Problem Set 5

C. Durso, D. Parada

March 07, 2019

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 "boulder_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 "boulder_precip.txt" and format it. The data are tab-separated. 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, cary 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 porblems will be provided separately to enable you to work on those before completing question 1.

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

```
dat<-read.table("boulder_precip.txt",stringsAsFactors = FALSE,sep="\t",header=TRUE)
```

Question 1, part 1

(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)".

```
# 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"   "may"   "jun"   "jul"
##  [9] "aug"   "sep"   "oct"   "nov"   "dec"   "total"
```

Question 1, part 2

(1 point)

Replace all occurrences of "Tr" with 0. Verify that this was successful by running the command "sum(str_detect(unlist(dat),"Tr"))".

```
# Replace "Tr" with "0"
dat<-mutate_all(dat,str_replace,"Tr","0")
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
sum(str_detect(unlist(dat),"Tr"))
```

```
## [1] 0
```

Question 1, part 3

(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.

```
sum(str_detect(unlist(dat),"\\*"))
```

```
## [1] 8
```

```
# function to return TRUE if a string vector x contains any entries with an "*".
any_stars<-function(x){
  sum(str_detect(x,"\\*"))>0
}

# Identify the rows in the data with at least 1 "*".
iffy<-apply(dat,1,any_stars)

dat<-mutate_all(dat,str_replace_all,"\\*","")
dat$year[iffy]
```

```
## [1] "1893" "1989" "1990"
```

Question 1, part 4

(2 points)

Set all precipitation columns to be of "numeric" class. Make the "year" column to be of class "integer". Please verify the success of this by running "sapply(dat,class)". Also, run "sum(is.na(dat))" to verify that all entries could be converted to numeric values.

```
dat<-mutate_all(dat,as.numeric)
dat[,1]<-as.integer(dat[,1])
sapply(dat,class)
```

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

```
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
```

```
sum(is.na(dat))
```

```
## [1] 0
```

Question 1, part 5

(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.

```
dat.trim<-dat[!iffy,]
mean(dat.trim$total)
```

```
## [1] 19.1535
```

Question 1, Part 6

(2 points)

Are all the years between 1893 and 2018 represented in the data? Please check this using code, not by hand.

```
unique(dat$year-1893:2018)
```

```
## [1] 0
```

Because these differences are all 0, the years in the data are 1893-2018 in that order.

Question 2

Question 2, part 1

(5 points)

The point of this question is to study 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 "dat_full.csv" like (but not exactly like) the one generated in question 1. A preliminary visualization is provided.

