Problem Set 7

A. Dehesa

## 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> downloaded 2/17/2022.

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, “chr”.

dat<-read.table("precip\_2021.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"

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” if 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,using this vector, reduce the data set to only those rows without asterisks.

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

<https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html>

# 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

## Question 1, part 1

(10 points)

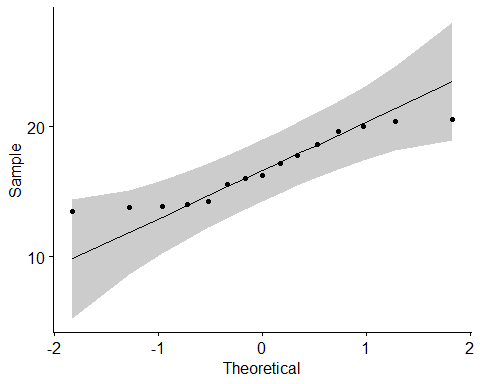
Since values in successive years may be related by persistent weather patterns, the data are thinned to every third entry in “dat.s”

For Welch’s test to be a valid test of the null hypothesis of equality of population means, both populations should be (approximately) Normally distributed.

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. Within each period, are these data consistent with Normality?

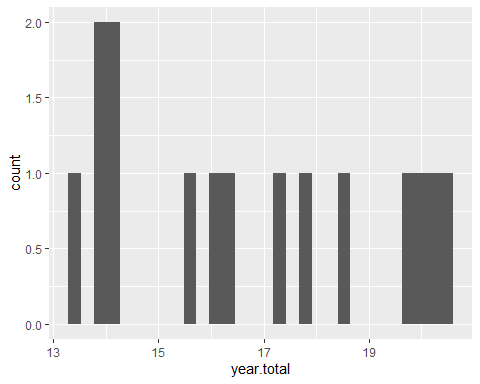
dat.s<-filter(dat.trim,year%%3==2)  
  
dat.sep<-dat.s[c(1:15,(nrow(dat.s)-14):nrow(dat.s)),]  
dat.sep$era<-rep(c("early","recent"),  
 times=c(15,15))  
dat.early <- filter(dat.sep, era == "early")  
dat.recent <- filter(dat.sep, era == "recent")  
# your plotting code here  
ggqqplot(dat.sep$year.total[dat.sep$era == "early"])

## Warning: The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?  
## The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?



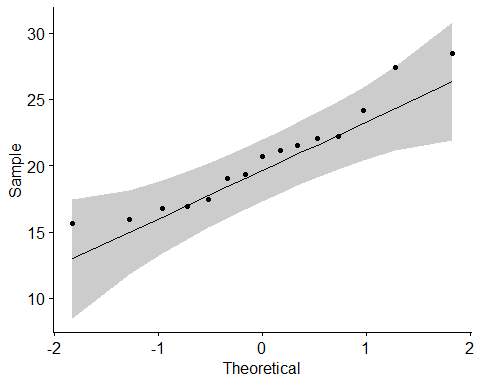
ggplot(data = dat.early, aes(x=year.total)) + geom\_histogram()

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



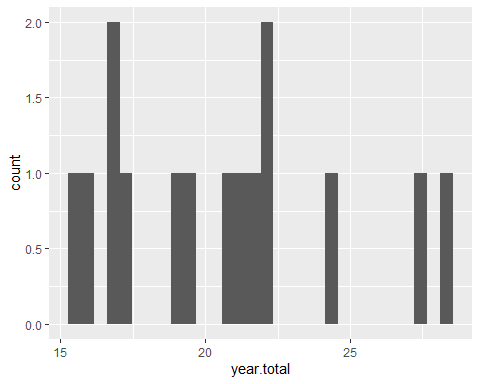
ggqqplot(dat.sep$year.total[dat.sep$era == "recent"])

## Warning: The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?  
## The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?



ggplot(data = dat.recent, aes(x=year.total)) + geom\_histogram()

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

 They dont seem to be very consistent. The histogram does not look to be normally distributed.

## Question 1, part 2

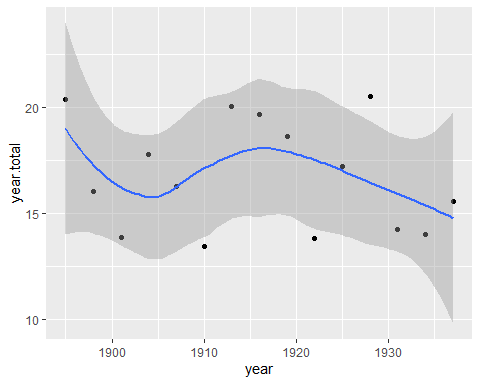
(10 points)

For Welch’s test to be a valid test of the null hypothesis of equality of population means, the values in each group should be independent of one another.

Please provide a visualization to examine whether the “year.total” values show smooth variation over time, an indication of dependence, or whether the “year.total” values at consecutive time points in “dat.s” within the early (first 15 in dat.s, 1895-1937) period appear to be independent of one another and the “year.total” values at consecutive time points in “dat.s” within the recent(last 15 in dat.s, 1979-2021) periods appear to be independent. Please state your assessment.

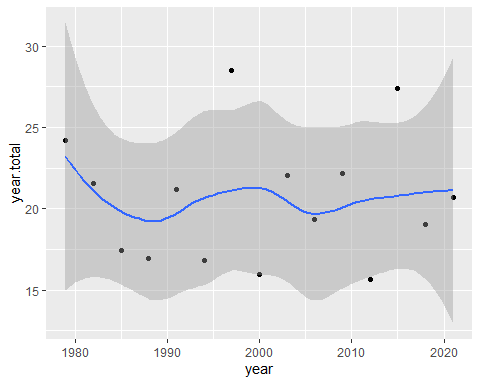
# Visualization for the early era  
ggplot(data = dat.early, aes(x=year, y= year.total)) +   
 geom\_point() +  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



# Visualization for the recent era  
ggplot(data = dat.recent, aes(x=year, y= year.total)) +   
 geom\_point() +  
 geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

 They both look decently smooth, which means we should be able to use the Welch´s test.

