DATA 606 Homework 6

Ilya Kats

April 9, 2017

6.6 2010 Healthcare Law.

a. **FALSE**. A confidence interval is constructed to estimate the population proportion, not the sample proportion.

b. **TRUE**. A confidence interval is constructed to estimate the population propotion which is 46% ± 3%, or 43% to 49%.

c. **TRUE**. By the definition of the confidence level.

d. **FALSE**. At a 90% confidence level, the margin of error will be less than 3% and confidence interval will be more narrow since we need to be less confident that the population probability is within the interval.

6.12 Legalization of marijuana, Part I.

a. 48% is a sample statistic that estimates the population parameter since it was derived from the sample data.

b.

```
n <- 1259
p <- 0.48
SE <- sqrt((p * (1-p))/n)
( ME <- 1.96 * SE )
```

```
## [1] 0.02759723
```

At 95% confidence level, the confidence interval is $(0.4524028, 0.5075972)$. We are 95% confident that the proportion of US residents who think marijuana should be made legal is between 45.24% and 50.76%.

c. Although we have no information about how residents were selected for the survey, it is reasonble to assume that they were selected using a simple random process. Additionally, at 1259 observations sample size is definitely lower than 10% of the population. Observations can be considered independent. We have obeserved $pn = 0.48 * 1259 = 604.32$ and $(1 - p)n = 0.52 * 1259 = 654.68$ successes and failures. Both are over 10, so normal model is a good approximation.

d. News piece's statement is not justified. Based on the confidence interval it is possible that the population probablity is over 50%, but it is also possible that it is noticeably lower than 50% (in fact most of confidence interval is below 50%).

6.20 Legalize Marijuana, Part II.

```
p <- 0.48
ME <- 0.02

# ME = 1.96 * SE
SE <- ME / 1.96

# SE = sqrt((p * (1-p)) / n)
# SE^2 = (p * (1-p)) / n
( n <- (p * (1-p)) / (SE^2) )
```

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

We need to survey 2398 Americans.

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

The sample was selected simple random process and it repsents less than 10% of the population. We have at least 10 successes and failures for both states, so the distribution can be approximated using the normal model.

```
p_ca <- 0.08
p_or <- 0.088

p <- p_ca - p_or

n_ca <- 11545
n_or <- 4691

SE2_ca <- (p_ca * (1-p_ca)) / n_ca
SE2_or <- (p_or * (1-p_or)) / n_or

SE <- sqrt(SE2_ca + SE2_or)
( ME <- 1.96 * SE )
```

```
## [1] 0.009498128
```

The confidence interval is $(-0.0174981, 0.0014981)$.

We are 95% confident that the difference between the proportion of Californians and Oregonians who are sleep deprived is between -0.0174981 and 0.0014981.

6.44 Barking deer.

```
observed <- c(4, 16, 61, 345, 426)
expected_prop <- c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396, 1)
expected <- expected_prop * 426
deer <- rbind(observed, expected)
colnames(deer) <- c("woods", "grassplot", "forests", "other", "total")
deer
```

```
##           woods grassplot forests   other total
## observed  4.000    16.000  61.000 345.000   426
## expected 20.448    62.622 168.696 174.234   426
```

a. H_0 : Barking deer has no preference of certain habitats for foraging. H_A : Barking deer prefers some habitats over others for foraging.

b. We can use chi-square goodness of fit test to this hypothesis.

c. Although it is possible that something in the behavior of barking deer makes cases dependent on each other, it is more likely that the cases are independent. Each expected value is above 5.

d.

```
k <- 4
df <- k-1

chi2 <- sum(((deer[1,] - deer[2,])^2)/deer[2,])
( p_value <- 1 - pchisq(chi2, df) )
```

```
## [1] 0
```

The $p-value$ is practically 0. Even at 99% confidence level, this value is below the significance level, so we reject the null hypothesis. Barking deer prefers to forage in some habitats over others.

6.48 Coffee and Depression.

a. Chi-square test for the two-way table can be used to evaluate if there is an association between coffee intake and depression.
b. H_0 : The risk of depression in women is the same regardless of amount of coffee consumed. H_A : The risk of depression in women varies depending on amount of coffee consumed.
c. Proportion of women who suffer from depression is $2607/50739 = 0.0513806$, and proportion of women who do not suffer from depression is $48132/50739 = 0.9486194$.
d. Highlighted cell show an observed count for women who suffer from depression and drink 2-6 cups of coffee per week. $Expected\ Count_{1,2} = \frac{2607*6617}{50739} = 339.9853958$
e. For $\chi^2 = 20.93$ and $df = (2-1)*(5-1) = 4$, the $p-value$ is 0.0003.

```
1 - pchisq(20.93, 4)
```

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

f. Even with a significance of 0.01, the p-value is less, so we reject the null hypothesis. The data provide convincing evidence that there is some difference in the risk of depression for women based on various levels of coffee consumption.
g. I agree with author's statement because this was an observational study. It cannot be used to demonstrate causation.