Ward\_abigail\_HW7

Abigail Ward

8/3/2021

## Question 1

dat<-read.table("/Users/abbyward/Downloads/precip.txt",sep="\t",header = TRUE)  
  
# 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"

dat<-filter(dat,year!="2021")  
  
# Replace "Tr".  
dat<-mutate\_all(dat,str\_replace,"Tr","0")  
# Count all occurrences of "Tr".  
sum(str\_detect(unlist(dat),"Tr"))

## [1] 0

# 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 value type  
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

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

## [1] 1989 1990

### 1.a.

Both the first 15 and last 15 data points appear normal respectively. All 15 data points fit within the permitted normal region.

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

## [1] 42

nrow(dat.s)-14

## [1] 28

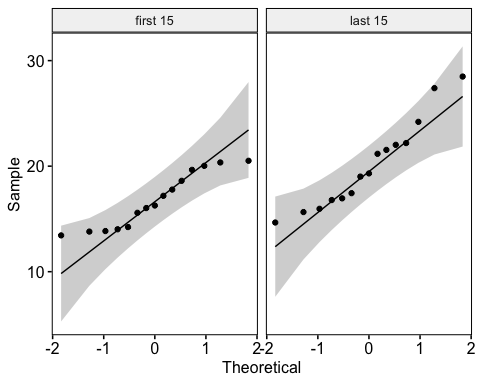
(nrow(dat.s)-14):nrow(dat.s)

## [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))

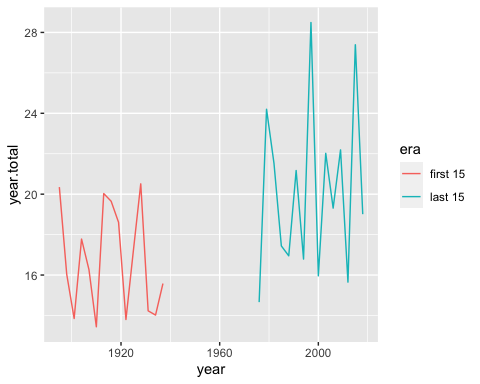
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 28 29 30 31 32 33 34 35 36 37  
## [26] 38 39 40 41 42

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

 ## 1.b.

There is no smooth variation in the data over time. There is no consistent upward or downward trend in the first or last data and therefore the year total values appear to be independent.

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



## 1.c.

Since the P-value is 1.3%, there is strong evidence against 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 . This is also re-enforced by parts a and b showing the data is both normal and independent.

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 first 15 and group last 15 is not equal to 0  
## 95 percent confidence interval:  
## -6.0693853 -0.7879481  
## sample estimates:  
## mean in group first 15 mean in group last 15   
## 16.75600 20.18467

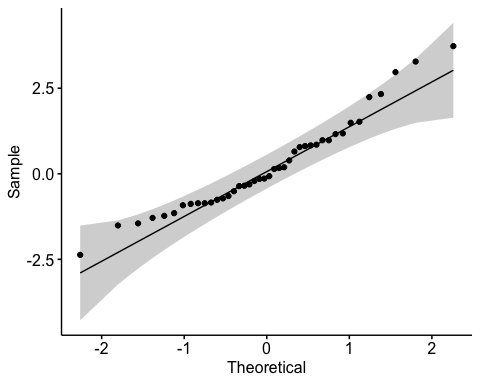
## Question 2

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

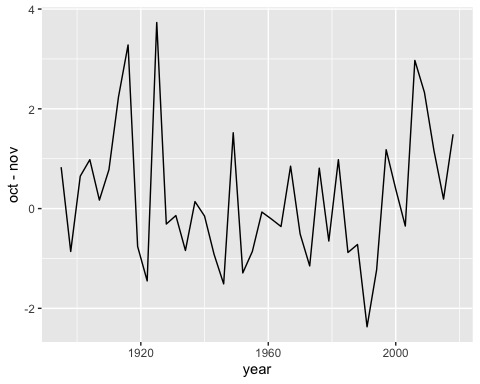
### 2.a.

The data points all fall within or very close to the permitted normal region in the first plot and there is no trends in the data in the second plot which indicate that the data is both normal and independent. With a p-value of abouy 8%, we fail to reject the null hypothesis.

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

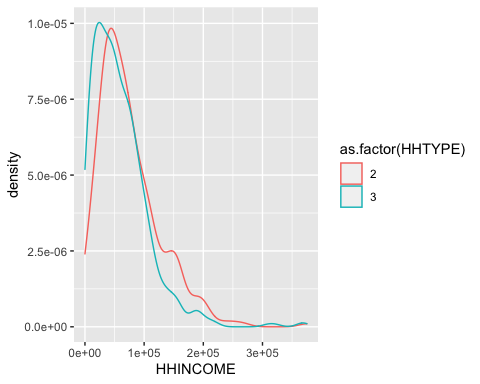
With a p value of 0.3064, we fail to reject the null hypothesis that the true difference int the means is equal to 0.

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

## Question 3

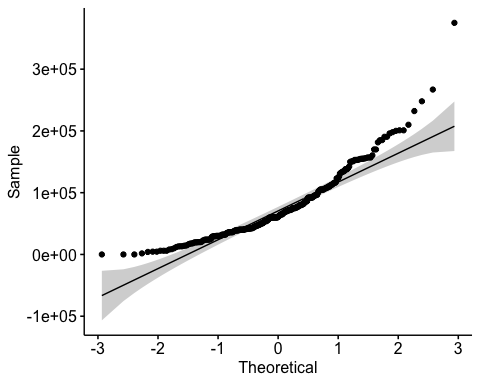
load("/Users/abbyward/Downloads/dat\_mf.RData")  
ggplot(dat.mf, aes(x=HHINCOME, color=as.factor(HHTYPE))) +  
 geom\_density()



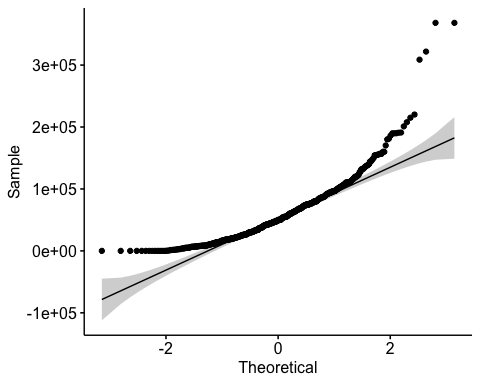
### 3.a.

The household incomes for the male-headed households are not approximately Normally distributed and neither are the household incomes for the female-headed households. The data points for both sets fall outside the permitted normal range.

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



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



### 3.b.

Since the p-value is 0.000001427 we reject the null hypothesis that there is no difference in the household incomes for the male-headed households and the household incomes for the female-headed households. This is a test of center as it is comparing the medians of the two data sets.

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

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

## [1] 60750

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

## [1] 49150

## 3.c

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)  
  
  
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%   
## 9181.245 23373.715

# 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% ( 9307, 23541 )   
## Calculations and Intervals on Original Scale