Problem Set 8

C. Durso

15 August, 2020

These questions were rendered in R markdown through RStudio (<<https://urldefense.com/v3/__https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf__>;!!NCZxaNi9jForCP\_SxBKJCA!H2o4AVX3C2X-TmYBjS04rM\_qVhCWeHQ0mcn9vyoT7AvlzUgSuJFJ379LaZsNInZefg$ >, <<https://urldefense.com/v3/__http://rmarkdown.rstudio.com__>;!!NCZxaNi9jForCP\_SxBKJCA!H2o4AVX3C2X-TmYBjS04rM\_qVhCWeHQ0mcn9vyoT7AvlzUgSuJFJ379LaZtOrdH-oA$ > ).

Please complete the following tasks, in R where applicable. Please generate a solution document in R markdown and upload the rendered .doc, .docx, or .pdf document. You may add hand computations to a .doc or .docx if you prefer. In the rendered document, please show your code. That is, don’t use “echo=FALSE”.

In either case, your work should be based on the data’s being in the same folder as the R files. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

## 1. (5 points)

Let and be independent random variables with distributions with and degrees of freedom respectively. What is the distribution of ?. (5 points)

The distribution of X+Y is itself a chi-square distribution with n+m degrees of freedom.

## 2.

Correlation may be reported as a measure of the linear association between two variables. An understanding of typical correlations between independent variables aids in understanding these reports.

Consider two iid samples and , each of size , from a standard Normal distribution. The following question investigates the range of correlations that are typical for such samples.

### 2a.(10 points)

Perform a numerical experiment to estimate the quantiles of the sample correlations between two iid samples of sizes 10, 20, 50, and 100 from the standard Normal distribution.Display the quantiles in a table. Suggestion: for each size,simulate the process of drawing a pair of samples calculating their correlation many times then report the requested quantiles of the observed correlations.(5 points)

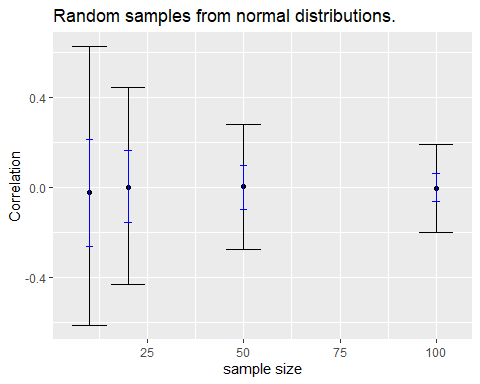
n10\_cors<- rep(NA, 1000)  
n20\_cors<- rep(NA, 1000)  
n50\_cors<- rep(NA, 1000)  
n100\_cors<- rep(NA, 1000)  
  
for (i in 1:1000) {  
x1 <- c(rnorm(n=10))  
x2 <- c(rnorm(n=10))  
x3 <- c(rnorm(n=20))  
x4 <- c(rnorm(n=20))  
x5 <- c(rnorm(n=50))  
x6 <- c(rnorm(n=50))  
x7 <- c(rnorm(n=100))  
x8 <- c(rnorm(n=100))  
  
  
n10\_cors[i] <- cor(x1,x2)  
n20\_cors[i] <- cor(x3,x4)  
n50\_cors[i] <- cor(x5,x6)  
n100\_cors[i] <- cor(x7,x8)  
}  
  
  
n10 <- quantile(n10\_cors,c(.025,.25,.5,.75,.975))  
n20 <- quantile(n20\_cors,c(.025,.25,.5,.75,.975))  
n50 <- quantile(n50\_cors,c(.025,.25,.5,.75,.975))  
n100 <- quantile(n100\_cors,c(.025,.25,.5,.75,.975))  
quantileframe <- data.frame(n10,n20,n50,n100)%>% add\_row(n10=10L, n20=20L,n50=50L,n100=100L)   
  
rownames(quantileframe) <- c(".025",".25",".5",".75",".975","n")  
m1 <- t(quantileframe)  
qf2 <- data.frame(r1= row.names(m1), m1, row.names=NULL)   
colnames(qf2) <- c("nvals",".025",".25",".5",".75",".975","n")  
qf2

## nvals .025 .25 .5 .75 .975 n  
## 1 n10 -0.6125346 -0.25915657 -0.021262910 0.2159205 0.6285852 10  
## 2 n20 -0.4297880 -0.15391734 0.001897988 0.1641617 0.4445392 20  
## 3 n50 -0.2764692 -0.09639986 0.007332346 0.1014039 0.2809407 50  
## 4 n100 -0.1975877 -0.06017613 -0.004777285 0.0621869 0.1914183 100

### 2b.(10 points)

Display the quantiles for each in a plot that has the value of on the x-axis, and the correlation quantiles on the y-axis. Please use bars to indicate the range between the 0.25 and 0.75 quantile and the range between the 0.025 and 0.975 quantile. Please plot a point at the 0.5 qunatile for each . Please title the plot and label the axes with informative text. You may find “geom\_errorbar” useful. You can position ticks on the x-axis using “scale\_x\_continuous(breaks=c(10,20,50,100))” for example.

