Problem Set 5

C. Durso

February 20, 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)".

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"))".

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
```

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.

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.

Question 1, Part 6

(2 points)

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

Question 2

Question 2, part 1

(5 points)

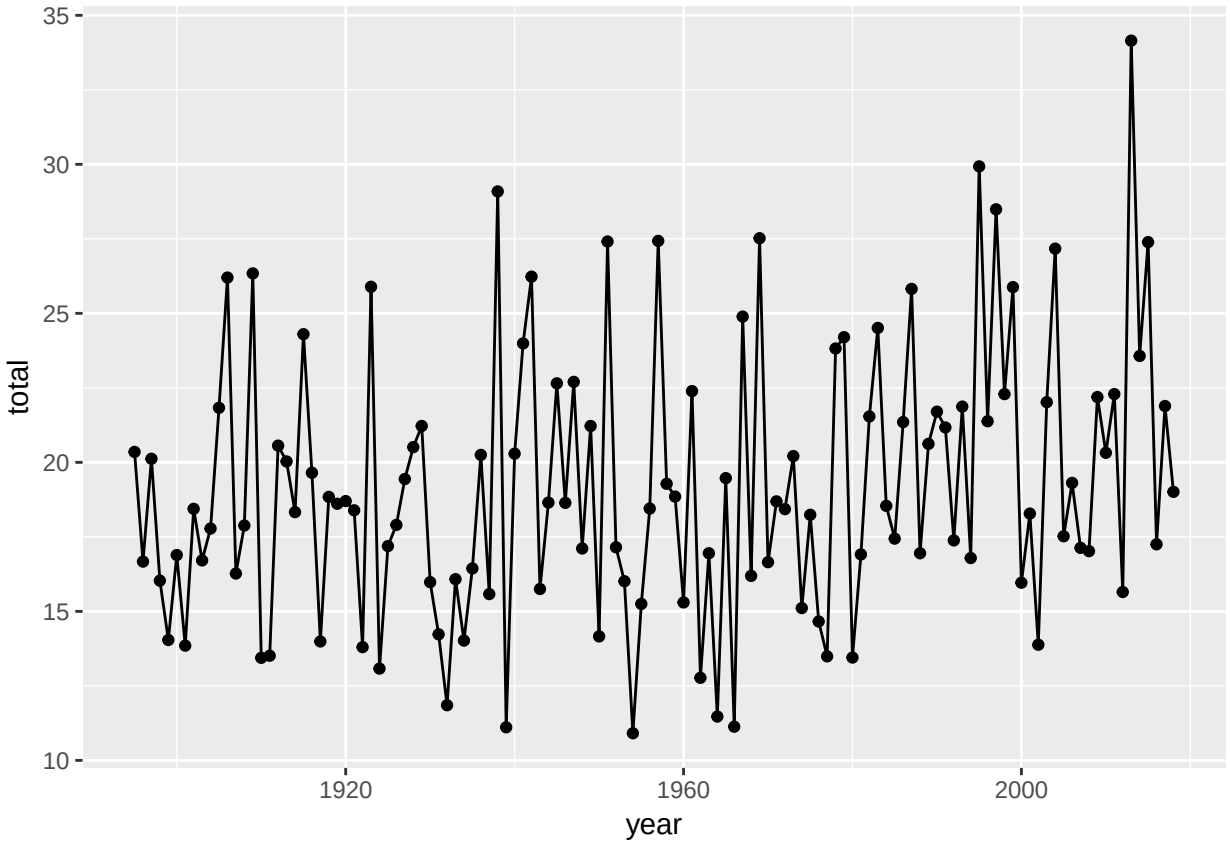
The point of this question is to study the extent to which total precipitation is one year is more like the total in the following year than it is like the total in a randomly selected year. The question uses a provided data set "dat_full.csv" like (but not exactly like) the one generated in question 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 samples 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?

```
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)
```

Question 2, Part 2

(5 points)

Please repeat the analysis from part 2 using dat$jul-dat$aug in place of dat$total, then dat$jan-dat$jul in place of dat$total. Please use the same "mat".

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 observations, 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?

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)
```

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.

Please find the 95% confidence interval for the difference.

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.

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

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.

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.

Question 5

(10 points)

Please take 10000 bootstrap samples of dat.5, that is, sample the rows of the data set with replacement to create a simulated data set of the same size as the original. 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 of the values from the samples for lower and upper end points, rather than using the "boot" package. Please compare this interval to the one obtained using the t-test.