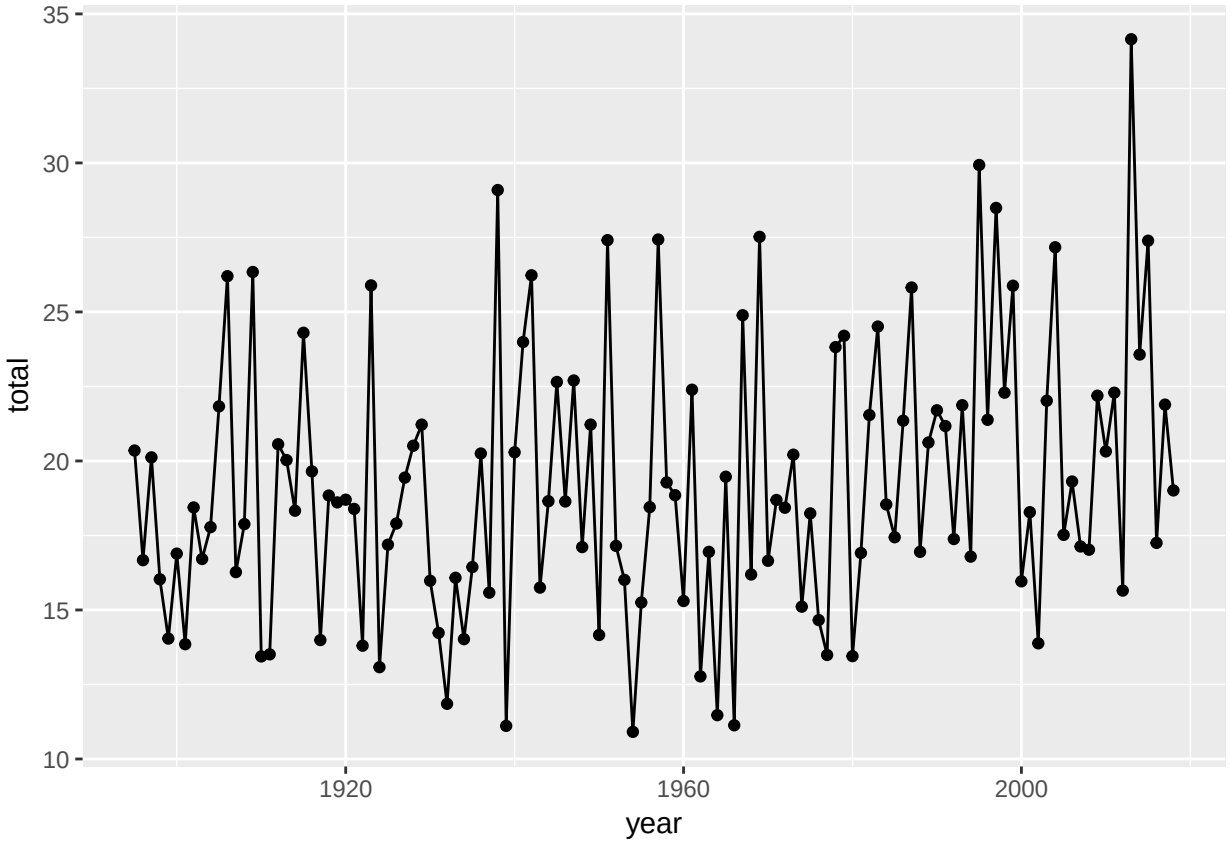
Please calculate the mean of the absolute value of the difference between the total precipitation in each odd row and the following even row.

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 totals in successive years are unrelated?

No. When taking the mean of the absolute value of the difference between the odd and random even years which are less than the consecutive even and odd years, we obtain a mean that is around 0.7194 which means that by randomly selecting years we get a mean that is closer to 0.

```
dat<-read.csv("dat_full.csv")
ggplot(data=dat,aes(x=year,y=total))+geom_line()+geom_point()
```




```
n<-nrow(dat)
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)

# Odd indices
odd<-dat$total[seq(1,n,by=2)]
# Even indices
even<-dat$total[seq(2,n,by=2)]

# Calculating the mean absolute value of the difference between odd years and the matrix of even years
means<-apply(mat,1,function(x){mean(abs(odd-dat$total[x]))})

# Comparison of even and odd years
mean.obs<-mean(abs(odd-even))


mean(means<=mean.obs)
```

```
## [1] 0.7194
```

Question 2, Part 2

(5 points)

Please repeat the analysis from part 1 using dat$jul-dat$aug in place of dat$total, then dat$jan-dat$jul in place of dat$total

```
dat$jul.aug<-dat$jul-dat$aug

odd<-dat$jul.aug[seq(1,n,by=2)]
means<-apply(mat,1,function(x){mean(abs(odd-dat$jul.aug[x]))})

even<-dat$jul.aug[seq(2,n,by=2)]
mean.obs<-mean(abs(odd-even))


mean(means<=mean.obs)
```

```
## [1] 0.7628
```

```
dat$jan.jul<-dat$jan-dat$jul

odd<-dat$jan.jul[seq(1,n,by=2)]
means<-apply(mat,1,function(x){mean(abs(odd-dat$jan.jul[x]))})

even<-dat$jan.jul[seq(2,n,by=2)]
mean.obs<-mean(abs(odd-even))


mean(means<=mean.obs)
```

```
## [1] 0.7256
```

The conclusions are the same. The mean absolute difference for the consecutive years is not unusually small compared to the mean absolute difference for random years. This is not strong evidence against the hypothesis that the totals in successive years are unrelated

