Problem Set 4

Ariel Huckabay and Kendall Weistroffer

2/9/2018

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 "precipitation_boulder.csv" 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 code provided below reads in the precipitation data. Most columns are assigned the string class. To make conversion to numeric values correct, the code replaces the value "Tr", for "trace amount" with 0, eliminates all "*"s, makes all columns numeric, and creates one data frame, dat, with all the values in the data, and one data frame, dat.trim, with only those years for which the measurements were all made at a standard site.

Question 1, Part 1

Using dat.trim, calculate the mean and sample standard deviation of the year.total precipitation for the fully valid years in the range 1900-1950 and the mean for the fully valid years in the range 2000-2017,

```
dat<-read.csv("precipitation_boulder.csv",stringsAsFactors = FALSE)
# Change all characters in the variable names to lower case.
names(dat)<-str_to_lower(names(dat))

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

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

# Replace "Tr" with "0"
dat<-mutate_all(dat,str_replace,"Tr","0")
dat<-mutate_all(dat,str_replace_all,"\\*","")
dat<-mutate_all(dat,as.numeric)
#dat$year[iffy]
dat.trim<-dat[!iffy,]

#retrieve the 1900-1950 years
dat.1950<-filter(dat.trim, dat.trim$year <= 1950)
dat.1950<-filter(dat.1950, dat.1950$year >= 1900)

#retrieve the 2000-2017
dat.2017<-filter(dat.trim, dat.trim$year <= 2017)
dat.2017<-filter(dat.2017, dat.2017$year >= 2000)

#get mean and sd from 1900-1950 data
mean.total.1950 <- mean(dat.1950$year.total)
sd.total.1950 <- sd(dat.1950$year.total)

#get mean and sd from 2000-2017 data
mean.total.2017 <- mean(dat.2017$year.total)
sd.total.2017 <- sd(dat.2017$year.total)
```

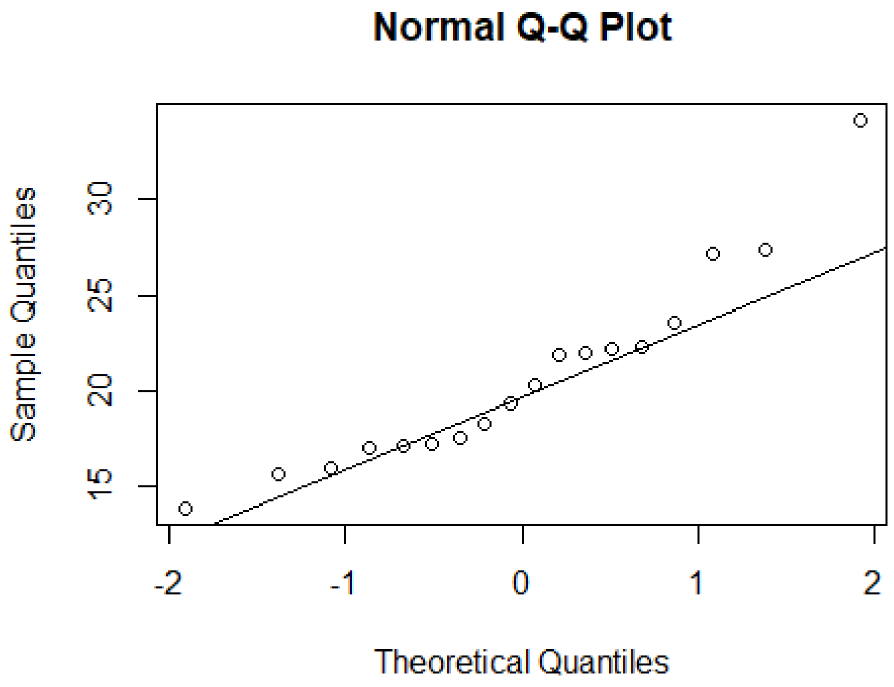
Response

For the 1900-1950s, the mean amount of yearly precipitation was 18.5409804 and the standard deviation was 4.0968003. For the 2000-2017, the mean amount of yearly precipitation was 20.7216667 and the standard deviation was 5.0548302.

Question 1, Part 2

Test the hypothesis that the mean of the annual precipitation in 2000-2017 is consistent with the null hypothesis that the precipitation in these years is a simple random sample from a Normal distribution with mean equal to the sample mean for the years in 1900-1950 and standard deviation equal to the sample standard deviation for the years in 1900-1950..

```r
#graph the normal distribution

qqnorm(dat.2017$year.total)
qqline(dat.2017$year.total)
```



```
#S-Shape in the qqnorm plot shows non-normailty of our data!
```

Response

The s-shape of the data in the qqnorm plot indicates that it is not normal, and thus it is not the case that the annual precipitation in 2000-2017is consistent with the null hypothesis.

