Lin-HW6

Bin Lin

2016-10-30

6.6 2010 Healthcare Law.

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

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

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, if we have n random samples of 1012 American, we will end up with n 95% confidence interval. 95% of those confidence intervals will contain the true population proportion.

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

False, since the we are less confident, which means the interval is narrower, so it is less likely to catch the population mean. If the interval is narrower, so that the margin of error will have to be smaller.

6.12 Legalization of marijuana, Part I. (a) Is 48% a sample statistic or a population parameter? Explain.

It is sample statistics, because the survey only asked 1259 US residents which does not represent the entire population of USA. There are way more than 1259 people who were not got interviewed.

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.

```
SE <- sqrt(0.48 * (1 - 0.48) / 1259)
z <- 1.96
lower_bound <- 0.48 - 1 * z * SE
upper_bound <- 0.48 + z * SE
c (lower_bound, upper_bound)
```

```
## [1] 0.4524028 0.5075972
```

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.

Distribution does have to be 100% normal. As long as all the conditions for confidence interval are met. For example, the sample need to be independent within groups and between groups. The success-failure condition should also need to be satisfied (true in this case). Furthermore, the distribution can even bit little skwed if the sample size is over 30 (also true in this case).

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? No, because the confidence interval across 50%, which means it could be majority or minority of Americans think mariyuana should be legalized.

6.20 Legalize Marijuana, Part II. 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 ?

We need to survey at least 2398 americans

```
margin_of_error <- 0.02
z <- 1.96
n <- z^2 * 0.48 * (1 - 0.48) / margin_of_error^2
n
```

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

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

The 95% confidence interval is (-0.001497954, 0.017497954). This confidence interval means californian who are sleep deprived can be 0.15% less than oregonian, or it can be 1.7% higher than oregonian. The confidence interval across 0, so that it is not statistically sginificant to claim who has the higher or lower sleep deprivation rate.

```
SE <- sqrt((0.08 * (1 - 0.08) / 11545) + (0.088 * (1- 0.088) / 4691))
SE
```

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

```
point_estimate <- 0.088 - 0.08
z <- qnorm(0.975)
lower <- point_estimate - 1 * z * SE
upper <- point_estimate + 1 * z * SE
c(lower, upper)
```

```
## [1] -0.001497954  0.017497954
```

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

H0: Barking deer does not prefer forage in certain habitats over others. HA: Barking deer does prefer forage in certain habitats over others.

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

Chi-square (X^2) statistics test. Because we want to observe if any group is significantly different from the other groups. In addition, we are also dealing with counts for each group.

c. Check if the assumptions and conditions required for this test are satisfied. There are 3 conditions required.

1. Independence: each deer's behavior does not influence other deer's.
2. Sample size: each particular scenaria (group) must have at least 5 expected cases. But for the woods habitat, it only has count of 4, but all the other habitat has counts higher than 5.
3. df > 1: degree of freedom in this case is 3.

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.

H0: Barking deer does not prefer forage in certain habitats over others. HA: Barking deer does prefer forage in certain habitats over others. After conducting the chi-statistic testing, we can find out the chi-square value is 98, so that the corresponding p-value is closed to 0 when degree of freedom is 3. In short, we can reject the null hypothesis and conclude the barking deer does prefer forage in certain habitats over others.

```
100 - 4.8 - 14.7 - 39.6
```

```
## [1] 40.9
```

```
chi_statistic1 <- ((4 / 426) - 4.8)^2 / 4.8
chi_statistic2 <- ((16 / 426) - 14.7)^2 / 14.7
chi_statistic3 <- ((61 / 426) - 39.6)^2 / 39.6
chi_statistic4 <- ((345 / 426) - 40.9)^2 / 40.9
chi_total <- chi_statistic1 + chi_statistic2 + chi_statistic3 + chi_statistic4
chi_total
```

```
## [1] 98.01667
```

```
p <- pchisq(chi_total, df=3, lower.tail=FALSE)
p
```

```
## [1] 4.148642e-21
```

6.48 Coffee and Depression. (a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression? chi-square test of independence for two-way table.

b. Write the hypotheses for the test you identified in part (a). H0: There is not association between coffee intake and depression HA: There is association between coffee intake and depression

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

5.138% of women suffer from depression, while 94.86% of women do not suffer.

```
2607 / 50739
```

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

```
1 - 2607 / 50739
```

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

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 count for the highlighted cell is 339.9854. The contribution of this cell to the test statistic is 3.206

```
(373 - (6617 * 2607 / 50739))^2 / ((6617 * 2607) / 50739)
```

```
## [1] 3.205914
```

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

p-value equals 0.0003269507

```
df <- (5 - 1) * (2 - 1)
p <- pchisq(20.93, df=df, lower.tail=FALSE)
p
```

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

f. What is the conclusion of the hypothesis test? Since the p-value is smaller than 0.05, therefore, we can reject the null hypothesis and conclude that there is association between coffee intake 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 agree with one of the authors, because there might be confounding variables that we did not take into consideration. For example, maybe those women who drink a lot of coffee also smoke a lot. Therefore, it could actually because of cigarette that is associated with depression rather than coffee.