Problem Set 7

Adam Ten Hoeve

November 04, 2020

## 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.

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> ).

## Question 1

The precipitation data in “precip.txt” are precipitation values for Boulder, CO from <https://www.esrl.noaa.gov/psd/boulder/Boulder.mm.precip.html>. Precipitation includes rain, snow, and hail. Snow/ice water amounts are either directly measured or a ratio of 1/10 applied for inches of snow to water equivalent. The symbol “Tr” represents a trace amount of precipitation. Observations marked by a "\*" were made at a non-standard site.

In this question, you will read in the data from “precip.txt” and format it. In the formatting steps, the string manipulation commands in the “stringr”" package, part of “tidyverse” may be helpful. The functions “str\_to\_lower”, “str\_detect”, “str\_replace”, and “str\_replace\_all” are particularly relevant. The dplyr function "mutate \_all" may be useful.

This problem is intended as practice in light-duty data formatting. Ordinarily, one would examine the data, decide what formatting needed to be done, carry out the formatting, then use the data in analyses. For educational purposes, the problem directs you through several formatting steps. The data for the remaining problems will be provided separately to enable you to work on those before completing question 1.

The code provided below reads in the precipitation data. The values are tab-separated.Most columns are assigned the string class.

dat<-read.table("precip.csv",stringsAsFactors = FALSE,  
 sep=",",header = TRUE)

## Question 1

### 1a. (1 point)

Replace all column names with all lower case versions. For example, “TOTAL” becomes “total”. Please use a function to do this. Note that the names of a data frame dat can be accessed and modified using the names(dat) syntax. Verify that the reformatting succeeded by outputting the names of the columns using the command “names(dat)”.

names(dat) = str\_to\_lower(names(dat))  
names(dat)

## [1] "year" "jan" "feb" "mar" "apr"   
## [6] "may" "jun" "jul" "aug" "sep"   
## [11] "oct" "nov" "dec" "year.total"

### 1b. (1 point)

Replace all occurrences of “Tr” with 0. Verify that this was successful by running the command “sum(str\_detect(unlist(dat),”Tr“))”.

dat[dat == "Tr"] = 0  
sum(str\_detect(unlist(dat),"Tr"))

## [1] 0

## 1c. (2 points)

Make a boolean vector that indicates which rows have at least one “\*”. Then remove all “\*”s in “dat”. Note that “\*” is a special character in string manipulation and must be proceeded by two back slashes to be used literally, "\\\*“. (This last instruction should be read in the rendered document, because back slashes and stars are also special characters in R markdown.) Please print out the years that include non-standard observations. Also, please verify that no”\*"s remain in the data set by running the command below.

# Find the years that have "\*" in them  
row.has.star = apply(dat, 2, function(x) str\_detect(x, "\\\*") )  
nonstandard.years = dat$year[rowSums(row.has.star) > 0]  
nonstandard.years

## [1] 1893 1989 1990

# Then remove any "\*"  
dat = apply(dat, 2, function(x) str\_remove(x, "\\\*+"))  
sum(str\_detect(unlist(dat),"\\\*"))

## [1] 0

### 1d. (2 points)

Set all precipitation columns to be of “numeric” class using the “as.numeric” function. Make the “year” column to be of class “integer”. Please verify the success of this by running “sapply(dat,class)” and displaying the results.

Also, identify any resulting “NA” values and confirm that the “NA” categorization is correct. The only unavailable values in the dataset are in 2019. Note that “is.na” is a boolean function returning “TRUE” on “NA” values and “FALSE” otherwise.

dat = data.frame(apply(dat, 2, as.numeric))

## Warning in apply(dat, 2, as.numeric): NAs introduced by coercion

dat$year = as.integer(dat$year)  
  
sapply(dat,class)

## year jan feb mar apr may jun   
## "integer" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"   
## jul aug sep oct nov dec year.total   
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"

sum(is.na(dat))

## [1] 4

which(is.na(dat), arr.ind=TRUE)

## row col  
## [1,] 127 11  
## [2,] 127 12  
## [3,] 127 13  
## [4,] 127 14

### 1e. (2 points)

Create another data set dat.trim that has only those years in which all measurements were made at a standard site. Please display the mean total precipitation for the trimmed data set, excluding 2019.