Question 3

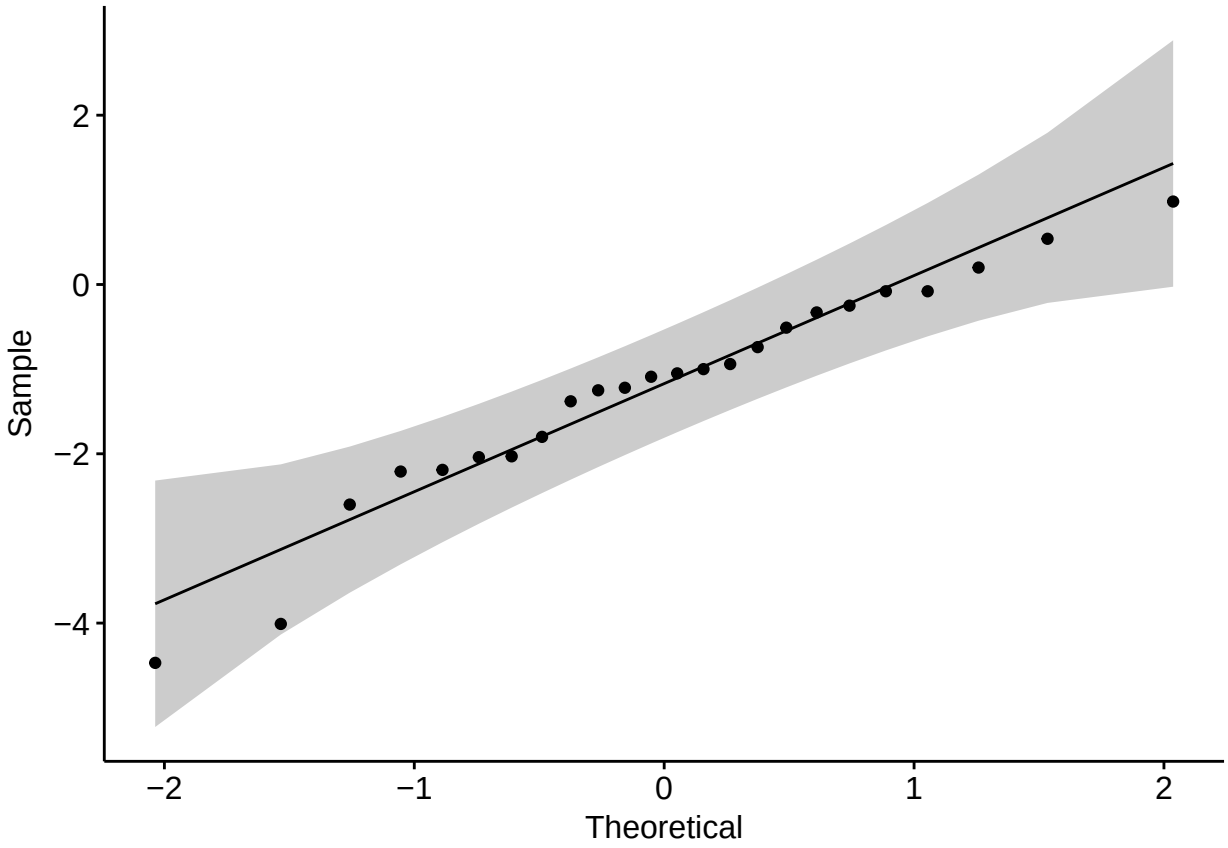
Question 3, part 1

(4 points)

