DATA606 Chapter 6 Inference for Categorical Data

Alexis Mekueko

10/9/2020

Github link: https://github.com/asmozo24/DATA606_Homework6

Web link: https://rpubs.com/amekueko/674072

```
library(tidyverse) #loading all library needed for this assignment
library(openintro)
library(infer)
library(gplots)
#head(fastfood)
#library(readxl)
#library(data.table)
#library(readr)
#library(plyr)
#library(dplyr)
#library(dice)
# #library(VennDiagram)
# #library(help = "dice")
#library(DBI)
#library(dbplyr)

#library(rstudioapi)
#library(RJDBC)
#library(odbc)
#library(RSQLite)
#library(rvest)
#library(stringr)
#library(readtext)
#library(ggpubr)
#library(fitdistrplus)
#library(ggplot2)
#library(moments)
#library(qualityTools)
library(normalp)
#library(utils)
#library(MASS)
#library(qqplotr)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
```

```
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

```
getLabs()
```

```
##  [1] "Lab1"  "Lab2"  "Lab3"  "Lab4"  "Lab5a" "Lab5b" "Lab6"  "Lab7"  "Lab8"
## [10] "Lab9"
```

```
#library(StMoSim)
```

Github Lin https://github.com/asmozo24/DATA606_Homework6 Web Link:

Exercice 1 : 6.48 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% con

dence 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.49

#(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. Answe: False. Based on the of the problem. 46% of 1,012 Americans agree with this decision and that is the 95% confidence level.

#(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. Answer: true. I want to agree on the fact that for a large sample population (in this case population = Americans) the points of estimate of the population proportion will be about 46% ± % 3% margin of error.so, (43, 49)

#(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%. Answer: true, we went through this question in lab5. Running random sample of 1,012 Americans many time, then 95% of those sample proportions will be between 43% and 49%. Actually, we found out if the rerun is at large scale then, we will be about the 46%.

#(d) The margin of error at a 90% confidence level would be higher than 3%. Answer: False, from lab5 we found out the margin error decreases if the confidence level decrease and the confidence interval will be narrowed.

Exercice 2: 6.10 Legalization of marijuana, Part I.

The 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.

Answer: this is a sample statistic (a defined number of observations randomly selected from a population). I think the the labor bureau or US homeLand department probably has the U.S. population parameter.

(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.

Answer: 95% confidence level for the proportion of US residents who think marijuana should be made legal is (45.24, 50.76)

(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. Answer: true , we will assume the random sample is independant and follow normal distribution since there is no other reason to think otherwise.

(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? Answer: I don't have evidence that this news piece's statement is justified since we don't know what the majority of the Americans is.

```
# (b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should

n <- 1259
p <- 0.48
se <- sqrt((p*(1-p))/n)
CI_high <- p + 1.96*se
CI_high
```

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

```
CI_low <- p- 1.96*se
CI_low
```

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

Exercice 3: 6.16 Legalize Marijuana, Part II.

As discussed in Exercise 6.10, the 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? Answer: se = (0.50-0.48)/1.96 = sqrt((0.48*(1-0.48))/n).....n = 2,399 ... so about 2400 Americans.

```
# If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many American

# n <- ?
p <- 0.48
se <- sqrt((p*(1-p))/n)
CI_high <- p + 1.96*se
CI_high <- 0.50

n = (((CI_high - p)/1.96)^2)*(1/(p*(1-p)))
```

##Exercice 4: 6.22 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 insufficient 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.

#Answer: this is where Diffference of two proportions applies. Assume all conditions for this formula is met. SE = sqrt((p1(1-p1)/n1)+(p2(1-p2)/n2))... point estimate ± z* X SE..... p1-p2 ± z* X SE

We ARE 95% confidence that the proportions of Californians and Oregonians HAVE a difference of -0.15% to 1.75% percentage point for sleep deprived

```
# Diffference of two proportions

n2 <- 11545
n1 <- 4691
p2 <- 0.08
p1 <- 0.088
z <- 1.96

SE <- sqrt((p1*(1-p1)/n1)+(p2*(1-p2)/n2))

CI_hight <- p1-p2 + z * SE
CI_hight
```

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

```
CI_low <- p1-p2 - z * SE
CI_low
```

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

Exercice 4; Barking deer.(6.34, p.239) Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make 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.

Woods Cultivated_grassplot Deciduous_forests Other Total 4 16 61 345 426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others. Answer: H_null : Banking deer does not prefer to forage in certain habitat H_alt : Banking deer does prefer to forage in certain habitat

(b) What type of test can we use to answer this research question? Answer: chi-square will fit this test

(c) Check if the assumptions and conditions required for this test are satisfied. Answer: Indepedence: does each case that contributes a count to the table must be independent of all the other cases in the table in this banking deer? yes, observations are not dependent to each other. sample size / distribution: do we have at least 05 expectation cases in this banking deer? yes

(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. ANswer: pvalue for this chi-square test is 0, therefore we reject the H_null (null hypothesis)

```
Areas <-  c("Woods", "Cultivated_grassplot", "Deciduous_forests", "Other")
values <-  c(4, 16, 61, 345)
# Bank_deer <-  cbind (Ares, values)
Bank_deer <- data.frame(Areas, values)

Bank_deer1 <- Bank_deer %>%
  pivot_wider(names_from = Areas, values_from = c(values))
Bank_deer1
```

```
## # A tibble: 1 x 4
##   Woods Cultivated_grassplot Deciduous_forests Other
##   <dbl>                <dbl>             <dbl> <dbl>
## 1     4                   16                61   345
```

```
other_exp = 1-0.048-0.147-0.396
expected_prop <- c(0.048, 0.147, 0.396, other_exp)
expected <- expected_prop * 426
Bank_deer2 <- length(values) - 1
chi_test <- (values - expected)^2 / expected
chi_test <-sum (chi_test)

pvalue <- 1-pchisq(chi_test, Bank_deer2)
pvalue
```

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

Exercice 5 6.50 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.

	Caffeinated coffee consumption				
	< 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	>4 cups/day	Total
Clinical Yes	670	373	905	564	95	2,607
depression 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? Answer: We have a two-way table here, so 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). Answer: H_null: There is no association between caffeinated coffee consumption and risk of depression in women H_alt : There is association between caffeinated coffee consumption and risk of depression in women

(c) Calculate the overall proportion of women who do and do not suffer from depression. Answer: The total proportion of women who do and do not is 2607+48132 = 50739 the overall proportion of women who do suffer from depression is (2607/50739)*100 = 5.14 % the overall proportion of women who do not suffer from depression is (48132/50739)*100 = 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. Answer: the highlighted cell from the Caffeinated coffee consumption table is cups/week which totla is 6617... So, the contribution of this cell to the test statistic is expected_count = 6617*0.0514 = 340, ((observed_count - expected_count)^2)/expected_count = 3.18

(e) The test statistic is X^2 = 20.93. What is the p-value? df <- (5-1)*(2-1), chi_test2 <- 20.93, pvalue <- 1- pchisq(chi_test2, df)

p-value is 0 .

(f) What is the conclusion of the hypothesis test? Answer: since the p-value is 0, we reject the null hypothesis

(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. ANswer: I agree, the study reject the null hypothesis, but to what extinct we don't know. So, we cannot generalize this study to a large population.

```
observed_count <- 373

expected_count <- 6617*0.0514
x2 <- ((observed_count - expected_count)^2)/expected_count
df <- (5-1)*(2-1)
chi_test2 <- 20.93
pvalue <-  1- pchisq(chi_test2, df)
```