Problem Set 7

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.

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!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezvxll0GP6$ >, <<https://urldefense.com/v3/__http://rmarkdown.rstudio.com__>;!!NCZxaNi9jForCP\_SxBKJCA!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezv9h7TJbG$ > ).

## Question 1

The precipitation data in “precip.txt” are precipitation values for Boulder, CO from <https://urldefense.com/v3/__https://www.esrl.noaa.gov/psd/boulder/Boulder.mm.precip.html__>;!!NCZxaNi9jForCP\_SxBKJCA!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezvwtfgeAo$ . 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 purpose of this analysis is to assess the null hypothesis that the total annual rainfalls in the early portion and the total annual rainfalls in the recent portion of the data are each independent identically distributed (i.i.d.) samples from Normally distributed populations with equal means, and .

Unlike in a class setting, in practice, data formatting is often a major component of a data analysis project. Some basic formatting of the data in “precip.txt” is included below for reference.

The symbol “Tr” represents a trace amount of precipitation. Observations marked by a “\*” were made at a non-standard site. Some light-duty data formatting appears below that sets “Tr” values to and drops years that include an observation made at a non-standard site.

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.txt",sep="\t",header = TRUE)

The following replaces all column names with lower case versions. For example, “TOTAL” becomes “total”. The command “names(dat)” is used to verify that the replacement has succeeded.

# Change all characters in the variable names to lower case.  
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"

Drop observations for the year 2021 which has multiple missing values.

dat<-filter(dat,year!="2021")

Replace all occurrences of “Tr” with 0. Verify that this was successful.

# Replace "Tr".  
dat<-mutate\_all(dat,str\_replace,"Tr","0")  
# Count all occurrences of "Tr".  
sum(str\_detect(unlist(dat),"Tr"))

## [1] 0

Drop all rows that include an asterisk indicating an observation at a non-standard location. The method for this is to write a function that takes a vector of strings as its argument and returns “TRUE” inf none of the strings contains an asterisk, “FALSE” otherwise. Then apply this function to each row of the data to generate a Boolean vector. Finally, reduce the data set to only those rows without asterisks using this vector.

Note that the asterisk has a special meaning in string manipulation so the backslashes are used to look for a literal asterisk.

<https://urldefense.com/v3/__https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html__>;!!NCZxaNi9jForCP\_SxBKJCA!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezv1pAhB-m$

# function to return TRUE if a string vector x contains no entries with an "\*".  
no\_stars<-function(x){  
 sum(str\_detect(x,"\\\*"))==0  
}  
  
# Count asterisks in the data.  
sum(str\_detect(unlist(dat),"\\\*"))

## [1] 8

# Identify the rows in the data with at least 1 "\*".  
all.standard<-apply(dat,1,no\_stars)  
dat.trim<-dat[all.standard,]  
# Count asterisks in the trimmed data.  
sum(str\_detect(unlist(dat.trim),"\\\*"))

## [1] 0

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

Verify that converting the strings to numeric values didn’t produce any “NA”s.

dat.trim<-mutate\_all(dat.trim,as.numeric)  
dat.trim[,1]<-as.integer(dat.trim[,1])  
  
sapply(dat.trim,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] 0

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

## row col

Identify the omitted years in “dat.trim”.

setdiff(min(dat.trim$year):max(dat.trim$year),dat.trim$year)

## [1] 1989 1990

### 1.a.

(10 points)

Since values in successive years may be related by persistent weather patterns, the data are thinned to every third entry in “dat.s”

Please provide a visual assessment of the consistency with Normality of the first 15 values for “year.total” in “dat.s” and of the consistency with Normality of the last 15 values for “year.total” in “dat.s”. Please give a verbal assessment based on the visualization.

dat.s<-filter(dat.trim,year%%3==2)  
  
nrow(dat.s)

## [1] 42

nrow(dat.s)-14

## [1] 28

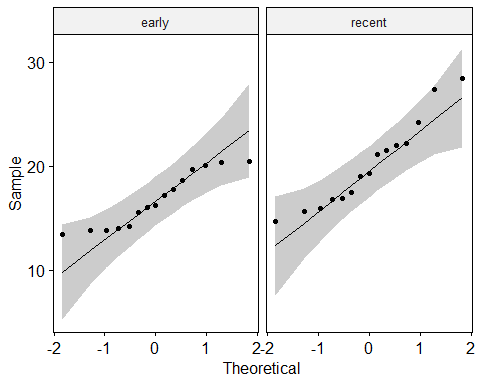
(nrow(dat.s)-14):nrow(dat.s)

