Problem Set 8

Megan Hoeksema

## Introduction

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your work should be based on the data’s being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots. Each part is worth 10 points.

These questions were rendered in R markdown through RStudio (<https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>, <http://rmarkdown.rstudio.com> ).

## 1

Recall that if is a random variable with the standard Normal distribution, then the distribution of (the result of applying the function to the outcomes of ) has a distribution. If are independent random variables with the standard Normal distribution, then the distribution of is a distribution.

Also, recall the fact from the week 5 slides that if and are random variables and then .

### 1.a

(5 points)

Please compute the expected value of , the distribution with 1 degree of freedom.

### 1.b

(5 points)

Based on your answer to 1.a, please compute the expected value of , the distribution with n degrees of freedom.

## 2

This problem set uses 2019 data primarily for Denver county accessed through IPUMS-USA, University of Minnesota, www.ipums.org ,

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 7.0 [dataset]. Minneapolis, MN: IPUMS, 2020. <https://doi.org/10.18128/D010.V10.0>

The PUMA-to-county restriction was done using MABLE, <http://mcdc.missouri.edu/websas/geocorr12.html>

This problem set uses a subsample of demographic data for Denver.

The sample was drawn according to the values in the variable “perwt”. This is a weight value provided by the US Census Bureau to correct for differences between the sampled population and the target population. It is called a sample weight or an expansion weight.

It can be thought of as the number of people in Colorado that the one observation represents in terms of demographic characteristics. For example, if you add all the weights in the original sample for all of Colorado, you will get an approximation of the population of Colorado in the sample year. If you multiply the “age” variable by “perwt” then divide by the sum of the “perwt” values, you will get an approximation of the average age in the state, whether or not the ages of the cases are present in the same proportion in the sample as in the population.

Thus a sample drawn using “perwt” as the probability will be a better approximation of the population than a simple random sample in which each case has an equal likelihood of being selected.

The category “educ”=7 corresponds to 1 year of college. The category “educ”=10 corresponds to 4 years of college.

Samples of size 40 are drawn from the responses with “educ”=7 and with “educ”=10 according to the weights in the data set and saved in “dat\_7\_10.RData”.

### IPUMS Data

Read in the subsample of the IPUMS data.

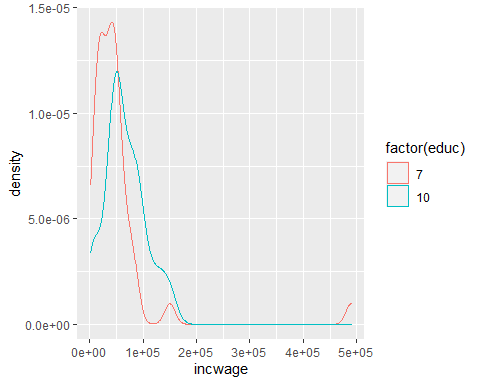
load("dat\_7\_10.RData")

### 2.a

(10 points)

Please run and interpret a Mann-Whitney U test comparing “incwage” for the observations with “educ” equal to 7 and with “educ” equal to 10. In your interpretation, please consider the case in which you treat the distributions of the two populations as related by translation and the case in which you don’t make this assumption.

ggplot(dat.7.10,aes(x=incwage,color=factor(educ)))+geom\_density()



wilcox.test(incwage~educ,dat.7.10)

## Warning in wilcox.test.default(x = c(78000L, 78000L, 48000L, 60000L, 45000L, :  
## cannot compute exact p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: incwage by educ  
## W = 467.5, p-value = 0.001395  
## alternative hypothesis: true location shift is not equal to 0

### P-value is very small. We can reject the null hypothesis in favor of the alternative hypothesis.

### 2.b

(10 points)

Please run a Mann-Whitney U test comparing log(incwage) for the observations with “educ” equal to 7 and with “educ” equal to 10 and compare to the result in part a. Please explain what you observe about the two tests.

wilcox.test(log(incwage)~educ,dat.7.10)