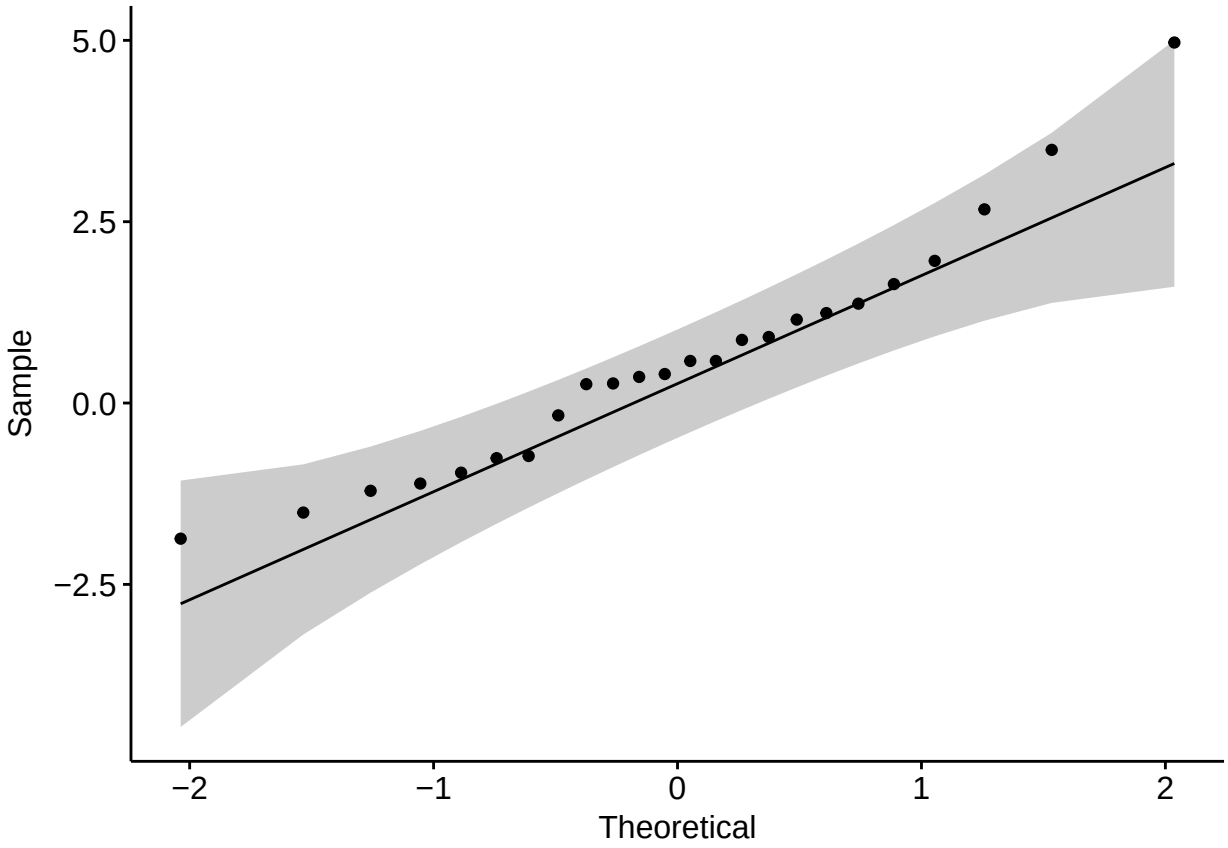
Below, a subset dat.5 of the observations in "dat_ok.csv" is selected by taking the years divisible by 5.(We are using the smaller data set partially to reduce dependence among the observation, if any. Also, the Student's t is typically a small-to-moderate sample test, so a smaller sample is used for pedagogic reasons.) Visually assess the Normality of the annual differences between jan and jul in dat.5. Does the distribution of the annual difference appear to be approximately Normal? Yes, we can verify through the qqplot that the distribution of our data is roughly normally distributed.

Do the same for the annual differences between jul and aug.

```
dat.trim<-read.csv("dat_ok.csv")
dat.5<-filter(dat.trim,year%%5==0)
dat.5$jan.jul<-dat.5$jan-dat.5$jul
ggqqplot(data=dat.5,x="jan.jul")
```




```
dat.5$jul.aug<-dat.5$jul-dat.5$aug
ggqqplot(data=dat.5,x="jul.aug")
```



The Normal quantile plots are consistent with the samples' coming from a Normally distributed population.

Question 3, part 2

(3 points)

Use the single sample Student's t test on the difference between the column dat.5$jan and the column dat.5$jul to test the hypothesis that difference of the precipitation in January and July of the same year has mean equal to 0. Please interpret your results, including your thoughts on whether the hypotheses of the test are satisfied.

The hypotheses for the test are satisfied as we previously checked the normality of the data. When performing the one sample t-test we can see that we have a p-value : 0.0001174 which is less than 0.05 so we may reject the null hypothesis that the mean is equal to 0. Alternatively, we may also look at the 95% confidence interval ([-1.7814872,-0.6810128]), and observe that the value 0 is not included which also allows us to conclude that we may reject the null hypothesis.

