Problem Set 8

Adam Ten Hoeve

11 November, 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)

We know that for and to be random variables, they would result from the sum of and , respectively, iid standard normal distributions. In other words, if for then and .

Now we can see that our sum is actually the sum of many iid standard normal distributions, i.e. , which would be distributed by another distribution with 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)

set.seed(0)  
  
sample\_correlation = function(sample\_size){  
 # Generate random samples from the standard normal  
 sample1 = rnorm(sample\_size)  
 sample2 = rnorm(sample\_size)  
 # Return the correlation between the samples  
 return(cor(sample1, sample2))  
}  
  
n = 5000  
sample\_sizes = c(10, 20, 50, 100)  
quantiles = c(0.025, 0.25, 0.5, 0.75, 0.975)  
df = data.frame(matrix(0, nrow=length(sample\_sizes), ncol=length(quantiles)))  
  
index = 1  
# For each sample size, simulate the correlation  
for (sample\_size in sample\_sizes){  
 # Generate n correlations from std.norms of the sample\_size  
 cors = replicate(n, sample\_correlation(sample\_size))  
   
 # Calculate the quantiles for the correlations  
 sample\_quantiles = quantile(cors, probs=quantiles)  
 # Add the quantiles to the dataframe  
 df[index, ] = sample\_quantiles  
 index = index + 1  
}  
  
names(df) = quantiles  
row.names(df) = sample\_sizes  
df

## 0.025 0.25 0.5 0.75 0.975  
## 10 -0.6243023 -0.24761654 -0.0059635221 0.24333546 0.6346495  
## 20 -0.4356216 -0.15526165 0.0054074033 0.16625530 0.4423686  
## 50 -0.2847129 -0.09561399 0.0023159249 0.10139518 0.2787268  
## 100 -0.1936590 -0.06709137 -0.0003046034 0.06930288 0.1968088

### 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.

g = ggplot(df, aes(x=c(1, 2, 3, 4)))  
  
g +   
 # Add errorsbars for [0.025, 0.975] range  
 geom\_errorbar(aes(ymin=df$"0.025", ymax=df$"0.975", colour="blue"), width=0.2) +  
 # Add errorbars for [0.25, 0.75] range  
 geom\_errorbar(aes(ymin=df$"0.25", ymax=df$"0.75", colour="red"), width=0.2) +   
 # Add points for the 0.5 quantile  
 geom\_point(aes(x=c(1, 2, 3, 4), y=df$"0.5")) +  
 # Change the xaxis label and ticks  
 scale\_x\_discrete(limits=c("10", "20", "50", "100")) +  
 # Change the title and axis labels  
 labs(title="Quantiles of Correlations among Std. Normal Samples of Different Sample Sizes",   
 x="Sample Size", y="Correlation Quantile")

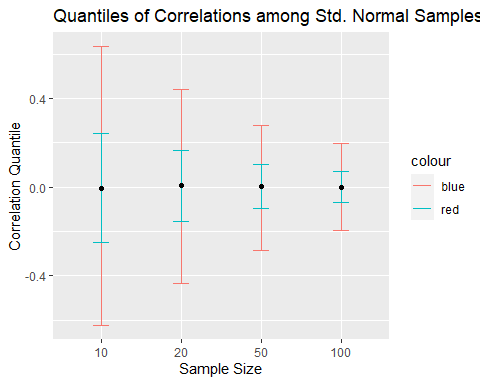
## Warning: Use of `df$"0.025"` is discouraged. Use `0.025` instead.

## Warning: Use of `df$"0.975"` is discouraged. Use `0.975` instead.

## Warning: Use of `df$"0.25"` is discouraged. Use `0.25` instead.

## Warning: Use of `df$"0.75"` is discouraged. Use `0.75` instead.

## Warning: Use of `df$"0.5"` is discouraged. Use `0.5` instead.



## 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-2.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”.

# Calculate the count of each grouping of worktype and response  
counts = samp %>%   
 group\_by(WORKTYPE\_FINAL, SCH7) %>%   
 summarize(n = n())

## `summarise()` regrouping output by 'WORKTYPE\_FINAL' (override with `.groups` argument)

# Format the counts into a 2-way table  
prop.table = matrix(c(counts$n[1:3], sum(counts$n[4:5]), counts$n[6:8], sum(counts$n[9:10])), ncol=2)  
two\_way = data.frame(t(prop.table))  
names(two\_way) = names=c("too hard", "too boring", "not useful", "other")  
row.names(two\_way) = c("stem", "nonstem")  
two\_way

