IS606 Homework 6

Daniel Dittenhafer

October 27, 2015

6.6 2010 Healthcare Law (p313)

(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. More accurately, I would say we are 100% confident that 46% of the sample support 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. Given the 3% margin of error at 95% confidence and point estimate of 46%, the interval of 43% - 49% seems correct to me as an estimate of the population parameter.

(c) If we considered many random samples of 1,012 American, 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%. Not so much. Rather, 95% of those proportions would would contain the population proportion.

(d) The margin of error at a 90% confidence level would be higher than 3%. The reverse would be true. At a lower confidence level (90% vs 95%), the margin of error would narrow (be smaller) due to the use of a smaller Z score in margin of error computation.

6.12 Legalization of marijuana (Part I)

The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana shoudl 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. The 48% is a sample statistic because it is the portion of respondents to the survey… the 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. The following `R` code constructs the 95% confidence interval.

```r
n <- 1259
p <- 0.48
z <- qnorm(0.975)
se <- sqrt( (p * (1-p) )/ n)
# Assuming the sample observations are independent, and
# checking success-failure conditino of inference.
succeses <- p * n
succeses > 10
```

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

```
failures <- (1-p) * n
failures > 10
```

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

```
# Compute the margin of error
me <- z * (se)
me
```

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

```
# Construct the 95% Confidence Interval
ci95 <- data.frame(lb=p - me, ub=p + me)
ci95
```

```
##          lb        ub
## 1 0.4524033 0.5075967
```

(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. From the text, “the distribution of $\hat{p}$” is nearly normal when the distribution of 0’s and 1’ is not too strongly skewed for the sample size." In our case, the sample proportion is close to 0.5, and so the distribution of 0’s and 1’s would not be skewed very much at all. Additionally, we have already checked the success-failure condition.

(d) A news piece on this survey’s findings states, “Majority of American’s think marijuana should be legalized.” Based on your confidence interval, is this news piece’s statement justified? I don’t believe the news piece’s statement is justified. The proportion of American could be as low as 45.2%.

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 makde 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)

se <- me / z

# Compute the required n
n <- (p * (1-p)) / se^2
n
```

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

We would need to survey 2398 Americans to achieve a 2% margin of error for a 95% confidence interval.

6.28 Sleep deprivation, CA vs OR Part I (p319)

Calculate a 95% confidence interval for the difference between the proportions of Californias and Oregonians who are sleep deprived and interpret it in context of the data.

```r
pCA <- 0.08
pOR <- 0.088
pDiff <- pOR - pCA
nCA <- 11545
nOR <- 4691
# Compute standard error and margin of error for the proportion difference.
SE <- sqrt( ((pCA * (1 - pCA)) / nCA) +  ((pOR * (1 - pOR)) / nOR))
me <- qnorm(0.975) * SE
# Construct the 95% confidence interval.
ci95 <- data.frame(lb=pDiff - me, ub=pDiff + me )
ci95
```

```
##             lb         ub
## 1 -0.001497954 0.01749795
```

The 95% confidence interval on the difference in proportions between CA and OR is -0.001498 - 0.017498. 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 (p323)

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 there 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. The hypotheses associated with testing if barking deer prefer to forage in certain habitats follows:

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

H_a : The sites where barking deer forage is not evely distributed across the habitiats of the region.

$\alpha = 0.05$

The null hypothesis even distribution is shown in the table below:

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

b) What type of test care 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. Conditions and assumptions are described below:

Independence: Are the cases independent? We will have to assume so since no assertion is given in the description.

Sample size / distribution: In our expected cases scenario related to the null hypothesis, all habitats have at least 5 expected cases, therefore this condition is 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. The following `R` code works through the hypothesis test:

```r
habitats <- c(4, 16, 67, 345)
expected <- c(20.45, 62.62, 168.70, 174.23)
k <- length(habitats)
df <- k - 1
# Loop over the bin values to compute the chi2 test statistic
chi2 <- 0
for(i in 1:length(habitats))
{
  chi2 <- chi2 + ((habitats[i] - expected[i])^2 / expected[i])
}
chi2
```

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

```r
# Anyway, now check the chi2 test statistic and lookup p-val
pVal <- pchisq(chi2, df=df, lower.tail=FALSE)
pVal
```

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

The χ^2 value is so large the p-value is effectively 0. I conclude there is convincing evidence the barking deer forage in certain habitats over others.

6.48 Coffee and Depression (p325)

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 symptons at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionaires 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.

Depression	$\leq$ 1 cup/wk	2-6 cups/wk	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
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? 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). 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. 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. The expected counts for the two-way table are shown below:

Depression	$\leq$ 1 cup/wk	2-6 cups/wk	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Yes	627.851	340.1138	885.8276	631.706	122.4862	5.14%
No	11,587.149	6,276.8862	16,348.1724	11,658.294	2,260.5138	94.86%
Total	12,215	6,617	17,234	12,290	2,383	100%

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

expCnt <- 340.1138
cellContrib <- (373 - expCnt)^2 / expCnt
```

The contribution to the test statistic for the highlighted cell is 3.1798244.

e) The test statistic is $\chi^2 = 20.93$. What is the p-value? We lookup the p-value using the `pchisq` function:

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

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

f) What is the conclusion of the hypothesis test? Based on the p-value of ~ 0.0003 is less than 0.05, I conclude there is an association between caffeinated coffee consumption 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. I do agree with the statement. One study can provide an interesting theory, but peer review and replicating the results would lend more support to the conclusion. Additionally, additional caffeine might be other adverse affects on women not accounted for in this study.