Problem Set 5

Ariel Huckabay & Kendall Weistroffer

Thursday, January 21, 2018

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).

Please complete the following tasks, in R where applicable. Please generate a solution document in R markdown and upload the rendered .doc, .docx, or .pdf document. You may add hand computations to a .doc or .docx if you prefer. In the rendered document, please show your code. That is, don't use "echo=FALSE".

In either case, your work should be based on the data's being in the same folder as the R files. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. Let Y be the random variable X^2 where X has a standard Normal distribution $N(0,1)$. Calculate the mean of Y. You may use what you know about the mean and variance of X. (10 points)

The variance of X equals $E[X^2] - E[X]^2$. Given that $Var[X] = 1$ and $E[X] = 0$, conclude $E[Y] = E[X^2] = 1$.

2. Returning to the data from "precipitation_boulder.csv", the file "precipitation.RData" contains the data.frame dat.trim created from the code in ps4, followed by the command

save(dat.trim,file="precipitation.RData")

The command

load("precipitation.RData")

imports dat.trim directly into the current environment.

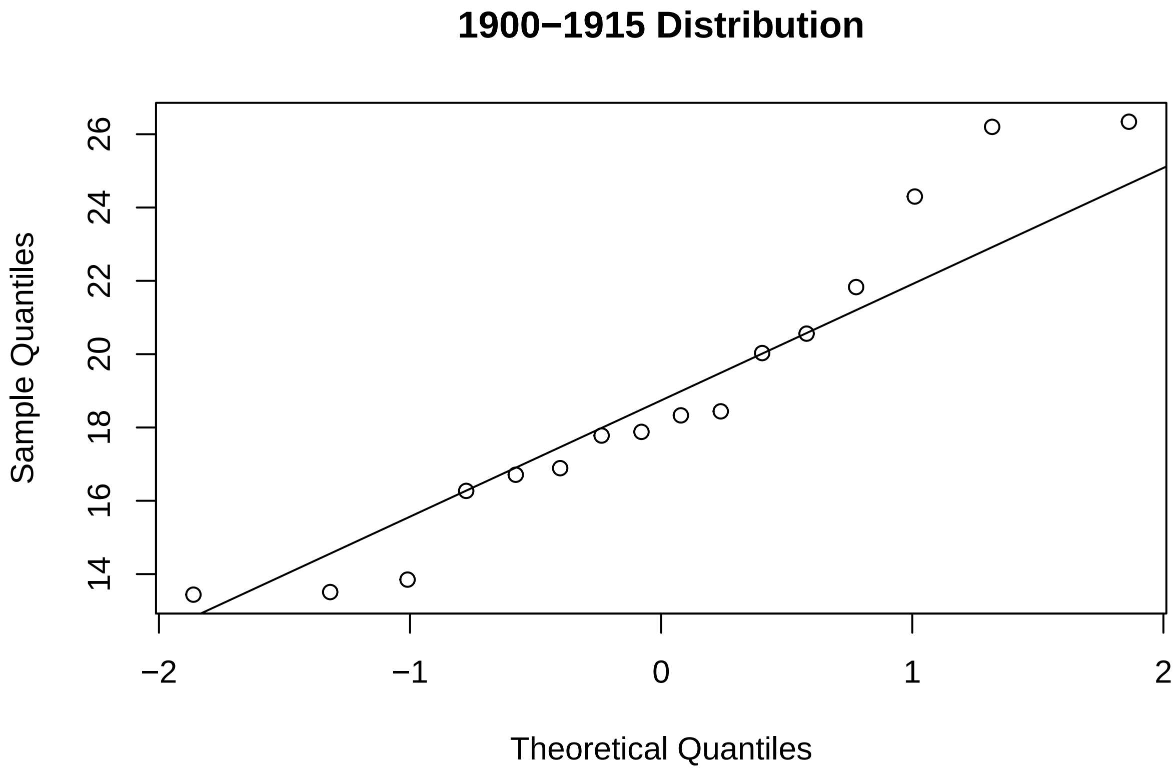
- 2.a. Are the annual precipitation totals in 1900-1915 approximately Normally distributed? Are the annual precipitation totals in 2000-2010 approximately Normally distributed? Please provide visual support for your conclusions. (10 points)

```
library(tidyverse)
```

