DATA 606 HW6

Ahsanul Choudhury

October 30, 2016

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 010 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, the confidence interval refers to the entire population not just the sample.

(b) We are 95% confident that between 43% and 49% of Americans support the decision of theU.S. Supreme Court on the 2010 healthcare law.

True, with 3% margin of error 46% ranges bwtween 43% and 49% and we are 95% confident.

(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, the cofidence interval is a range of possible values for population porportion.

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

False, lowering the confidence level would lower the margin of 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.

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

48% is a sample statistic, of 1,259 US residents sampled 48% responded positively.

(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 <- .48
se <- sqrt((p*(1-p))/n)
z <- 1.96
me <- z * se
p - me; p + me
```

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

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

We are 95% confident that between 45.24% and 50.76% of US residents support legalization of marijuana.

(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.

The sample proportion is close to 0.5 which indicates data will not be too skewed in any side and will be normally disributed. The sample is randomly selected and is indepedent, size is fairly large but not greater then total population and both positive and negative are over 10.

(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?

We cannot confirm if the news piece's statement justified or not as our 95% confidence interval ranges from 45.24% to 50.76%, we can neither accept or reject null hypothesis.

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 <- .48
z <- 1.96
me <- .02
se <- me/z
n <- round(p * (1-p) / se^2)
n
```

```
## [1] 2397
```

we world need to survey 2397 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 insu"cient 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.

```
nca <- 11545
nor <- 4691
pca <- 0.08
por <- 0.088
pdiff <- por - pca

se <- sqrt( ((pca * (1 - pca)) / nca) +  ((por * (1 - por)) / nor))
se
```

```
## [1] 0.004845984
```

```
z <- 1.96
me <- z * se

ci <- (pdiff - me); (pdiff + me )
```

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

```
ci
```

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

We are 95% confident the difference between Californians and Oregonians who are sleep deprived is between -0.14% and 1.75%. The interval overlaps 0, which indicates there are no significant diffrence in proportion.

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.

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

H_0: Barking deer has no preference in foraging on certain habitast over others.

H_a: Barking deer has preference.

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

We can use a chi-square test to answer this research question.

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

There are two conditions that must be satisfied for a chi-square test:

- Independence: All 4 deer habitat variables are independent of each other
- Sample size / distribution: Apart from the woods habitat variable all the other habitats are over expected minimum of 5 cases. The woods habitat has only 4.8 cases, we would assume in this case this is an acceptable count.

(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.

```r
tot <- 426
woods <- c(4, tot * .048)
grassplot <- c(16, tot * .147)
forests <- c(67, tot * .396)
other <- c(345, tot * .409)
total <- c(tot, tot)
df <- data.frame(rbind(woods, grassplot, forests, other, total))
colnames(df) <- c("observed", "expected")
knitr::kable(df)
```

	observed	expected
woods	4	20.448
grassplot	16	62.622
forests	67	168.696
other	345	174.234
total	426	426.000

```r
chisq.test(df)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df
## X-squared = 138.73, df = 4, p-value < 2.2e-16
```

The p-value is almost 0 which points us to conclude barking deer in fact prefer to forage in certain habitat.

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.

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

A chi-square 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).

H_0:There is no relation between amount of coffee drunk by woman and depression. H_a:There is a relation between amount of coffee drunk by woman and depression.

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

```
tot <- 50739
totcd <- 2607

pcd <- 2607/50739
pcd
```

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

```
pncd <- (1-pcd)
pncd
```

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

Overall proportion of women who do suffer from depression is 5.14% and Overall proportion of women who do not suffer from depression is 94.87%.

d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $\frac{(O_k - E_k)^2}{E_k}$

```
k <- 5
df <- k - 1
expcnt <- 340.1138
cellcontrib <- (373 - expcnt)^2 / expcnt
cellcontrib
```

```
## [1] 3.179824
```

The contribution of this cell to the test statistic 3.18.

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

***p-value**

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

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

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

Based on the p-value in previous section we reject null hypothesis and conclude there is a relation between amount of coffee drunk by woman 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.64 Do you agree with this statement? Explain your reasoning.

I agree, although this is a interesting theory drinking coffee may have other adverse effects that needs further studies.