# Create a trimmed data where all data came from the standard site.  
dat.trim = dat[rowSums(row.has.star) == 0, ]  
# Calculate the mean of the total precitpitation, excluding 2019.  
mean(dat.trim[dat.trim$year != 2019 ,]$year.total)

## [1] 19.1535

### 1f. (2 points)

Are all the years between 1893 and 2019 represented in the data? Please check this using code, not by hand, displaying your results.

# Determine if all years between 1893 and 2019 are accounted for.  
all(1893:2019 == dat$year)

## [1] TRUE

## Question 2

### 2a. (5 points)

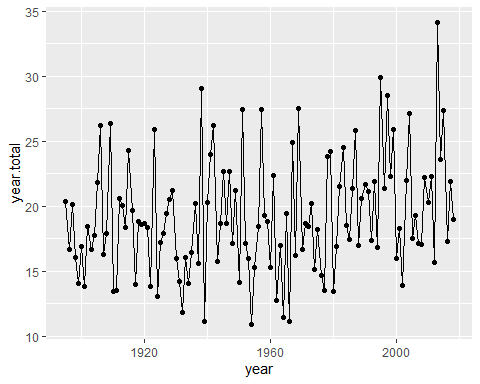
The purpose of this question is to apply a method to examine the extent to which total precipitation is one year is more like the total in the following year than it is like the total in a randomly selected year.The question uses a provided data set “precip\_dat.RData” like (but not exactly like) the one generated in question 1. A preliminary visualization is provided.

Please calculate the mean of the absolute value of the difference between the total precipitation in each odd row with an odd index and the following even-indexed row.

A matrix of 10,000 permutations of the even indices is provided below. Please also calculate the mean of the absolute value of the difference between the total precipitation in odd rows and each simulated even row generated by using the rows given by indices in a row of “mat” in the order in “mat”.

Does comparison of the observed difference in successive years to the simulated differences provide strong evidence against the hypothesis that the precipitation total in successive years no more similar than differences between odd years and any permutation of even years?

load("precip\_dat.RData")  
dat.full<-dat.full[dat.full$year!=2019,]  
ggplot(data=dat.full,aes(x=year,y=year.total))+geom\_line()+geom\_point()

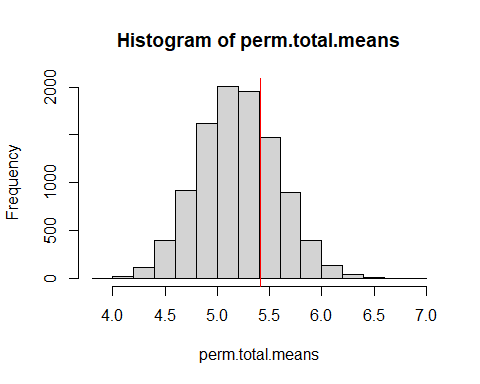


n<-nrow(dat.full)  
indices.even<-seq(2,n,by=2)  
l<-length(indices.even)  
set.seed(345678)  
mat<-matrix(sample(indices.even,10000\*l,replace=TRUE),ncol=l)

# Calculate the difference in total precipitation   
# between each odd index and the following even index  
odd.indexes = seq(1, nrow(dat.full)-1, 2)  
total.diff = dat.full[odd.indexes, ]$year.total - dat.full[odd.indexes+1, ]$year.total  
# Calculate the mean of the absolute values of these differences  
successive.mean = mean(abs(total.diff))  
successive.mean

## [1] 5.411452

# Find the total percepitation of the permuted even indexes  
# and the percipitation of the odd indeces for each row.  
odd.totals = matrix(rep(dat.full[odd.indexes, ]$year.total, 10000), ncol=l, byrow=TRUE)  
even.totals = matrix(dat.full[mat, ]$year.total, ncol=l, byrow=TRUE)  
# Calculate the differences and means between each permuted sample.  
perm.total.diff = abs(odd.totals - even.totals)  
perm.total.means = rowMeans(perm.total.diff)  
# Display the   
hist(perm.total.means)  
abline(v=successive.mean, col="red")

 We want to determine if the mean of the differences from successive years is significantly different than the mean of the differences of any permutation of years. To do that, we can use a Z-test. We know that the means of the perumations will be normally distributed because of the central limit theorem, and we have thousands of trials, so a Z-test is applicable.