Please find the 95% confidence interval for the difference.

```
t.test(dat.5$jan.jul)
```

```
##
##  One Sample t-test
##
## data:  dat.5$jan.jul
## t = -4.629, df = 23, p-value = 0.0001174
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  -1.7814872 -0.6810128
## sample estimates:
## mean of x
##  -1.23125
```

Question 3, part 3

(3 points)

Use the single sample Student's t test on the difference between the column dat.5$jul and the column dat.5$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 whether the hypotheses of the test are satisfied.

The hypotheses for the test are satisfied as we previously checked the normality of the data. When performing the one sample t-test we can see that we have a p-value : 0.08148 which is more than 0.05 so we accept the null hypothesis that the mean is equal to 0. Alternatively, we may also look at the 99% confidence interval ([-0.3244723,1.5244723]), and observe that the value 0 is encompassed in it, which also allows us to conclude that we may accept the null hypothesis.

Please find the 99% confidence interval for the difference. Note this is not the default for "t.test".

```
t.test(dat.5$jul.aug,conf.level = .99)
```

```
##
##  One Sample t-test
##
## data:  dat.5$jul.aug
## t = 1.822, df = 23, p-value = 0.08148
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -0.3244723  1.5244723
## sample estimates:
## mean of x
##       0.6
```

Question 4

Question 4, part 1

(5 points)

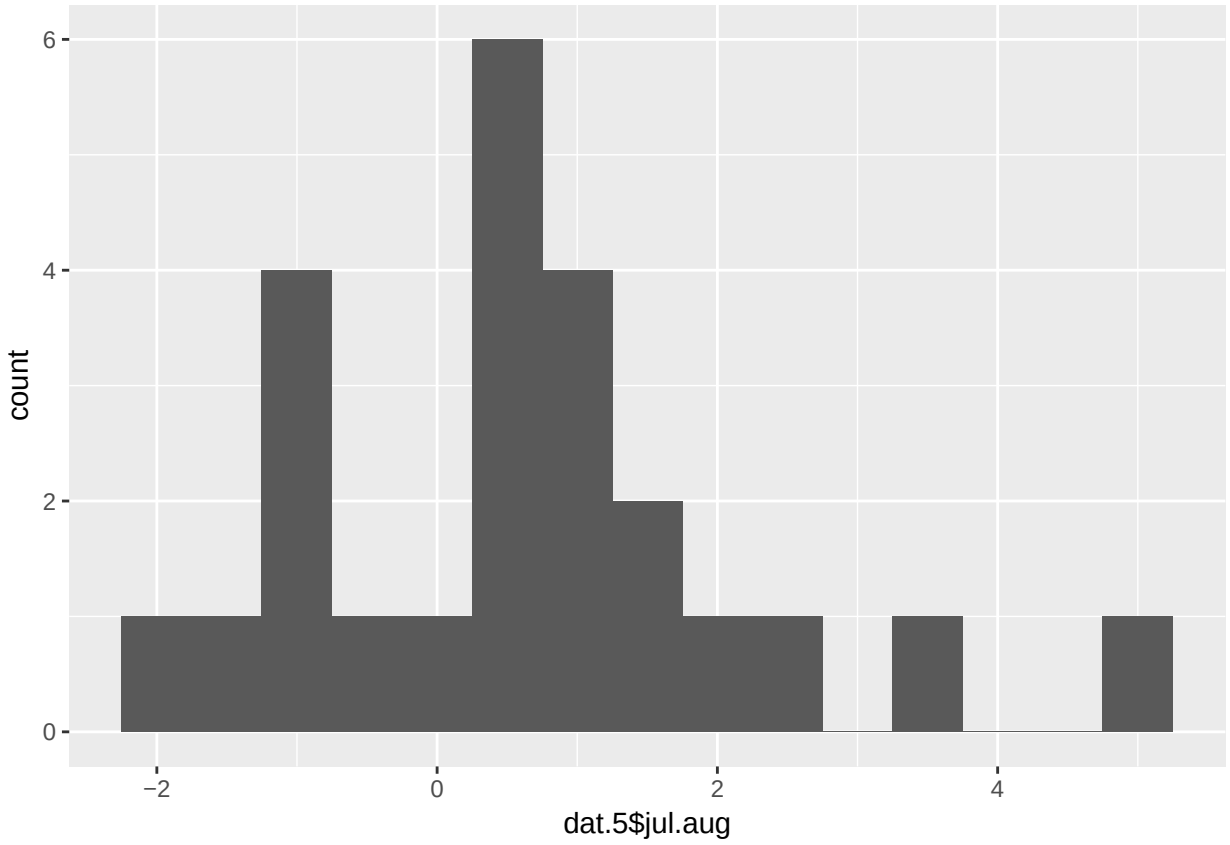
Please apply the Wilcoxon signed rank test to the difference between the column dat.5$jul and the column dat.5$aug to test the hypothesis that difference of the precipitation in July and August of the same year has mean equal to 0. Please interpret your results, including your thoughts on whether the hypotheses of the test are satisfied.