## Question 1, part 3

(10 points)

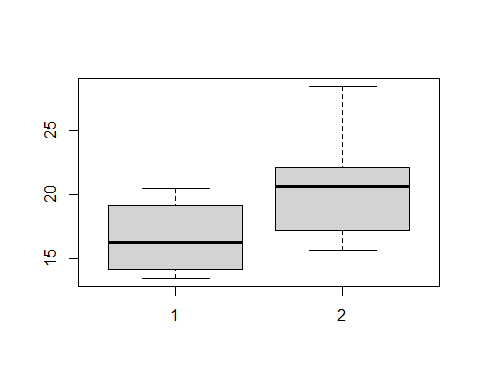
Please perform Welch’s test of the null hypothesis that the total annual rainfalls in the early portion (first 15 values of dat.s) and the total annual rainfalls in the recent portion (last 15 values of dat.s) are each i.i.d. samples from Normally distributed populations with equal means, and . Please state your conclusion based on 1.a. and 1.b. regarding the null hypothesis that the means in the two populations are equal.

# We will have mean = 0  
# We can simply use the integrated t.test in R  
t.test(dat.early$year.total, dat.recent$year.total)

##   
## Welch Two Sample t-test  
##   
## data: dat.early$year.total and dat.recent$year.total  
## t = -3.1518, df = 24.345, p-value = 0.004267  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.336094 -1.323906  
## sample estimates:  
## mean of x mean of y   
## 16.756 20.586

And as we can see, the p-value is way below 0.5, so we can discard the null hypothesis. We can actually see this using a boxplot

boxplot(dat.early$year.total, dat.recent$year.total)

 We can see here that the mean for the early data is way below the mean for the recent data.

# Question 2

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

## Question 2, part 1

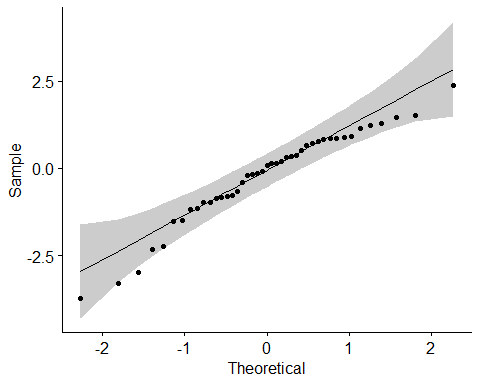
(10 points)

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

## Answer Q2P1

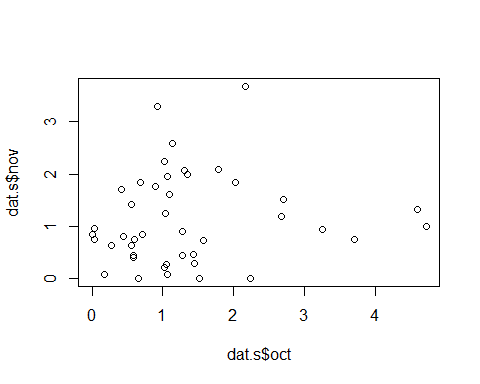
diff<-dat.s$nov-dat.s$oct  
ggqqplot(diff)

## Warning: The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?  
## The following aesthetics were dropped during statistical transformation: sample  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?

 With this plot, we can see the normality of the samples, which seem to be within limits.

To check for the independence of both samples, we could simply plot them, setting one to x, and the other to y. If both maintain a linear graph, they could be dependent (one dependent on the other).

plot(x=dat.s$oct, y=dat.s$nov)



We could also perform the chi square test.

(res<-chisq.test(x = dat.s$nov, y=dat.s$oct))

## Warning in chisq.test(x = dat.s$nov, y = dat.s$oct): Chi-squared approximation  
## may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: dat.s$nov and dat.s$oct  
## X-squared = 1515.7, df = 1482, p-value = 0.2652

The slide for week 8 mentioned that the rule of thumb for the chi square test is np(1-p) > 3:

p <- res$p.value  
n <- length(dat.s$nov)  
n\*p\*(1-p)

## [1] 8.378694

As we can see, our result is over 3, so we accept that they are independent.

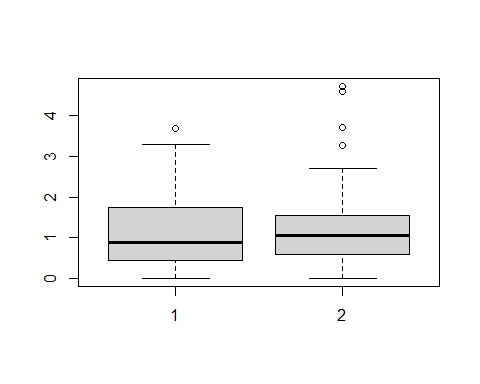
## Question 2, part 2

(10 points)

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

The best test for this case would be the t-test.

# By using the boxplot, we can see that the variance is similar, but not equal, so we set the variance to FALSE for the t-test.  
boxplot(dat.s$nov, dat.s$oct)



t.test(dat.s$nov, dat.s$oct, var.equal = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: dat.s$nov and dat.s$oct  
## t = -0.99187, df = 79.472, p-value = 0.3243  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.6418692 0.2148925  
## sample estimates:  
## mean of x mean of y   
## 1.128140 1.341628

The test itself tells us that the difference in means is not equal to 0. Therefore, the means are not equal.