Question 1, Part 3

Under the assumption that the precipitation totals in 2000-2017 are a simple random sample from a Normal distribution with unknown mean and variance equal to 4, please give a 95% confidence interval for the mean.

```
n = nrow(dat.2017)
sigma = sqrt(4) #Use variance to find sd
sem = sigma/sqrt(n) #Standard error of the mean

#Using .975 as 95% confidence level implies the 97.5th percentile of the norm
al dist. at the upper tail
E = qnorm(.975)*sem; #Margin of Error
xbar = mean(dat.2017$year.total)
cat("Margin of Error: ", E, "\n")

## Margin of Error:  0.9239359

cat("Sample mean: ", xbar, "\n")
```

```
## Sample mean:  20.72167

cat("Confidence Interval: ", xbar + c(-E, E), "\n\n")

## Confidence Interval:  19.79773 21.6456

#An easier, built-in solution:
library(TeachingDemos)
z.test(dat.2017$year.total, sd = sigma)

## 
##  One Sample z-test
## 
## data:  dat.2017$year.total
## z = 43.957, n = 18.0000, Std. Dev. = 2.0000, Std. Dev. of the
## sample mean = 0.4714, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.79773 21.64560
## sample estimates:
## mean of dat.2017$year.total 
##                    20.72167
```

Response

See confidence interval above.

Question 1, Part 4

Now test the hypothesis that precipitation totals in 2000-2017 are a simple random sample from a Normal distribution with mean=18.5 and unknown variance. Note the 95% confidence interval for the mean.

```
n = nrow(dat.2017)
xbar2 = 18.5 #Given mean
sampleSD = sd(dat.2017$year.total) #Sample Mean
sampleMean <- mean(dat.2017$year.total)#Sample Standard Deviation
SE = sampleSD/sqrt(n) #Standard Error Estimate
cat("Standard Error Estimate: ", SE, "\n")

## Standard Error Estimate:  1.191435

E = qt(.975, df = n-1)*SE #Margin of Error
cat("Margin of Error: ", E, "\n")

## Margin of Error:  2.513708

cat("Sample SD: ", sampleSD, "\n")

## Sample SD:  5.05483

cat("Confidence Interval: ", sampleMean+c(-E, E), "\n\n")
```

```
## Confidence Interval:  18.20796 23.23537
```

```
#An easier, built-in solution, indicates that 18.5 is not the true mean:
t.test(dat.2017$year.total, mu = xbar2, conf.level = 0.95)
```

```
## 
##  One Sample t-test
## 
## data:  dat.2017$year.total
## t = 1.8647, df = 17, p-value = 0.07958
## alternative hypothesis: true mean is not equal to 18.5
## 95 percent confidence interval:
##  18.20796 23.23537
## sample estimates:
## mean of x 
##  20.72167
```

Response

see confidence interval above.