qfplot <- ggplot(qf2, aes(x = n, y = qf2[,4])) + geom\_point()+ geom\_errorbar(aes(ymin=qf2[,2], ymax=qf2[,6]))+geom\_errorbar(aes(ymin=qf2[,3],ymax=qf2[,5]),width=2, color="blue")+ggtitle("Random samples from normal distributions.") +  
 xlab("sample size") + ylab("Correlation")  
qfplot



## 3. (5 points each)

The data in “pew\_sampled\_data.RData” come from the PEW RESEARCH CENTER 2017 STEM SURVEY downloaded from <https://urldefense.com/v3/__https://www.pewsocialtrends.org/dataset/2017-pew-research-center-stem-survey/__>;!!NCZxaNi9jForCP\_SxBKJCA!H2o4AVX3C2X-TmYBjS04rM\_qVhCWeHQ0mcn9vyoT7AvlzUgSuJFJ379LaZupcRJ\_5w$ 8/5/2019.

The data here are sampled from the full survey data according to the survey weights.

load("pew\_sampled\_data.RData")

### 3a.

The variable recording whether the respondent was a STEM worker or a non-STEM worker is “WORKTYPE\_FINAL”.

The variable “SCH7” records the respondents’ selection based on the following question and options:

What’s the main reason many young people don’t pursue college degrees in science, technology, engineering and mathematics? Is it mostly because.

1 They think these subjects are too hard

2 They think these subjects are too boring

3 They think these subjects are not useful for their careers

4 Some other reason (please specify)

5 Refused

Please create a 2-way table that gives the proportion of respondents among Stem workers that selected each response and the proportion of respondents among non-Stem workers that selected each response.

For the purposes of table display, please shorten the responses to “too hard”, “too boring”, “not useful”, “other”.

stem <- samp$SCH7[which(samp$WORKTYPE\_FINAL=="STEM Worker")]  
nonstem <- samp$SCH7[which(samp$WORKTYPE\_FINAL!="STEM Worker")]  
  
stemcounts <- c(length(which(stem=="They think these subjects are too hard")),length(which(stem=="They think these subjects are too boring")),length(which(stem=="They think these subjects are not useful for their careers")),length(which(stem=="Some other reason (please specify)")),length(which(stem=="Refused")))  
stemprops <- round((stemcounts/293),digits=2)  
  
nonstemcounts <- c(length(which(nonstem=="They think these subjects are too hard")),length(which(nonstem=="They think these subjects are too boring")),length(which(nonstem=="They think these subjects are not useful for their careers")),length(which(nonstem=="Some other reason (please specify)")),length(which(nonstem=="Refused")))  
nonstemprops <- round((nonstemcounts/293),digits=2)  
  
  
countframe <- data.frame(stemcounts,nonstemcounts)  
rownames(countframe) <- c("Too Hard","Too Boring","Not Useful","Other","Refused")  
propframe <- data.frame(stemprops,nonstemprops)  
rownames(propframe) <- c("Too Hard","Too Boring","Not Useful","Other","Refused")  
propframe

## stemprops nonstemprops  
## Too Hard 0.58 0.51  
## Too Boring 0.14 0.11  
## Not Useful 0.15 0.25  
## Other 0.11 0.12  
## Refused 0.02 0.01

## 3b.

Please do a test and a Fisher’s Exact Test to determine if the responses to “SCH7” are independent of “WORKTYPE\_FINAL”, and compare the results.

The two tests generated similar P values, .04 & .037, respectively. In each case, we would reject the null hypothesis. There is compelling statistical evidence that the two categorical variables are not independent.

countframe

## stemcounts nonstemcounts  
## Too Hard 170 150  
## Too Boring 41 31  
## Not Useful 45 74  
## Other 32 36  
## Refused 5 4

chisq.test(countframe)

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

##   
## Pearson's Chi-squared test  
##   
## data: countframe  
## X-squared = 10.046, df = 4, p-value = 0.03966

fisher.test(countframe)

##   
## Fisher's Exact Test for Count Data  
##   
## data: countframe  
## p-value = 0.03726  
## alternative hypothesis: two.sided

## 3c.

Please repeat 2b. omitting respondents who answered “refused”.

When the “refused” option is omitted, the conclusion is still the same, rejection of the presumption of independence. In this case the results are even more convincing, as the p-values are lower.

newframe <- countframe[1:4,]  
newframe

## stemcounts nonstemcounts  
## Too Hard 170 150  
## Too Boring 41 31  
## Not Useful 45 74  
## Other 32 36

chisq.test(newframe)

##   
## Pearson's Chi-squared test  
##   
## data: newframe  
## X-squared = 9.9261, df = 3, p-value = 0.0192

fisher.test(newframe)

##   
## Fisher's Exact Test for Count Data  
##   
## data: newframe  
## p-value = 0.01892  
## alternative hypothesis: two.sided

### Harassment

The variable “HARASS3” records the respondents’ selection based on the following question and options:

Have you ever personally experienced sexual harassment at work, or have you not experienced this?