Z = (successive.mean - mean(perm.total.means)) / sd(perm.total.means)  
# P-val for a 2 tailed test  
2\*(1 - pnorm(Z))

## [1] 0.5732473

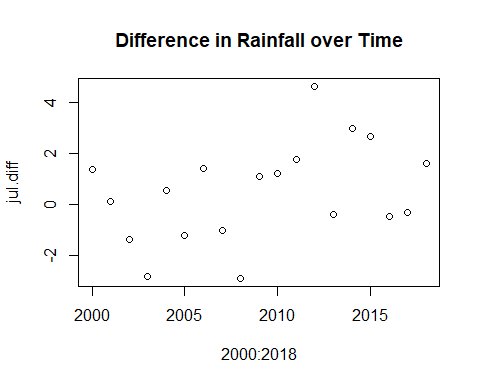
We get a p-value of , which is large. Therefor, we fail to reject the null hypothesis and continue to assert that there is not difference between the successive or randomly permuted differences in percepitation.

### 2b. (5 points)

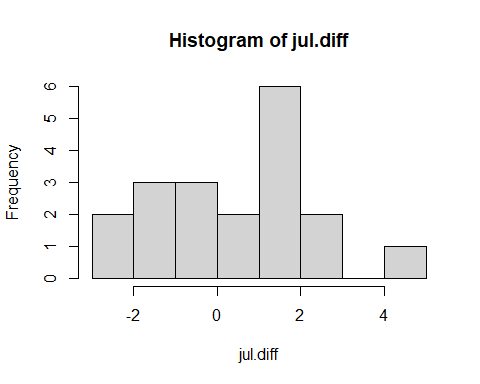
In the code below, a data set, “dat.ok”, similar to the trimmed data from Question 1, is loaded. Please restrict to values between 2000 and 2018. Please apply Student’s t-test and the Wilcoxon signed rank test to the difference between the column dat.ok$jul and the column dat.ok$aug to test the hypothesis that difference of the precipitation in July and August of the same year has mean equal to 0. Please interpret your results, including your thoughts on which test, if either, is appropriate.

Please consider whether the distribution of the annual differences appears to be approximately Normal, whether the distribution of the annual differences appears to be approximately symmetric, and whether the differences in successive years appear to be independent.

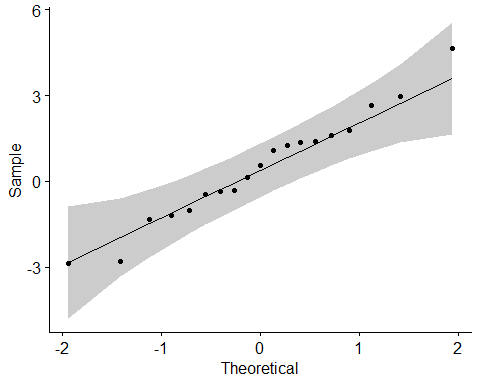
load("precip\_dat\_trim.RData")  
# Restrict to values between 2000 and 2018  
dat.ok = dat.ok[dat.ok$year >= 2000 & dat.ok$year <= 2018, ]  
jul.diff = dat.ok$jul - dat.ok$aug  
  
# Look at the data to see if it meets assumtions.  
plot(x=2000:2018, y=jul.diff, type="p", main="Difference in Rainfall over Time")



hist(jul.diff)



ggqqplot(jul.diff)



t.test(jul.diff)

##   
## One Sample t-test  
##   
## data: jul.diff  
## t = 1.0761, df = 18, p-value = 0.2961  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.4526116 1.4031379  
## sample estimates:  
## mean of x   
## 0.4752632

wilcox.test(jul.diff)

##   
## Wilcoxon signed rank exact test  
##   
## data: jul.diff  
## V = 124, p-value = 0.2579  
## alternative hypothesis: true location is not equal to 0

From the histogram, the distribution doesn’t look that normal. However, the QQ-plot shows that the values do fall within the the theoretical bounds of a normal distribution. Therefor, we can make the assumption that the sample is normally distributed.