From the following plot we can observe that the distribution of dat.5$jul.aug approximately symmetric therefore we can interpret the Wilcoxon signed rank test as testing the null hypothesis that the mean of the observations is 0.

The Wilcoxon sign rank test has a p-value = 0.128 which is greater than 0.05 so we accept the null hypothesis that the mean is equal to 0.

```
dat.trim$jul.aug<-dat.trim$jul-dat.trim$aug

qplot(dat.5$jul.aug,binwidth=.5)
```




```
wilcox.test(dat.5$jul.aug)
```

```
##
##  Wilcoxon signed rank test
##
## data:  dat.5$jul.aug
## V = 204, p-value = 0.128
## alternative hypothesis: true location is not equal to 0
```

Question 4, part 2

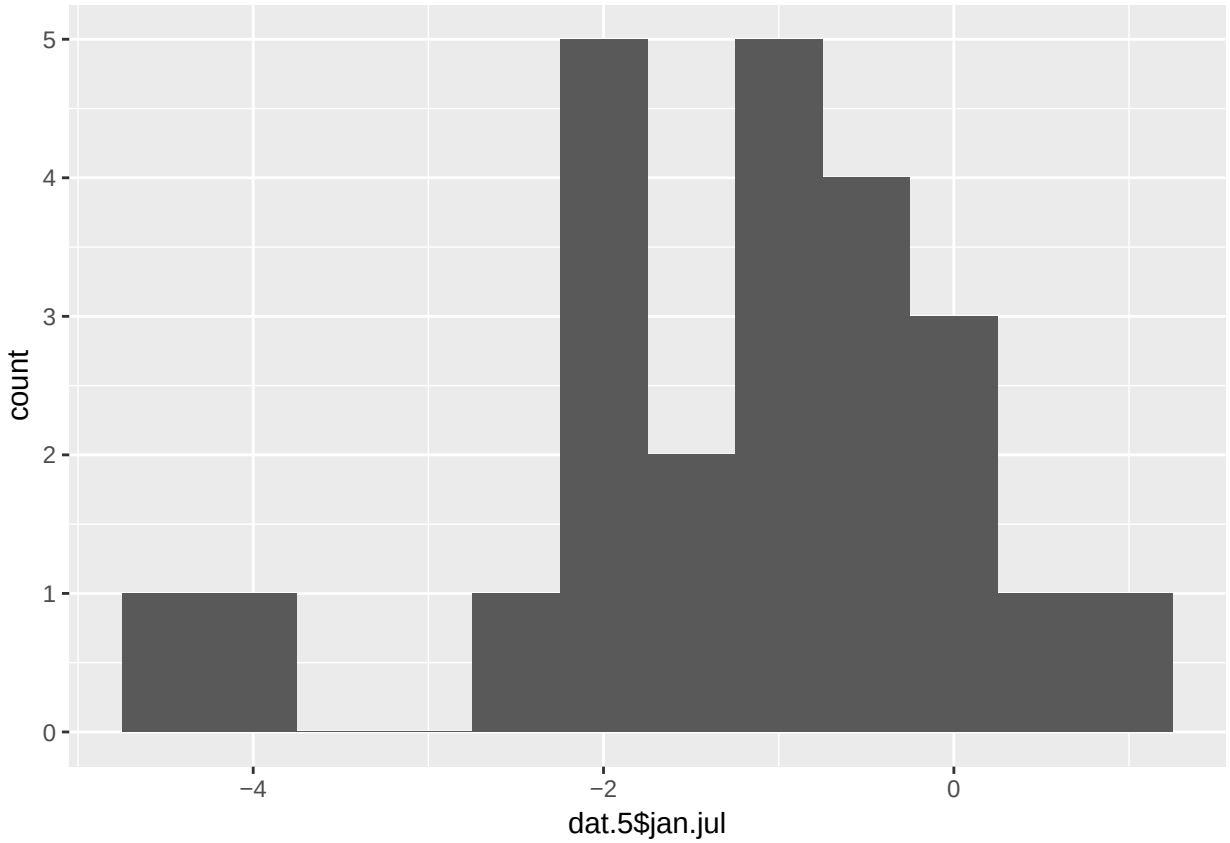
(5 points)