1 Yes, I have experienced sexual harassment at work

2 No, I have not experienced sexual harassment at work

3 Refused

(Note that the responses are self-reported. Respondents may differ in their perception of harassment and their willingness to report harassment.)

The genders of the respondents are recorded as “Male” or “Female” in the variable “PPGENDER”

(Note that limitation to binary gender reports makes these data ineffective for study of the association of non-binary gender identification with the remaining variables. )

One categorization of the ethnicities of the respondents is recorded in the variable “PPETHM”

### 3d.

Please create a subset of the data that omits responses from the small categories “HARASS3”==“Refused”. The function “droplevels” applied to a factor variable will omit levels that aren’t represented in the data.

Please run and report the result of an appropriate test of the null hypothesis that the response to “HARASS3” is independent from the response to “PPGENDER” on this subset. Please give a statistical interpretation of the results.

For either test, at either the alpha .05 or .01 level, the null hypothesis is rejected. There is overwhelmingly convincing evidence that gender and experiencing harassment at work are not independent. Using the results of the chi-squared test, which rendered a p-value = 0.0009919, if the two categorical variables truly were independent, the probability of getting survey results this disparate, or greater, is less than 1/10 of 1%.

male <- samp$HARASS3[which(samp$PPGENDER=="Male")]  
female <- samp$HARASS3[which(samp$PPGENDER=="Female")]  
malecounts <- c(length(which(male=="Yes, I have experienced sexual harassment at work")),length(which(male=="No, I have not experienced sexual harassment at work")))  
femalecounts <- c(length(which(female=="Yes, I have experienced sexual harassment at work")),length(which(female=="No, I have not experienced sexual harassment at work")))  
harassframe <- data.frame(malecounts,femalecounts)  
rownames(harassframe) <- c("Yes","No")  
harassframe

## malecounts femalecounts  
## Yes 30 59  
## No 265 232

chisq.test(harassframe)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: harassframe  
## X-squared = 10.843, df = 1, p-value = 0.0009919

fisher.test(harassframe)

##   
## Fisher's Exact Test for Count Data  
##   
## data: harassframe  
## p-value = 0.000785  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.2674558 0.7305750  
## sample estimates:  
## odds ratio   
## 0.4457607

### 3e.

Please restrict the data from 3a the respondents who report “Female” in the “PPGENDER” variable. Please provide a table of the proportion of each category of “PPETHM” that responds “Yes, I have experienced sexual harassment at work”. Please run and report the result of an appropriate test of the null hypothesis that the response to “HARASS3” is independent from the response to “PPETHM” on this subset. Please give a statistical interpretation of the results.

With p values of .017 and .019, respectively, we reject the null hypothesis at the .05 level, but not the .01 level. There is some evidence, at the alpha=.05 level, that the variables, ethnicity experiencing workplace harassment (within females) are not independent. If the variables, truly were independent, the probability of getting survey results this disparate or greater, between the different ethnicities is .017, or .019, depending on which test results we are using.

wnf <- samp$HARASS3[which(samp$PPGENDER=="Female" & samp$PPETHM=="White, Non-Hispanic")]  
wnfcounts <- c(length(which(wnf=="Yes, I have experienced sexual harassment at work")), length(which(wnf=="No, I have not experienced sexual harassment at work")))  
  
bnf<- samp$HARASS3[which(samp$PPGENDER=="Female" & samp$PPETHM=="Black, Non-Hispanic")]  
bnfcounts <- c(length(which(bnf=="Yes, I have experienced sexual harassment at work")), length(which(bnf=="No, I have not experienced sexual harassment at work")))  
  
  
onf<- samp$HARASS3[which(samp$PPGENDER=="Female" & samp$PPETHM=="Other, Non-Hispanic")]  
onfcounts <- c(length(which(onf=="Yes, I have experienced sexual harassment at work")), length(which(onf=="No, I have not experienced sexual harassment at work")))  
  
hispf <- samp$HARASS3[which(samp$PPGENDER=="Female" & samp$PPETHM=="Hispanic")]  
hispfcounts <- c(length(which(hispf=="Yes, I have experienced sexual harassment at work")), length(which(hispf=="No, I have not experienced sexual harassment at work")))  
  
  
twoplf <- samp$HARASS3[which(samp$PPGENDER=="Female" & samp$PPETHM=="2+ Races, Non-Hispanic")]  
twoplfcounts <- c(length(which(twoplf=="Yes, I have experienced sexual harassment at work")), length(which(twoplf=="No, I have not experienced sexual harassment at work")))  
  
ethnframe <- data.frame(wnfcounts,bnfcounts,onfcounts,hispfcounts,twoplfcounts)  
rownames(ethnframe) <- c("Yes","No")  
chisq.test(ethnframe)

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

##   
## Pearson's Chi-squared test  
##   
## data: ethnframe  
## X-squared = 12.005, df = 4, p-value = 0.01731

fisher.test(ethnframe)

##   
## Fisher's Exact Test for Count Data  
##   
## data: ethnframe  
## p-value = 0.01869  
## alternative hypothesis: two.sided