## too hard too boring not useful other  
## stem 170 41 45 37  
## nonstem 150 31 74 40

round(two\_way/rowSums(two\_way), 4)

## too hard too boring not useful other  
## stem 0.5802 0.1399 0.1536 0.1263  
## nonstem 0.5085 0.1051 0.2508 0.1356

## 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.

# Chi-squared test on two\_way data  
chisq.test(two\_way)

##   
## Pearson's Chi-squared test  
##   
## data: two\_way  
## X-squared = 9.8163, df = 3, p-value = 0.02019

# Fisher test on two\_way data  
fisher.test(two\_way)

##   
## Fisher's Exact Test for Count Data  
##   
## data: two\_way  
## p-value = 0.0198  
## alternative hypothesis: two.sided

Both tests have very small p-values, each meaning that we should reject the null hypothesis and assert that there is some statistically significant difference between the STEM and NON-STEM groups.

## 3c.

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

chisq.test(samp$WORKTYPE\_FINAL[samp$SCH7 != "Refused"],   
 samp$SCH7[samp$SCH7 != "Refused"])

##   
## Pearson's Chi-squared test  
##   
## data: samp$WORKTYPE\_FINAL[samp$SCH7 != "Refused"] and samp$SCH7[samp$SCH7 != "Refused"]  
## X-squared = 9.9261, df = 3, p-value = 0.0192

fisher.test(samp$WORKTYPE\_FINAL[samp$SCH7 != "Refused"],   
 samp$SCH7[samp$SCH7 != "Refused"])

##   
## Fisher's Exact Test for Count Data  
##   
## data: samp$WORKTYPE\_FINAL[samp$SCH7 != "Refused"] and samp$SCH7[samp$SCH7 != "Refused"]  
## p-value = 0.01892  
## alternative hypothesis: two.sided

When we remove the “Refused” answers, we get even smaller p-values than when we left them in, which leads us to the same conclusion as in part b. We should reject the null hypothesis and assert that there is some statistically significant difference between the STEM and NON-STEM groups.

### 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.

# Remove any data tuples where HARASS = "Refused"  
samp.harass = samp[samp$HARASS3 != "Refused", ]  
  
# To determine independence, we can use a chi-squared test  
chisq.test(samp.harass$HARASS3, samp.harass$PPGENDER)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: samp.harass$HARASS3 and samp.harass$PPGENDER  
## X-squared = 10.843, df = 1, p-value = 0.0009919

# Let's also check with the fisher test  
fisher.test(samp.harass$HARASS3, samp.harass$PPGENDER)

##   
## Fisher's Exact Test for Count Data  
##   
## data: samp.harass$HARASS3 and samp.harass$PPGENDER  
## p-value = 0.000785  
## alternative hypothesis: two.sided

Both of our tests resulted in a very small p-value, therefore we should reject the null hypothesis, and should assert that the HARASS3 and PPGENDER variables are not independent.

### 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.

# Restrict the data to just females  
samp.female = samp.harass[samp.harass$PPGENDER == "Female", ]  
  
# Construct a 2-way test for the proportions  
counts.female = samp.female %>%  
 group\_by(PPETHM, HARASS3) %>%  
 summarize(n = n())

## `summarise()` regrouping output by 'PPETHM' (override with `.groups` argument)

two.way.female = data.frame(matrix(head(counts.female$n, -1), ncol=2, byrow=TRUE),  
 row.names=c("White", "Black", "Other", "Hispanic"))  
names(two.way.female) = c("Yes", "No")  
# Print out the values and proportions  
two.way.female

## Yes No  
## White 51 147  
## Black 2 28  
## Other 4 29  
## Hispanic 2 25

round(two.way.female / rowSums(two.way.female), 4)

## Yes No  
## White 0.2576 0.7424  
## Black 0.0667 0.9333  
## Other 0.1212 0.8788  
## Hispanic 0.0741 0.9259

# Conduct some tests to see if HARASS3 and PPETHM are independent  
chisq.test(two.way.female)

##   
## Pearson's Chi-squared test  
##   
## data: two.way.female  
## X-squared = 11.148, df = 3, p-value = 0.01095

fisher.test(two.way.female)

##   
## Fisher's Exact Test for Count Data  
##   
## data: two.way.female  
## p-value = 0.01078  
## alternative hypothesis: two.sided

Both tests result in a very small p-value, therefor we should reject the null and assert that the HARASS3 and PPETHM variables are not independent.