Sharat\_Sripada\_HW4.R

ssharat

2020-02-09

#  
# Course: IST-687  
# Name: Sharat Sripada  
# Homework #4  
# Due Date: 2/9/2020  
# Date Submitted: 2/9/2020  
# Topic: Samples HW  
  
# Install moments package for skewness calculation  
# install.packages("moments")  
  
# Step1-1: Summarizing function to understand distribution of a vector  
# Step1-2: Calculate mean, min, max, sd, quantile & skewness  
printVecInfo <- function(input){  
 my\_mean <- mean(input)  
 my\_median <- median(input)  
 my\_min <- min(input)  
 my\_max <- max(input)  
 my\_sd <- sd(input)  
 my\_quantile <- quantile(input, probs=c(0.05, 0.95))  
 library("moments")  
 my\_skewness <- skewness(input)  
 cat("Mean:", my\_mean, "Median:", my\_median, "Min:", my\_min,   
 "Max:", my\_max, "Std.Dev:", my\_sd,  
 "Quantile (0.05-0.95):", my\_quantile,   
 "Skewness:", my\_skewness)  
}  
  
# Step1-3:  
# Create a vector with c(1,2,3,4,5,6,7,8,9,10,50) & call   
# function printVecInfo  
myData <- c(1,2,3,4,5,6,7,8,9,10,50)  
printVecInfo(myData)

## Mean: 9.545455 Median: 6 Min: 1 Max: 50 Std.Dev: 13.72125 Quantile (0.05-0.95): 1.5 30 Skewness: 2.620396

# Step2-4:  
# Create a jar var with 50x red & 50x blue marbles/strings  
red <- replicate(50, 'red')  
blue <- replicate(50, 'blue')  
jar <- c(red, blue)  
  
# Step2-5:  
# Count if there are 50 red marbles in the jar  
count\_red <- length(jar[jar == 'red'])  
if (count\_red == 50) "There are 50 red marbles in the jar!"

## [1] "There are 50 red marbles in the jar!"

# Step2-6:  
# Sample 10 marbles from the jar & count % of red marbles  
sampleSize <- 10  
sampleSet <- sample(jar, sampleSize, replace = TRUE)  
sampleSet

## [1] "red" "blue" "red" "red" "red" "red" "blue" "blue" "red" "red"

count\_red\_sample <- length(sampleSet[sampleSet == 'red'])  
count\_red\_sample

## [1] 7

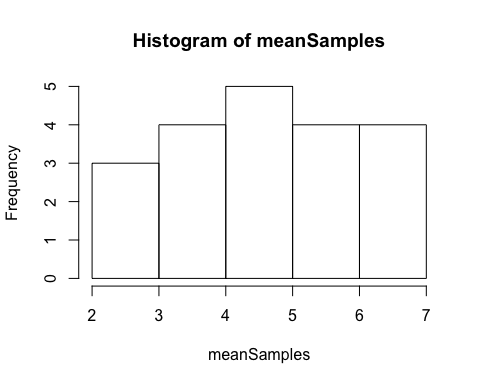
percent\_red <- count\_red\_sample/sampleSize \* 100  
percent\_red

## [1] 70

# Step2-7:  
# Sample the jar 20x times using the replicate() function  
# with sampleSize <- 10, each time counting red marbles in the sample.   
# Method to count red marbles in jar:  
# - grep for "red" - grep("red", sample(jar, sameplSize, replace=TRUE))  
# - count using length  
meanSamples <- replicate(20, mean(length(grep("red", sample(jar, sampleSize, replace = TRUE))),   
 simplify = TRUE))  
printVecInfo(meanSamples)

## Mean: 5.05 Median: 5 Min: 2 Max: 7 Std.Dev: 1.468081 Quantile (0.05-0.95): 2.95 7 Skewness: -0.2925982

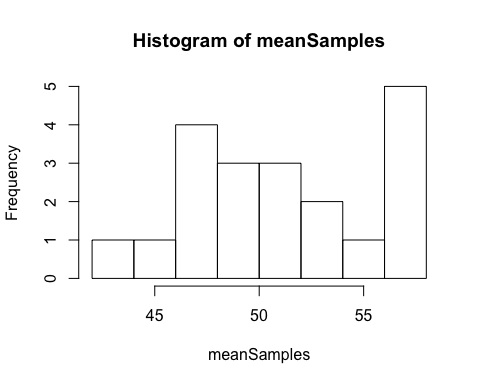
hist(meanSamples)



