distributed proportionately to the percentage of land.

H_A: The number of sites where the deers forage is not distributed proportionately to the percentage of land.

(b) We use a chi-square test for one way table.

(c) Below is the table based on the proportions:

```
wood_proportion <- round(426 * 0.048,2)
grassplot_proportion <- round(426 * 0.147,2)
forests_proportion <- round(426 * 0.396,2)
others_proportion <- round(426 * 0.409,2)
```

	wood	grassplot	Forests	others
proportions	20.45	62.62	168.7	174.23

To conduct the chi-square test the following conditions must be satisfied:

Independence - We assume that the cases are independent **Sample Size / Distribution** - All the cell counts are greater than 5 hence this conditions is also satisfied.

(d) calculating the chi-square and p-value:

```
Actual_sites <- c(4, 16, 67, 345)
Proportionate_sites <- c(20.45, 62.62, 168.70, 174.23)

k <- length(Actual_sites)
df <- k - 1

# Loop over the bin values to compute the chi2 test statistic
chi_square <- 0

for(i in 1:length(Actual_sites)){
  chi_square <- chi_square + ((Actual_sites[i] - Proportionate_sites[i])^2 / Proportionate_sites[i])
}

chi_square
```

```
## [1] 276.6286
```

```
# looking up p value
p_value <- pchisq(chi_square, df=df, lower.tail=FALSE)
p_value
```

```
## [1] 1.135815e-59
```

The chi-square value is so large the p-value is effectively 0. Therefore, we can conclude that there is 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.[63]

		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?

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

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

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

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

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

(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.[64] Do you agree with this statement? Explain your reasoning.

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

(b) The following is the hypothesis:

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) The following is the calculation:

```
depression_yes <- 2607 / 50739
depression_yes
```

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

```
depression_no <- 1 - depression_yes
depression_no
```

```
## [1] 0.9486194
```

The overall proportion of women who suffer from depression is 0.05 and the overall proportion of women who do not suffer from depression is 0.95

(d) Following are the calculations:

```
k <- 5
df <- k - 1

expected_count <- round(depression_yes * 6617,2)
expected_count
```

```
## [1] 339.99
```

```
cell_contribution <- (373 - expected_count)^2 / expected_count
cell_contribution
```

```
## [1] 3.204977
```

The exppected count is 339.99 and the cell contribution is 3.2049769.

(e) We can lookup the p-value using the below code:

```
p_value <- pchisq(20.93, df=df, lower.tail=FALSE)
p_value
```

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

(f) Since the p value is less than 0.05 we reject the null hypothesis in favor of the alternative hypothesis. Hence we can conclude that there is an association between caffeinated coffee consumption and depression.

(g) The study seems to be more observational in nature and not experimental. Given that the researchers had no control on the outcomes, there may have been other factors that may have led to the results. Hence I agree with the author.