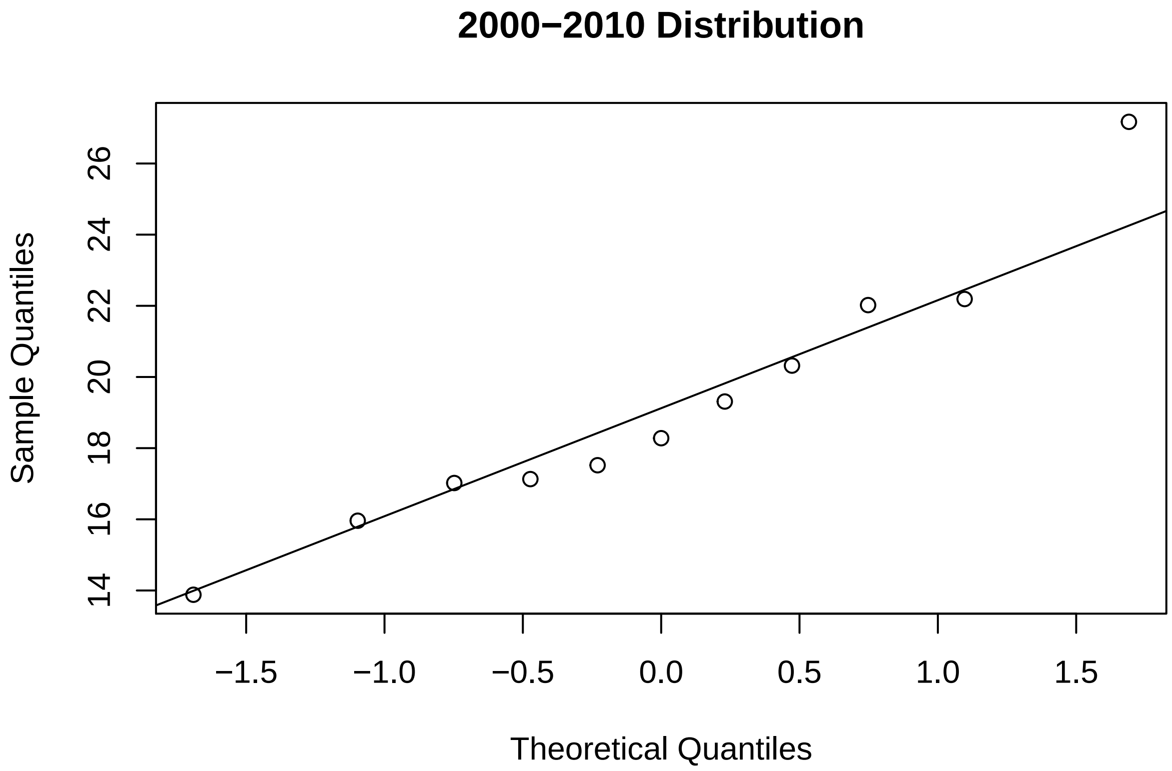
```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
load("precipitation.RData")

#retrieve the 1900-1915 years
dat.1915<-filter(dat.trim, dat.trim$year <= 1915)
dat.1915<-filter(dat.1915, dat.1915$year >= 1900)

#retrieve the 2000-2010
dat.2010<-filter(dat.trim, dat.trim$year <= 2010)
dat.2010<-filter(dat.2010, dat.2010$year >= 2000)
```

```
qqnorm(dat.1915$year.total, main="1900-1915 Distribution")
qqline(dat.1915$year.total)
```



```
qqnorm(dat.2010$year.total, main="2000-2010 Distribution")
qqline(dat.2010$year.total)
```



Response

The data from 1900 to 1915 does not appear to be normally distributed, but from 2000 to 2010 it does. This may be due to the small sample size that causes unusual years to have a more dramatic impact in the distribution.

- 2.b. Perform Welch's t test to test the hypothesis that the annual precipitation totals in 1900-1915 and the annual precipitation totals in 2000-2010 are drawn from populations with equal means. Please comment on the applicability of the test, based on your observations in 2.a. (5 points)

```
t.test(dat.1915$year.total,dat.2010$year.total, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat.1915$year.total and dat.2010$year.total
## t = -0.17678, df = 23.255, p-value = 0.8612
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.378495  2.846222
## sample estimates:
## mean of x mean of y
##  18.89750  19.16364
```

Response

From this test we would accept the null hypothesis that these samples are drawn from populations with roughly equal means.