To determine if the data is symmetric, we can look at the histogram. Similar to normality, the differences appear somewhat symmetric about the mean, but not extremely. More data would be useful in determining if this were true. However, the plot looks close enough to being symmetric that we can assume the symmetric assumption is maintained.

There is no discernible structure in the plot of differences over time, therefor we can assume that successive differences are independent.

From the conclusions that the data is roughly normal and successive differences are independent, we can use a t-test for this sample. The Wilcoxon test would still probably work, but the t-test will be more rigorous. The t-test results in a p-value of and the Wilcoxon test has a p-value of . As both of these p-values are large, both tests fall to reject the null that the difference between July and August precipitation is .

## Question 3

The data loaded below are sampled from IPUMS, <https://ipums.org/> , an interface for accessing survey and census data. These are drawn from U.S. census microdata in a way that approximates a simple random sample from Colorado households in 2017 that are headed by unmarried men and a simple random sample from Colorado household in 2017 that are headed by unmarried women.

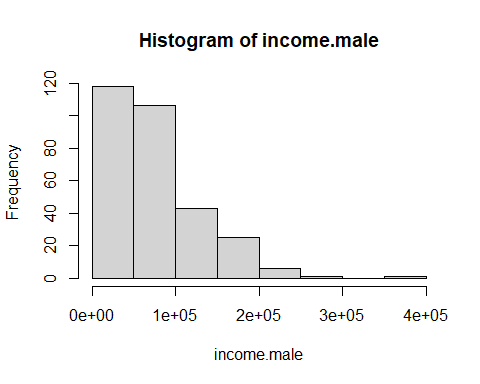
Steven Ruggles, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. Integrated Public Use Microdata Series: Version 6.0 [dataset]. Minneapolis: University of Minnesota, 2015. <http://doi.org/10.18128/D010.V6.0>.

The cases with HHTYPE equal to 2 make up the sample of male-headed households. The cases with HHTYPE equal to 3 make up the sample of female-headed households.

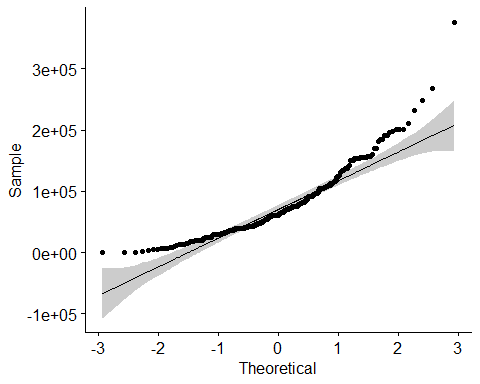
### 3a. (10 points)

Are the household incomes for the male-headed households approximately Normally distributed? Are the household incomes for the female-headed households approximately Normally distributed? Please provide visualizations to support your response.

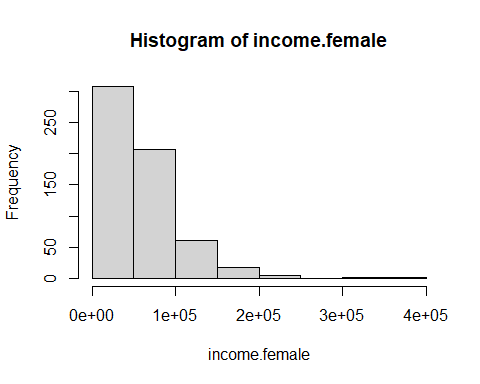
load("dat\_mf.RData")  
  
# Is the male incomes normally distributed?  
income.male = dat.mf[dat.mf$HHTYPE == 2, ]$HHINCOME  
hist(income.male)



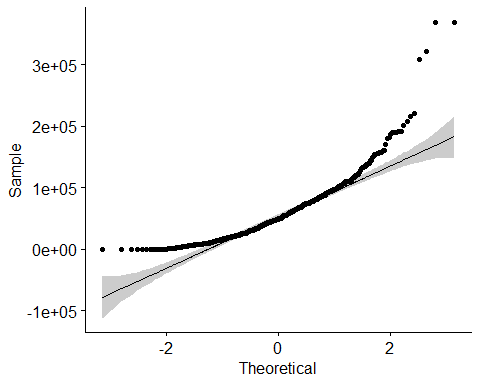
ggqqplot(income.male)



