**Unexecuted Code**

##Homework Week 4 - 30 Jul 2018

#Step 1: Write a summarizing function to understand the distribution of a vector

#----------------------------------------------------------------------

#Q1. The function, call it ‘printVecInfo’ should take a vector as input

#Creating the function

printVecInfo <- function(vec){

library(moments)

#Q2. The function should print the following information:

#a. Mean

cat("Mean:", mean(vec),"\n")

#b. Median

cat("Median:", median(vec),"\n")

#c. Min & max

cat("Min:", min(vec), "Max:", max(vec),"\n")

#d. Standard deviation

cat("Standard deviation:", sd(vec),"\n")

#e. Quantiles (at 0.05 and 0.95)

cat("quantile (0.05 - 0.95):", quantile(vec, 0.05), "--", quantile(vec, 0.95),"\n")

#f. Skewness

cat("skewness:", skewness(vec),"\n")

}

#Q3. Test the function with a vector that has (1,2,3,4,5,6,7,8,9,10,50).

vec <- c(1,2,3,4,5,6,7,8,9,10,50)

#testing on the vector

printVecInfo(vec)

#Step 2: Creating Samples in a Jar

#----------------------------------------------------------------------

#Q4. Create a variable ‘jar’ that has 50 red and 50 blue marbles

jar <- rep(c("red", "blue"), 50)

#Q5. Confirm there are 50 reds by summing the samples that are red

sum(jar == "red")

#Q6. Sample 10 ‘marbles’ (really strings) from the jar. How many are red? What was the

#percentage of red marbles?

#sampling the jar 10 times

oneSample <- sample(jar, 10, replace = TRUE)

#number of red and blue marbles in the sample

table(oneSample)

#printing the percentage of red marbles drawn

cat(sum(oneSample == "red") / length(oneSample) \* 100, "%", sep = "")

#Q7. Do the sampling 20 times, using the ‘replicate’ command. This should generate a

#list of 20 numbers. Each number is the mean of how many reds there were in 10 samples.

#Use your printVecInfo to see information of the samples. Also generate a histogram of

#the samples.

#Setting the seed for repeatable tasks

set.seed(1)

#Sampling the jar 10 times

n <- 10

#Repeating the sampling 20 times and getting the percentage of red marbles

manySamples <- replicate(20, sum(sample(jar, n, replace = TRUE) == "red") / n)

#printing the distribution statistics

printVecInfo(manySamples)

#plotting the histogram of the percentage red marbles drawn in each sample

hist(manySamples, main = "20 Samples of Red Marbles n = 10", xlab = "% Red Marbles")

#Q8. Repeat #7, but this time, sample the jar 100 times. You should get 20 numbers,

#this time each number represents the mean of how many reds there were in the 100

#samples. Use your printVecInfo to see information of the samples. Also generate a

#histogram of the samples.

#Sampling the jar 100 times

n <- 100

#Replicating the sampling of the jar 20 times, returning the percentage of red marbles

moreSamples <- replicate(20, sum(sample(jar, n, replace = TRUE) == "red") / n)

#printing information about the sampling distribution

printVecInfo(moreSamples)

#plotting the histogram of the percentage red marbles drawn in each sample

hist(moreSamples, main = "20 Samples of Red Marbles n = 100", xlab = "% Red Marbles")

#Q9. Repeat #8, but this time, replicate the sampling 100 times. You should get 100

#numbers, this time each number represents the mean of how many reds there were in

#the 100 samples. Use your printVecInfo to see information of the samples. Also

#generate a histogram of the samples.

#Setting numper of samples in the jar to 100

n <- 100

#Replicating the sampling 100 times

moreSamples <- replicate(100, sum(sample(jar, n, replace = TRUE) == "red") / n)

#Statistics of the distribution

printVecInfo(moreSamples)

#plotting the histogram of the percentage red marbles drawn in each sample

hist(moreSamples, main = "100 Samples of Red Marbles n = 100", xlab = "% Red Marbles")

#Step 3: Explore the airquality dataset

#----------------------------------------------------------------------

#Q10. Store the ‘airquality’ dataset into a temporary variable

airq <- airquality

#Q11. Clean the dataset (i.e. remove the NAs)

airq <- na.omit(airq)

#Q12. Explore Ozone, Wind and Temp by doing a ‘printVecInfo’ on each as well as

#generating a histogram for each

#Distribution of Ozone

printVecInfo(airq$Ozone)

#Histogram of Ozone

hist(airq$Ozone, main = "Ozone Histogram", xlab = "Ozone")

#Distribution of Wind

printVecInfo(airq$Wind)

#Histogram of Wind

hist(airq$Wind, main = "Wind Histogram", xlab = "Wind Speed")

#Distribution of Temperature

printVecInfo(airq$Temp)

#Histogram of Temperature

hist(airq$Temp, main = "Temperature Histogram", xlab = "Temperature")

**Console log w/plot**

**Executed code**

> ##Homework Week 4 - 30 Jul 2018

>

> #Step 1: Write a summarizing function to understand the distribution of a vector

> #----------------------------------------------------------------------

> #Q1. The function, call it ‘printVecInfo’ should take a vector as input

>

> #Creating the function

> printVecInfo <- function(vec){

+ library(moments)

+ #Q2. The function should print the following information:

+ #a. Mean

+ cat("Mean:", mean(vec),"\n")

+ #b. Median

+ cat("Median:", median(vec),"\n")

+ #c. Min & max

+ cat("Min:", min(vec), "Max:", max(vec),"\n")

+ #d. Standard deviation

+ cat("Standard deviation:", sd(vec),"\n")

+ #e. Quantiles (at 0.05 and 0.95)

+ cat("quantile (0.05 - 0.95):", quantile(vec, 0.05), "--", quantile(vec, 0.95),"\n")

+ #f. Skewness

+ cat("skewness:", skewness(vec),"\n")

+ }

> #Q3. Test the function with a vector that has (1,2,3,4,5,6,7,8,9,10,50).