- 2.c. Are the variances of the annual precipitation totals in 1900-1915 and the annual precipitation totals in 2000-2015 approximately equal? (5 points)

```
#The question asks for 2000-2015 rather than 2000-2010, something to clarify with Cathy
#Is it really this easy?

#retrieve the 2000-2015
dat.2015<-filter(dat.trim, dat.trim$year <= 2015)
dat.2015<-filter(dat.2015, dat.2015$year >= 2000)


t.test(dat.1915$year.total, dat.2015$year.total, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dat.1915$year.total and dat.2015$year.total
## t = -1.2823, df = 15, p-value = 0.2192
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.239444  1.303194
## sample estimates:
## mean of the differences
##               -1.968125
```

```
var.test(dat.1915$year.total, dat.2015$year.total)
```

```
##
##  F test to compare two variances
##
## data:  dat.1915$year.total and dat.2015$year.total
## F = 0.60237, num df = 15, denom df = 15, p-value = 0.3369
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2104655 1.7240435
## sample estimates:
## ratio of variances
##          0.6023717
```

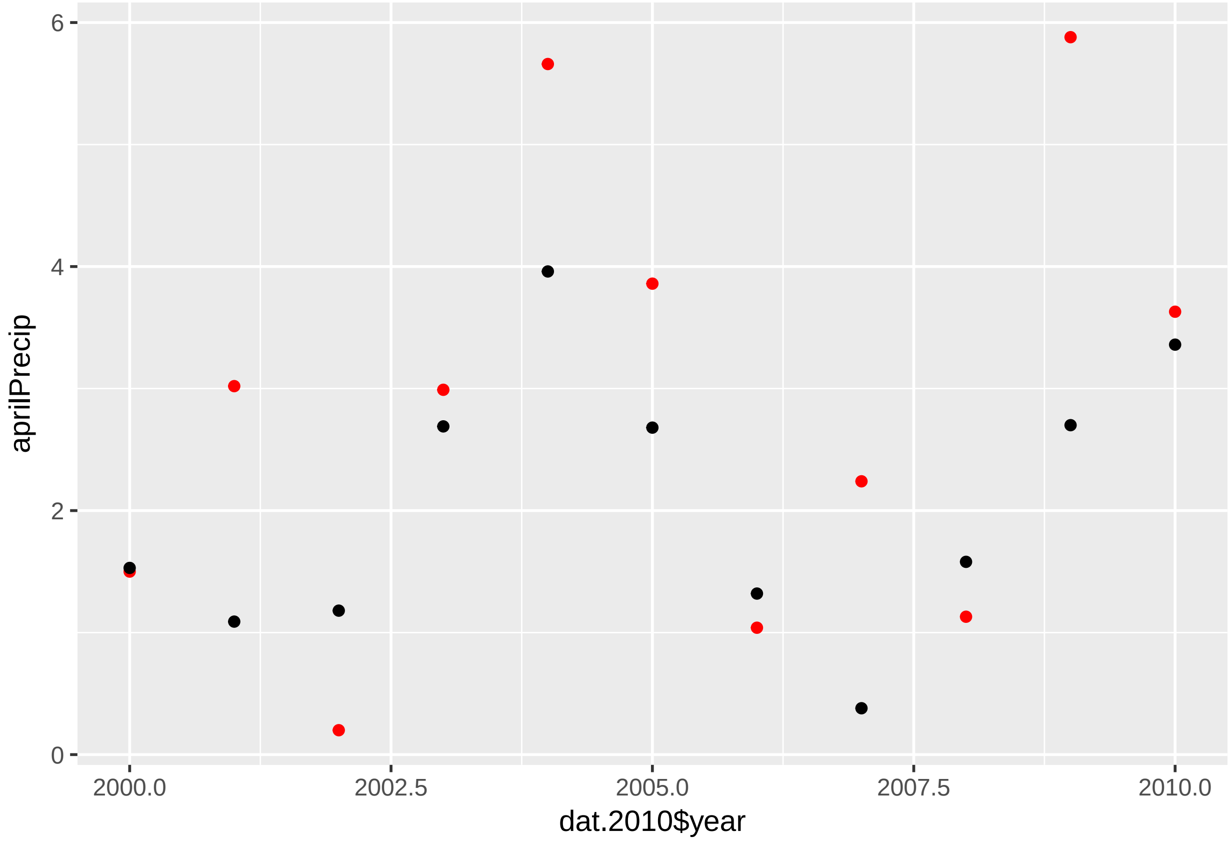
Response

Since the mean of the differences falls within the interval, we would accept the null hypothesis that the means are approximately equal rom the Paired t-test. To test the variances, the f-test was used, and it indicated that