Please apply the Wilcoxon signed rank test to the difference between the column dat.5$jan and the column dat.5$jul to test the hypothesis that difference of the precipitation in January and July of the same year has mean equal to 0. Please interpret your results, including your thoughts on whether the hypotheses of the test are satisfied.

From the following plot we can observe that the distribution of dat.5$jan.jul is approximately symmetric therefore we can interpret the Wilcoxon signed rank test as testing the null hypothesis that the mean of the observations is 0.

```
dat.trim$jan.jul<-dat.trim$jan-dat.trim$jul

qplot(dat.5$jan.jul,binwidth=.5)
```




```
wilcox.test(dat.5$jan.jul)
```

```
## Warning in wilcox.test.default(dat.5$jan.jul): cannot compute exact p-value
## with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  dat.5$jan.jul
## V = 20, p-value = 0.0002154
## alternative hypothesis: true location is not equal to 0
```

The p-value is very small, giving strong evidence against the null hypothesis that the mean is 0. In the presence of ties, wilcox.test uses an approximation method to estimate the p-value.

One can do a Monte Carlo estimation of the p-value in the case of ties as follows. *This is not required in a correct answer to this question, and is provided here as a service to the curious.*

The function randomranks takes the vector of ranks, rankvec, and randomly assigns each value independently with probability 0.5 to be considered positive. It returns to sum of the ranks designated to be positive.

```
ranks<-rank(abs(dat.5$jan.jul))
randomrank<-function(rankvec){
  posranks.this<-rbinom(length(rankvec),1,.5)
  return(sum(posranks.this*rankvec))
}
```

Calculate the signed rank statistic v for dat.5$jan.jul:

```
ranks<-rank(abs(dat.5$jan.jul))
posranks<-(dat.5$jan.jul>0)*1
v<-sum(ranks*posranks)
```

Generate values of the signed rank statistic under the null hypothesis for the observed ranks and compare these to v. Since very few are less than or equal to v, there is strong evidence against the null hypothesis that dat.5$jan.jul is centered at 0.

```
set.seed(45678876)
vs<-replicate(100000,randomrank(ranks))
2*mean(vs<=sum(ranks*posranks))
```

```
## [1] 2e-05
```

Question 5

Please take 10000 bootstrap samples of dat.5 to create bootstrap intervals for the mean of the difference between the column dat.5$jan and the column dat.5$jul to test the hypothesis that difference of the precipitation in January and July of the same year has mean equal to 0. For now, please code the sampling yourself and use quantiles, rather than using the "boot" package. Please compare this interval to the one obtained using the t-test.

The confidence interval obtained by the t-test is : [-1.7814872, -0.6810128] and for the bootstrap sampling it's : [-1.75375, -0.74125]. These two intervals closely resemble each other, confirming the validity of the t-test for finding the confidence interval.

```
n<-nrow(dat.5)
bootmean<-function(){
  temp<-dat.5[sample(1:n,n,replace = TRUE),]
  return(mean(temp$jan.jul))
}
set.seed(5678)
jan.jul.means<-replicate(10000,bootmean())
sort(jan.jul.means)[c(250,9750)]
```

```
## [1] -1.75375 -0.74125
```