> vec <- c(1,2,3,4,5,6,7,8,9,10,50)

> #testing on the vector

> printVecInfo(vec)

Mean: 9.545455

Median: 6

Min: 1 Max: 50

Standard deviation: 13.72125

quantile (0.05 - 0.95): 1.5 -- 30

skewness: 2.620396

>

> #Step 2: Creating Samples in a Jar

> #----------------------------------------------------------------------

> #Q4. Create a variable ‘jar’ that has 50 red and 50 blue marbles

> jar <- rep(c("red", "blue"), 50)

> #Q5. Confirm there are 50 reds by summing the samples that are red

> sum(jar == "red")

[1] 50

> #Q6. Sample 10 ‘marbles’ (really strings) from the jar. How many are red? What was the

> #percentage of red marbles?

> #sampling the jar 10 times

> oneSample <- sample(jar, 10, replace = TRUE)

> #number of red and blue marbles in the sample

> table(oneSample)

oneSample

blue red

4 6

> #printing the percentage of red marbles drawn

> cat(sum(oneSample == "red") / length(oneSample) \* 100, "%", sep = "")

60%> #Q7. Do the sampling 20 times, using the ‘replicate’ command. This should generate a

> #list of 20 numbers. Each number is the mean of how many reds there were in 10 samples.

> #Use your printVecInfo to see information of the samples. Also generate a histogram of

> #the samples.

> #Setting the seed for repeatable tasks

> set.seed(1)

> #Sampling the jar 10 times

> n <- 10

> #Repeating the sampling 20 times and getting the percentage of red marbles

> manySamples <- replicate(20, sum(sample(jar, n, replace = TRUE) == "red") / n)

> #printing the distribution statistics

> printVecInfo(manySamples)

Mean: 0.475

Median: 0.5

Min: 0.2 Max: 0.7

Standard deviation: 0.1551739

quantile (0.05 - 0.95): 0.2 -- 0.605

skewness: -0.6124041

> #plotting the histogram of the percentage red marbles drawn in each sample

> hist(manySamples, main = "20 Samples of Red Marbles n = 10", xlab = "% Red Marbles")



> #Q8. Repeat #7, but this time, sample the jar 100 times. You should get 20 numbers,

> #this time each number represents the mean of how many reds there were in the 100

> #samples. Use your printVecInfo to see information of the samples. Also generate a

> #histogram of the samples.

> #Sampling the jar 100 times

> n <- 100

> #Replicating the sampling of the jar 20 times, returning the percentage of red marbles

> moreSamples <- replicate(20, sum(sample(jar, n, replace = TRUE) == "red") / n)

> #printing information about the sampling distribution

> printVecInfo(moreSamples)

Mean: 0.492

Median: 0.495

Min: 0.4 Max: 0.59

Standard deviation: 0.05207282

quantile (0.05 - 0.95): 0.4095 -- 0.571

skewness: -0.1052752

> #plotting the histogram of the percentage red marbles drawn in each sample

> hist(moreSamples, main = "20 Samples of Red Marbles n = 100", xlab = "% Red Marbles")



> #Q9. Repeat #8, but this time, replicate the sampling 100 times. You should get 100

> #numbers, this time each number represents the mean of how many reds there were in

> #the 100 samples. Use your printVecInfo to see information of the samples. Also

> #generate a histogram of the samples.

> #Setting numper of samples in the jar to 100

> n <- 100

> #Replicating the sampling 100 times

> moreSamples <- replicate(100, sum(sample(jar, n, replace = TRUE) == "red") / n)

> #Statistics of the distribution

> printVecInfo(moreSamples)

Mean: 0.5007

Median: 0.51

Min: 0.36 Max: 0.62

Standard deviation: 0.04623731

quantile (0.05 - 0.95): 0.42 -- 0.56

skewness: -0.4078723

> #plotting the histogram of the percentage red marbles drawn in each sample

> hist(moreSamples, main = "100 Samples of Red Marbles n = 100", xlab = "% Red Marbles")



> #Step 3: Explore the airquality dataset

> #----------------------------------------------------------------------

> #Q10. Store the ‘airquality’ dataset into a temporary variable

> airq <- airquality

> #Q11. Clean the dataset (i.e. remove the NAs)

> airq <- na.omit(airq)

> #Q12. Explore Ozone, Wind and Temp by doing a ‘printVecInfo’ on each as well as

> #generating a histogram for each

> #Distribution of Ozone

> printVecInfo(airq$Ozone)

Mean: 42.0991

Median: 31

Min: 1 Max: 168

Standard deviation: 33.27597

quantile (0.05 - 0.95): 8.5 -- 109

skewness: 1.248104

> #Histogram of Ozone

> hist(airq$Ozone, main = "Ozone Histogram", xlab = "Ozone")



> #Distribution of Wind

> printVecInfo(airq$Wind)

Mean: 9.93964

Median: 9.7

Min: 2.3 Max: 20.7

Standard deviation: 3.557713

quantile (0.05 - 0.95): 4.6 -- 15.5

skewness: 0.4556414

> #Histogram of Wind

> hist(airq$Wind, main = "Wind Histogram", xlab = "Wind Speed")



> #Distribution of Temperature

> printVecInfo(airq$Temp)

Mean: 77.79279

Median: 79

Min: 57 Max: 97

Standard deviation: 9.529969

quantile (0.05 - 0.95): 61 -- 92.5

skewness: -0.2250959

> #Histogram of Temperature

> hist(airq$Temp, main = "Temperature Histogram", xlab = "Temperature")