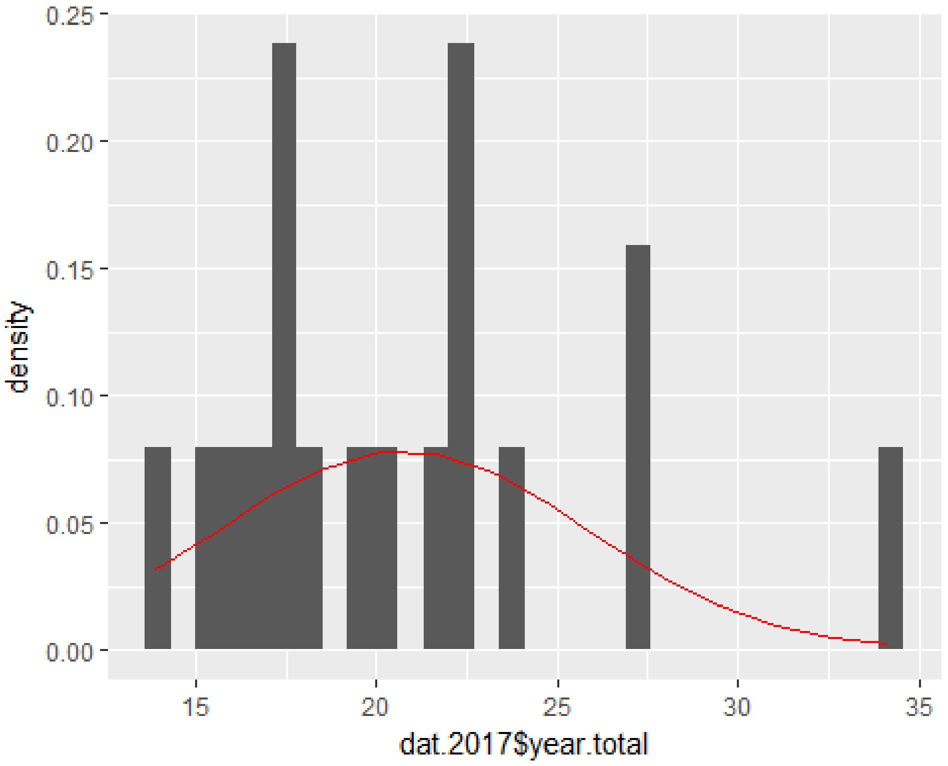
Question 1, part 5

Visually assess the Normality of the annual precipitation values in 2000-2017. Does the distribution of annual precipitation appear to be Normal?

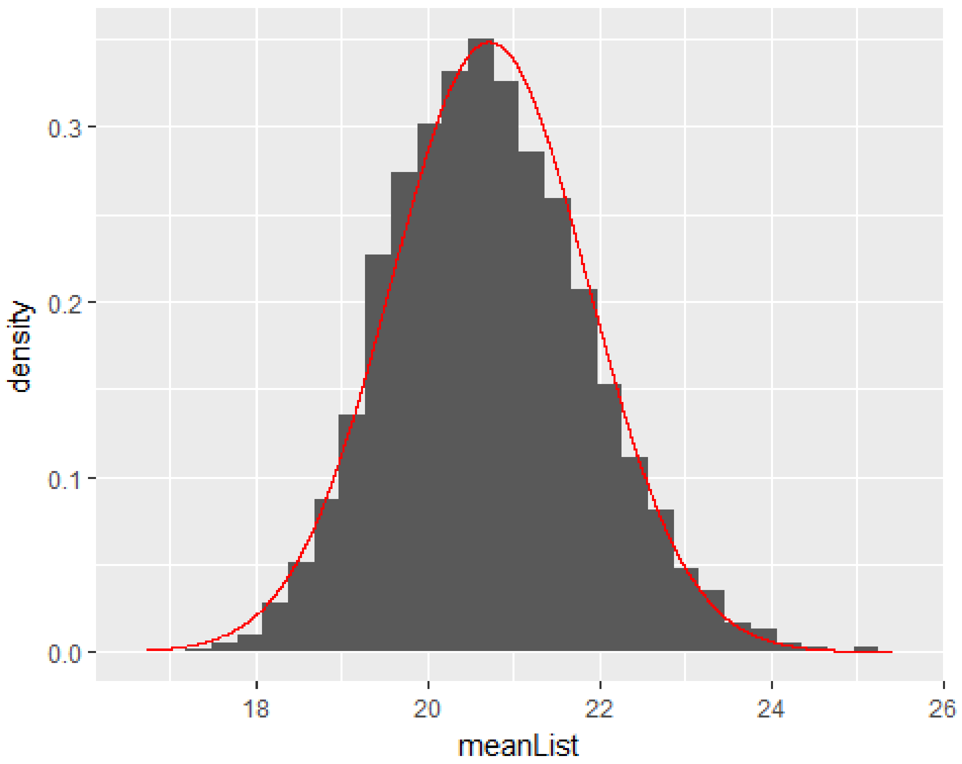
Draw 10,000 samples of size 18 from the annual precipitation values in 2000-2017. Do the means of samples of size 18 appear to be Normally distributed? Does the t-distribution seem to be a reasonable approximation?

```
#Plotting annual precipitation values for 2000-2017 to assess for normality:
frame2017 <- data.frame(dat.2017$year.total)
mean <- mean(dat.2017$year.total)
sd <- sd(dat.2017$year.total)

g <- ggplot(frame2017)+geom_histogram(aes(x = dat.2017$year.total, y = ..dens
ity..))+
  stat_function(fun = dnorm, n = length(dat.2017), args = list(mean = mean, s
d = sd), color = "red")
g
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
#Draw 10,000 samples of size 19 from annual precipitation values in 2000-2017
nSamples <- 10000
listSamples <- lapply(1:nSamples, function(x) sample(dat.2017$year.total, size = 18, replace = TRUE))

grabMeans <- function(x){
  sampleMean <- mean(x)
  return (sampleMean)

}

meanList <- sapply(listSamples, grabMeans)
meanFrame <- data.frame(meanList)

#Plot means
g1 <- ggplot(meanFrame)+geom_histogram(aes(x = meanList, y= ..density..))+stat_function(fun = dnorm, n = length(meanList), args = list(mean = mean(meanList), sd = sd(meanList)), color = "red")
g1

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Response

The annual precipitation amounts do not appear to be normally distributed. See first graph. The means appear to be normally distributed so the t-test seems to be a reasonable approximation. See second graph.

Question 1, part 6

Use the means from part 5 to give the 95% bootstrap interval for the mean of the 2000-2017 precipitation.