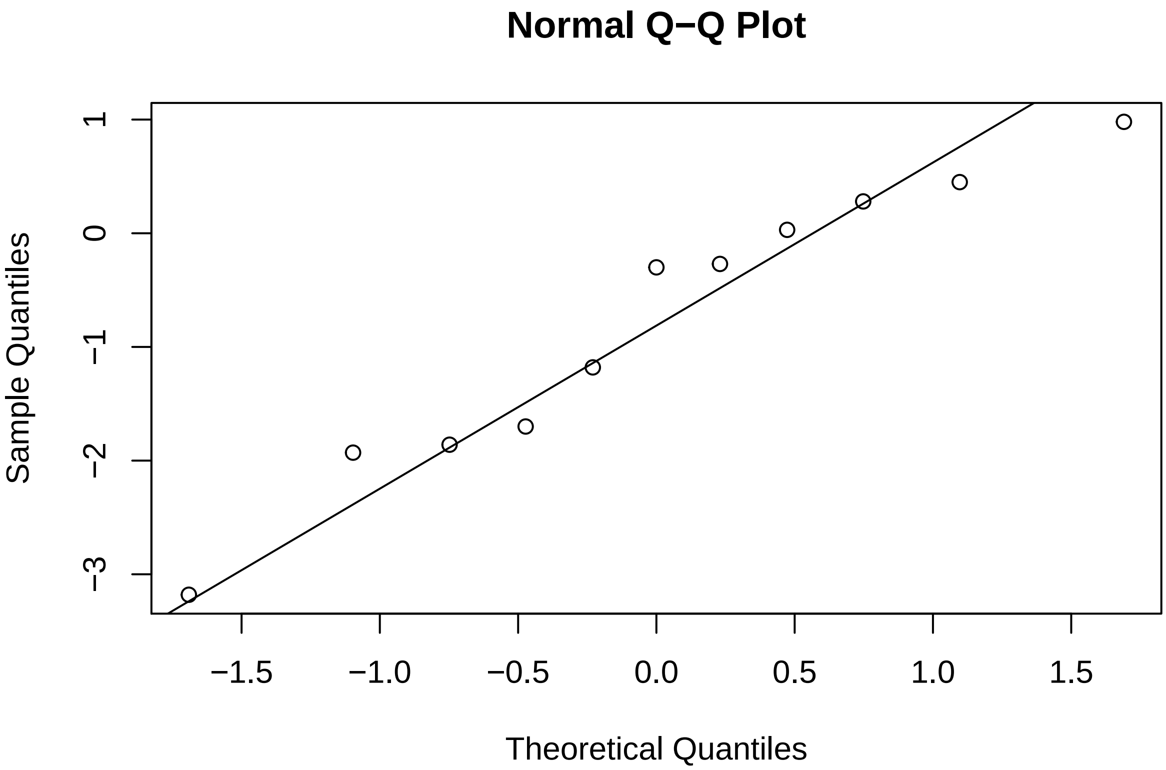
- 2.d. Test whether the April precipitation in the years 2000-2010 may reasonably be viewed as coming from a population with same mean as the corresponding June precipitation. (10 points)

```
aprilPrecip <- dat.2010$apr
junePrecip <- dat.2010$jun
```

```r
aFrame <- data.frame(aprilPrecip)
jFrame <- data.frame(junePrecip)

g1 <- ggplot(aFrame, aes(x=dat.2010$year, y = aprilPrecip)) +geom_point(color="red")
g2 <- g1+geom_point(data = jFrame, aes(x=dat.2010$year, y = junePrecip))
g2
```



```r
datDiff<-dat.2010$jun-dat.2010$apr
qqnorm(datDiff)
qqline(datDiff)
```



##Response The differences do not appear to be normally distributed, and upon visual inspection the two plots appear not to be independent. Consequently, we conclude that there is not a systematic difference between June and April in terms of rainfall in these years.

3. Please describe your project. If you are explaining a method not covered in class, what is the method? If you are addressing a question on a new data set, what is the data set? If you are working in a team of two, who is your teammate? (10 points)

Our project will use McNemar's test on Mendel's data set. McNemar's test is applied to paired data in a 2x2 contingency table wherein the outcomes of two tests performed on a sample of size n. The letter b and c denote the outcome of Test 1 being positive while Test 2 is negative and Test 1 beig negative while Test 2 is positive respectively. Its test statistic is $X^2 = (b-c)^2/(b+c)$. Its null hypothesis is that with a large enough b and c this statistic approximates the X^2 statistic. Its p-value is the twice the binomial distirubution with probabiltiy of .5 and size $b+c$.

We will evaluate Mendel's data from his experiments to determine if under this statistic it still seems suspiciously "cooked."

Partners: Ariel Huckabay and Kendall Weistroffer.