Homework_6

Salma Elshahawy

10/19/2019

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

(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! because we are 100% confidence for this particular sample

(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! because 43% and 49% is the confidence interval when generalizing to the whole population +/- 3% margin error

(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%.

TRUE! because we were 95% confidence with ME of 3% for whole population will be in this margin

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

FALSE! when decreasing the confidence interval, the margin of error will decrease, because we are getting less porbability

---

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 statistics, because there is no confidence interval to be generalized for whole population and the sample size was 1,259 of US residents

(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.

```r
n <- 1259
p <- 0.48
SE <- sqrt((p*(1-p)/n))
ME <- 1.96*SE
ME_per <- ME * 100
upper <- 48 + ME_per
lower <- 48 - ME_per
c(upper, lower)
```

```
## [1] 50.75972 45.24028
```

(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.

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) 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?

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. 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 ?

```r
ME <- 0.02
P <- 0.48
SE <- ME/1.96
n <- (p*(1-p)) / (SE)^2
n
```

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

The sample size should be 2,398 responders

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 insucient 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.

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_org <- 0.088
n_org <- 4691

p_cal <- 0.08
n_cal <- 11545

SE_org <- p_org*(1-p_org)/n_org
SE_cal <- p_cal*(1-p_cal)/n_cal

SE <- sqrt(SE_cal + SE_org)
ME <- 1.96 * SE
ME
```

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

```
p <- p_cal - p_org

upper <- p + ME
lower <- p - ME
c(upper, lower)
```

```
## [1]  0.001498128 -0.017498128
```

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.

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

H0 –> There is no preference for forage deer to habitat, Ha –> There are a preference for forage deer to habita

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

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

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

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) 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
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
```

```r
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. 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?

Chi-square test for the two-way table can be used to evaluate if there is an association between coffee intake and depression.

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

H0 –> The risk of depression in women is the same regardless of amount of coffee consumed.HA –> The risk of depression in women varies depending on amount of coffee consumed.

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

```r
p_women_dep <- 2697/50739
p_women_dep
```

```
## [1] 0.05315438
```

```
p_women_nodep <- 48123/50739
p_women_nodep
```

```
## [1] 0.948442
```

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

Highlighted cell show an observed count for women who suffer from depression and drink 2-6 cups of coffee per week.

```
exp_count <- (2607 * 6617)/50739
exp_count
```

```
## [1] 339.9854
```

(e) The test statistic is X2 = 20.93. What is the p-value

```
df <- (2-1)*(5-1)
p_value <- 1 - pchisq(20.93, df)
p_value
```

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

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

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) 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 author's statement because this was an observational study. It cannot be used to demonstrate causation.