## [1] 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

c(1:15,(nrow(dat.s)-14):nrow(dat.s))

## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 28 29 30 31 32 33 34 35 36 37  
## [26] 38 39 40 41 42

dat.sep<-dat.s[c(1:15,(nrow(dat.s)-14):nrow(dat.s)),]  
dat.sep$era<-rep(c("early","recent"),  
 times=c(15,15))  
ggqqplot(dat.sep$year.total,facet.by=c("dat.sep$era"))

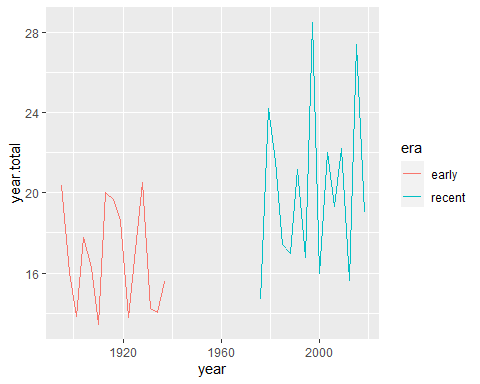


## 1.b.

(5 points)

Please provide a visualization to examine whether the “year.total” values show smooth variation over time or the “year.total” values at consecutive time points in “dat.s” appear to be independent. Please state your assessment.

ggplot(data=dat.sep,aes(x=year,y=year.total,color=era))+geom\_line()



## 1.c.

(10 points)

Please perform a test of the null hypothesis that the total annual rainfalls in the early portion and the total annual rainfalls in the recent portion are each i.i.d. samples from Normally distributed populations with equal means, and . Please state your conclusion based on 1.a. and 1.b. regarding the null hypothesis that the means in the two populations are equal.

t.test(year.total~era,data=dat.sep)

##   
## Welch Two Sample t-test  
##   
## data: year.total by era  
## t = -2.6836, df = 23.356, p-value = 0.01316  
## alternative hypothesis: true difference in means between group early and group recent is not equal to 0  
## 95 percent confidence interval:  
## -6.0693853 -0.7879481  
## sample estimates:  
## mean in group early mean in group recent   
## 16.75600 20.18467

### MDH Comment: We can conclude that both populations are normally distributed. There is a strong evidence against the null hypothesis.

## Question 2

The goal in this analysis is to perform the strongest suitable test of whether the precipitation amount differs between October and November.

### 2.a.

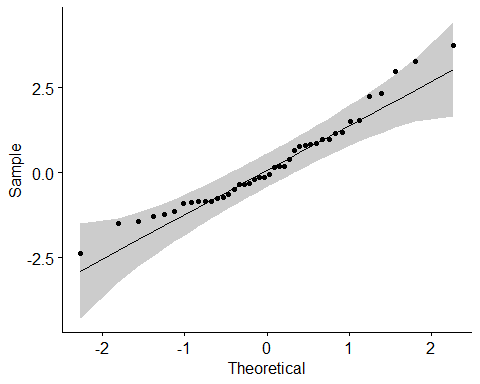
(10 points)

Please generate visualizations to address whether the differences between the precipitation in October the following November in “dat.trim” are consistent with being i.i.d. samples from a distribution. Please address independence and Normality.

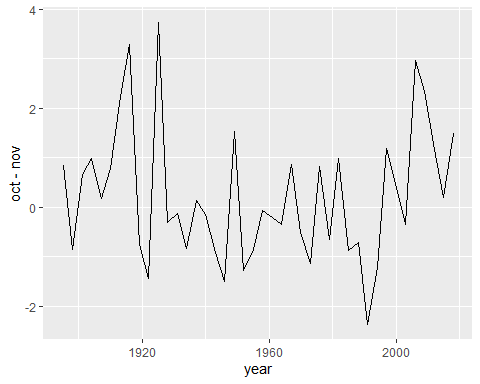
diff<-dat.s$oct-dat.s$nov  
diff

## [1] 0.83 -0.86 0.65 0.98 0.17 0.78 2.24 3.28 -0.76 -1.45 3.73 -0.31  
## [13] -0.14 -0.84 0.14 -0.15 -0.92 -1.51 1.52 -1.29 -0.86 -0.07 -0.21 -0.36  
## [25] 0.85 -0.51 -1.15 0.81 -0.65 0.98 -0.88 -0.72 -2.37 -1.23 1.18 0.39  
## [37] -0.35 2.97 2.33 1.16 0.19 1.49

ggqqplot(diff)



ggplot(dat.s,aes(x=year,y=oct-nov))+geom\_line()



shapiro.test(diff)