# Step2-8:  
# Repeat with replicate with a larger sampleSize (sample = 100)  
sampleSize <- 100  
meanSamples <- replicate(20, mean(length(grep("red", sample(jar, sampleSize, replace = TRUE))),   
 simplify = TRUE))  
printVecInfo(meanSamples)

## Mean: 51.75 Median: 51.5 Min: 43 Max: 58 Std.Dev: 4.482422 Quantile (0.05-0.95): 44.9 58 Skewness: -0.0613819

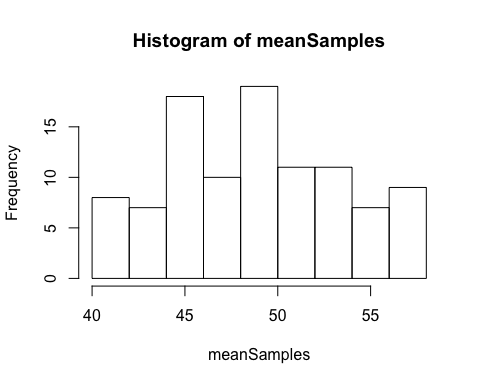
hist(meanSamples)



# Step2-9:   
# Repeat with larger replication size (replicate 100x times)  
meanSamples <- replicate(100, mean(length(grep("red", sample(jar, sampleSize, replace = TRUE))),   
 simplify = TRUE))  
printVecInfo(meanSamples)

## Mean: 49.26 Median: 49 Min: 40 Max: 58 Std.Dev: 4.600439 Quantile (0.05-0.95): 42 57 Skewness: 0.09988368

hist(meanSamples)



# Step3-10:  
# Store airquality data-set into a temp. var  
data()  
aq <- airquality  
  
# Step3-11:  
# Clean the data-set (remove NAs)  
summary(aq)

## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day   
## Min. :5.000 Min. : 1.0   
## 1st Qu.:6.000 1st Qu.: 8.0   
## Median :7.000 Median :16.0   
## Mean :6.993 Mean :15.8   
## 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :9.000 Max. :31.0   
##

dim(aq)

## [1] 153 6

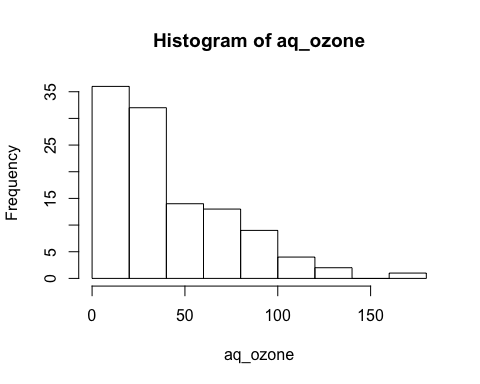
# Omit NAs from the data-set  
# NOTE - This will remove the row in full!  
aq\_omit\_na <- na.omit(aq)  
dim(aq\_omit\_na)

## [1] 111 6

# Step3-12:  
# Explore Ozone data by calling printVecInfo & hist()  
aq\_ozone <- aq\_omit\_na$Ozone  
printVecInfo(aq\_ozone)

## Mean: 42.0991 Median: 31 Min: 1 Max: 168 Std.Dev: 33.27597 Quantile (0.05-0.95): 8.5 109 Skewness: 1.248104

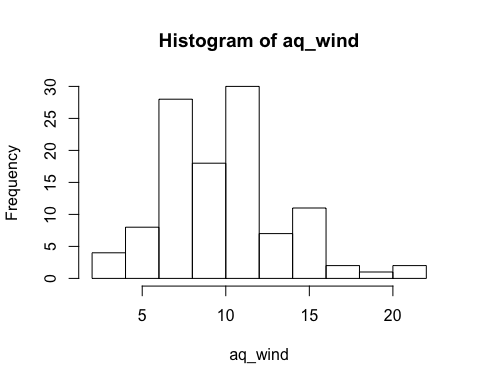
hist(aq\_ozone)



# Explore Wind data by calling printVecInfo & hist()  
aq\_wind <- aq\_omit\_na$Wind  
printVecInfo(aq\_wind)

## Mean: 9.93964 Median: 9.7 Min: 2.3 Max: 20.7 Std.Dev: 3.557713 Quantile (0.05-0.95): 4.6 15.5 Skewness: 0.4556414

hist(aq\_wind)



# Explore Temp data by calling printVecInfo & hist()  
aq\_temp <- aq\_omit\_na$Temp  
printVecInfo(aq\_temp)

## Mean: 77.79279 Median: 79 Min: 57 Max: 97 Std.Dev: 9.529969 Quantile (0.05-0.95): 61 92.5 Skewness: -0.2250959

hist(aq\_temp)