# Is the famel incomes normally distributed?  
income.female = dat.mf[dat.mf$HHTYPE == 3, ]$HHINCOME  
hist(income.female)



ggqqplot(income.female)



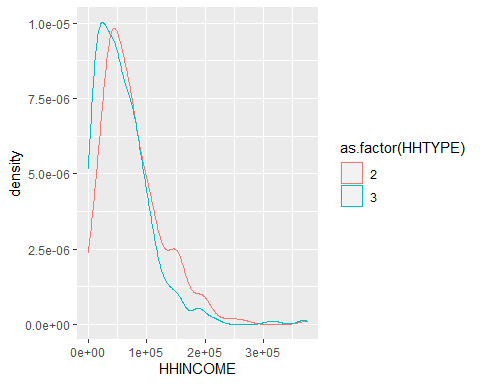
Niether the male or female incomes appear to be normally distributed. Both histograms of the incomes follow an exponential curve and the QQ plots show that the values fall way out of the theoretical bounds.

### 3b. (10 points)

Please carry out a Mann-Whitney U-test on the two data sets, the household incomes for the male-headed households and the household incomes for the female-headed households.

What can you conclude from the results? In particular, can this test be interpreted as a test of center in this case?

ggplot(dat.mf, aes(x=HHINCOME, color=as.factor(HHTYPE))) +  
 geom\_density()



wilcox.test(HHINCOME ~ HHTYPE, data=dat.mf)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: HHINCOME by HHTYPE  
## W = 107724, p-value = 1.427e-06  
## alternative hypothesis: true location shift is not equal to 0

From our Mann-Whitney U-test, we get a p-value of . This is a very small value and, assuming a 95% confidence level, means we should reject the null hypothesis. Therefor, we can assert that the mean incomes are different between men and women.

### 3c. (0 points)

The code below carries out a bootstrap test of the difference in means of the household incomes for the male-headed households and the household incomes for the female-headed households. Please study this and be prepared to ask questions about it in live session.

Basic bootstrap samples are samples with replacement of cases from the data. They are used to estimate confidence intervals on statistics non-parametrically. If the empirical distribution is close to the population distribution, then a bootstrap sample from the empirical distribution simulates a new sample. Computing the range of the statistic of interest for a large number of bootstrap samples gives an indication of the range of values that would be produced if the population actually was resampled.

# wrapper function for the difference in means in a format compatible with the "boot" sampling function  
  
# The return value is a vector with the number of households of each type followed by the difference in means. The counts of the households of each type is to check the stratified sampling.  
  
boot.mean.diff<-function(dat,indices){  
 dat.this<-dat[indices,]  
 gp2<-dat.this$HHINCOME[dat.this$HHTYPE==2]  
 gp3<-dat.this$HHINCOME[dat.this$HHTYPE==3]  
 return(c(length(gp2),length(gp3),mean(gp2)-mean(gp3)))  
}  
  
# Draw 5000 bootstrap samples stratified by household type.   
  
samp<-boot(dat.mf,boot.mean.diff,5000,strata=dat.mf$HHTYPE)  
  
# The sampling results are in the data member samp$t.  
  
# Check that the number of HHTYPE==2 and HHTYPE==3 households in each bootstrap sample equals the original number.  
  
unique(samp$t[,1])

## [1] 300

unique(samp$t[,2])

## [1] 600

# Look at quantiles of the mean difference  
  
quantile(samp$t[,3],c(.025,.975))

## 2.5% 97.5%   
## 9485.261 23328.986

# Another interval estimate  
  
boot.ci(samp,type="bca",index=3)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 5000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = samp, type = "bca", index = 3)  
##   
## Intervals :   
## Level BCa   
## 95% ( 9232, 23193 )   
## Calculations and Intervals on Original Scale

# Note that the lower bound is positive.