##   
## Shapiro-Wilk normality test  
##   
## data: diff  
## W = 0.95258, p-value = 0.07993

### 2.b.

(5 points)

Please perform the strongest test of the null hypothesis that the difference in precipitation between October and November has mean equal to .

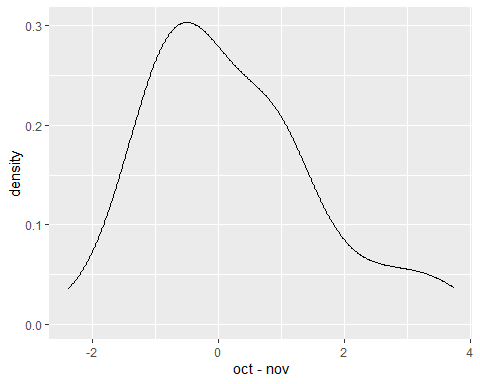
x<-dat.s$oct  
y<-dat.s$nov  
t.test(x,y, paired = TRUE, conf.level = .99)

##   
## Paired t-test  
##   
## data: x and y  
## t = 1.0358, df = 41, p-value = 0.3064  
## alternative hypothesis: true difference in means is not equal to 0  
## 99 percent confidence interval:  
## -0.3476201 0.7800010  
## sample estimates:  
## mean of the differences   
## 0.2161905

#we could also just do this:  
t.test(diff, )

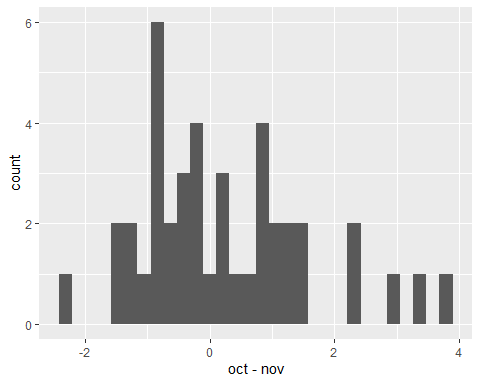
##   
## One Sample t-test  
##   
## data: diff  
## t = 1.0358, df = 41, p-value = 0.3064  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.2053431 0.6377240  
## sample estimates:  
## mean of x   
## 0.2161905

# weaker, requires symmetry to be a test of location of the mean.  
ggplot(dat.s,aes(x=oct-nov))+geom\_density()



ggplot(dat.s,aes(x=oct-nov))+geom\_histogram()

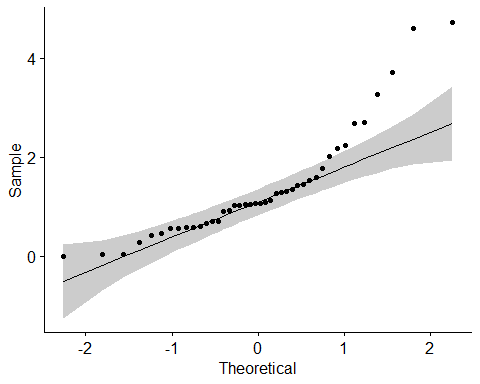
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



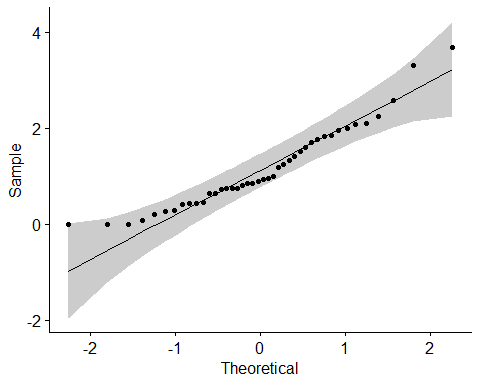
wilcox.test(diff)

##   
## Wilcoxon signed rank exact test  
##   
## data: diff  
## V = 493, p-value = 0.6117  
## alternative hypothesis: true location is not equal to 0