## Warning in wilcox.test.default(x = c(11.2644641056717, 11.2644641056717, :  
## cannot compute exact p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: log(incwage) by educ  
## W = 467.5, p-value = 0.001395  
## alternative hypothesis: true location shift is not equal to 0

### There is no difference between the two. The p-values are the same.

## 3

The raw data in this question is the “Pew Research Center’s American Trends Panel” Wave 69 Field dates: June 16 – June 22, 2020 Topics: Coronavirus tracking, politics, 2020 Census data and questionnaire downloaded 3/4/2021 from <https://www.pewresearch.org/politics/dataset/american-trends-panel-wave-69/>

The codebook was downloaded 3/5/2021 from <https://www.pewresearch.org/wp-content/uploads/2018/05/Codebook-and-instructions-for-working-with-ATP-data.pdf>

The Pew Research Center provides sample weights in the variable “WEIGHT\_W69”. These serve a similar purpose to the “perwt” variable in the IPUMS data, though these wieghts have the effect of readjusting the proportions of demographic groups in the sample to be approximately the proportions in the target population when the responses are viewed as representing the number of people given by the weight. The weights add up to the number of responses in the study.

The code below draws a sample from the full response set with probability based on the weight. Please use “dat.sub” in the questions below. The data frame “dat.sub” is provided with the assignment.

# dat.pew<-data.frame(read.spss("W69\_Jun20/ATP W69.sav"))  
# sum(dat.pew$WEIGHT\_W69)  
# set.seed(1234)  
# sub.index<-sample(1:nrow(dat.pew),200,prob = dat.pew$WEIGHT\_W69)  
# dat.sub<-dat.pew[sub.index,]  
# save(dat.sub,file="dat\_sub.RData")  
load("dat\_sub.RData")

The code below generates a contingency table for the answers to the question “How much of a problem do you think each of the following are in the country today?” applied to the coronavirus outbreak by the age category of the respondent. The respondents who refused to supply their age or an answer to the question are omitted.

For intuition, the percent within each age range selecting each response is shown.

If you would prefer to investigate the independence of another pair of variables, you may generate your own contingency table and base your answers to 3.a and 3.b on your own table.

t<-table(dat.sub$F\_AGECAT,dat.sub$NATPROBS\_b\_W69)  
t<-t[1:4,1:4] # Drop the "Refused" row and column  
  
round(100\*t/rowSums(t),0)

##   
## A very big problem A moderately big problem A small problem  
## 18-29 66 18 16  
## 30-49 39 47 12  
## 50-64 63 23 12  
## 65+ 63 28 9  
##   
## Not a problem at all  
## 18-29 0  
## 30-49 2  
## 50-64 2  
## 65+ 0

### 3.a

(10 points)

Please use the test to test the independence of the probability distribution with the outcomes in the rows and the probability distribution with the outcomes in the columns for the table you chose. Is this an appropriate test for your table? If not, why not? If so, please respond to the question of whether the data are consistent with the null hypothesis that the row distribution and the column distribution are independent.

chisq.test(t)

## Warning in chisq.test(t): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: t  
## X-squared = 16.261, df = 9, p-value = 0.06163

### This is not an appropriate test. According to the guidelines presented in the async lectures, the expected value should be at least 5.

### 3.b

(10 points)

Please use Fisher’s exact test to test the independence of the probability distribution with the outcomes in the rows and the probability distribution with the outcomes in the columns for the table you chose. Is this an appropriate test for your table? If not, why not? If so, please respond to the question of whether the data are consistent with the null hypothesis that the row distribution and the column distribution are independent. (Setting cache=TRUE means that the code in the block will not be reevaluated on subsequent “knit” applications unless the code in the block is changed. This speeds up text editing once the calculations are in place, but shouldn’t be used before than because the calculations won’t be updated to reflect changes elsewhere.)

fisher.test(t)# can be slow

##   
## Fisher's Exact Test for Count Data  
##   
## data: t  
## p-value = 0.03326  
## alternative hypothesis: two.sided

### This is an appropriate test. It gives us strong evidence against the null hypothesis.