```
  #Use quantile method from the book (pg 62):
  cat("Bootstrapped interval: ", quantile(meanList, c(0.025, 0.975)), "\n")

## Bootstrapped interval:  18.63771 23.10561

  cat("Sample mean of the list of means: ", mean(meanList), "\n")

## Sample mean of the list of means:  20.71909
```

Response

See interval and list of means above.

Question 2, required for 4441 only.

While plots show that if X is Normally distributed, then $Y = aX + b$ looks Normal as well, we haven't actually show this analytically. If the random variable X is Normally distributed with mean μ and variance σ^2, and the random variable Y equals $aX + b$, then the cumulative distribution for Y at y equals the probability that Y is less than or equal to y, which can be expressed as an event in the probability space for X. Please write the integral for this event, and use it to show that the density of Y is the density of a Normal random variable. Give the mean and variance of Y. Remember a may be negative.

To find the cumulative distribution for Y, start with the cumulative distribution for X:

$$P(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt\,.$$

There are three cases: $a > 0, a < 0,$ and $a = 0$. We will start by applying this transformation and assuming that $a > 0$:

$P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\frac{y-b}{a}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$. This is the integral for this event. To convert back to the PDF we will substitute $t = \frac{y-b}{a}$ so that $dt = \frac{1}{a} d\left(\frac{y-b}{a}\right)$:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\frac{y-b}{a}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \rightarrow \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}} \frac{1}{a} d\left(\frac{y-b}{a}\right) = \frac{1}{a} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{\left(\frac{1}{a}(y-b-a\mu)\right)^2}{2\sigma^2}} d\left(\frac{y-b}{a}\right) =$$

$$\frac{1}{\sqrt{2\pi a^2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{\frac{1}{a^2}(y-(a\mu+b))^2}{2\sigma^2}} d\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi a^2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}} d\left(\frac{y-b}{a}\right).$$

This gives us a mean of $a\mu + b$ and a variance of $a^2\sigma^2$.

If $a < 0$, then $P(aX + b \leq y) = P\left(X \geq \frac{y-b}{a}\right)$. Looking at the cumulative distribution, we have $P\left(X \geq \frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\frac{y-b}{a}}^{\infty} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$ as the integral for this event. To convert back to the PDF we will substitute $t = \frac{y-b}{a}$ so that $dt = \frac{1}{a} d\left(\frac{y-b}{a}\right)$. Note that $\frac{1}{a}$ is negative. We will consequently substitute $-\frac{1}{|a|}$ for this quantity for easier illustration. The density of $\frac{y-b}{a}$ is thus given by $\frac{1}{\sqrt{2\pi\sigma^2}} \int_{\infty}^{-\infty} e^{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}} (-\frac{1}{|a|}) d\left(\frac{y-b}{a}\right) =$

$$-\frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\infty}^{-\infty} e^{-\frac{\left(\left(-\frac{1}{|a|}(y-b-a\mu)\right)\right)^2}{2\sigma^2}} d\left(\frac{y-b}{a}\right) = \frac{1}{|a|} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{\left(\left(-\frac{1}{|a|}(y-b-a\mu)\right)\right)^2}{2\sigma^2}} d\left(\frac{y-b}{a}\right) =$$

$\frac{1}{\sqrt{2\pi a^2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{\frac{1}{a^2}(y-(a\mu+b))^2}{2\sigma^2}} d\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi a^2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}} d\left(\frac{y-b}{a}\right)$. This gives us a mean of $a\mu + b$ and a variance of $a^2\sigma^2$, just as in the previous case.

If $a = 0$ then the distribution is constant and the mean and variance are trivially b and 0 respectively.