# weaker, with departure from Normality in October data  
ggqqplot(dat.s$oct)



ggqqplot(dat.s$nov)



t.test(dat.s$oct,dat.s$nov)

##   
## Welch Two Sample t-test  
##   
## data: dat.s$oct and dat.s$nov  
## t = 0.9961, df = 77.428, p-value = 0.3223  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.2159490 0.6483299  
## sample estimates:  
## mean of x mean of y   
## 1.369524 1.153333

## Question 3

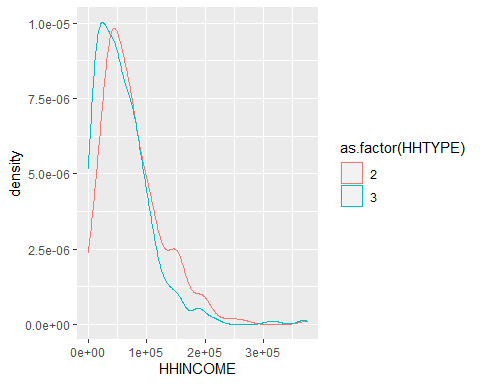
The data loaded below are sampled from IPUMS, <https://urldefense.com/v3/__https://ipums.org/__>;!!NCZxaNi9jForCP\_SxBKJCA!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezvwYW4pXi$ , an interface for accessing survey and census data. These are drawn from U.S. Census microdata in a way that 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. <https://urldefense.com/v3/__http://doi.org/10.18128/D010.V6.0__>;!!NCZxaNi9jForCP\_SxBKJCA!GQQUp0j2mpSjpcaumRUWpSeycTzmO-hoNw-xTCO6Q4vqAZcLBiNzp8qriFezv8T2lFa2$ .

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.

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

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

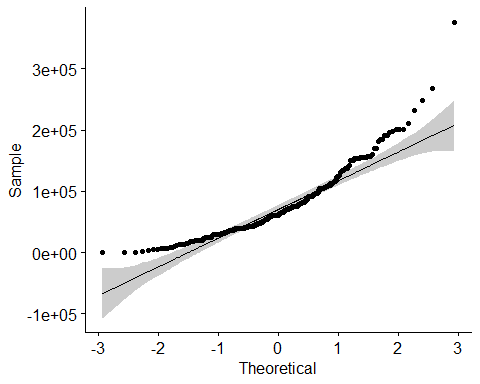


### 3.a.

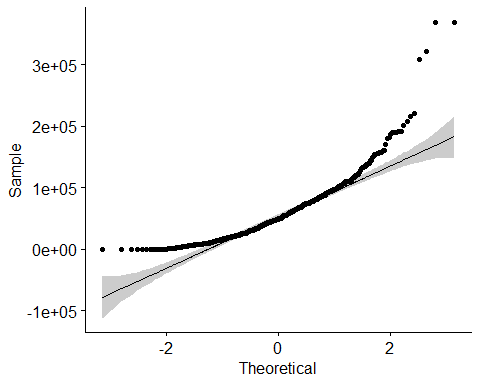
(5 points)

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.

ggqqplot(dat.mf$HHINCOME[dat.mf$HHTYPE==2])



ggqqplot(dat.mf$HHINCOME[dat.mf$HHTYPE==3])



### 3.b.

(5 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?

wilcox.test(dat.mf$HHINCOME[dat.mf$HHTYPE==2],  
 dat.mf$HHINCOME[dat.mf$HHTYPE==3])

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: dat.mf$HHINCOME[dat.mf$HHTYPE == 2] and dat.mf$HHINCOME[dat.mf$HHTYPE == 3]  
## W = 107724, p-value = 1.427e-06  
## alternative hypothesis: true location shift is not equal to 0

### MDH Comment: We can conclude that there is a significant difference between the two groups. The null hypothesis is rejected.

### 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 class.

Basic bootstrap samples are samples with replacement of cases from the data. They are used to estimate confidence intervals on statistics non-parametrically.

A data vector defines an *empirical* probability distribution as follows. The sample space is the set of distinct values in . The set of events is the power set of the sample space. The probability function is defined by the density function where is the number of occurrences of the value in and is the length of .

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%   
## 9199.341 23417.145

# 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% ( 9194, 23425 )   
## Calculations and Intervals on Original Scale

# Note that the lower bound is positive.