02\_Howes\_Laura\_2018.R

Laura

Sat Sep 22 14:41:21 2018

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#######  
#Homework Week 02  
#######

@author Laura Howes 2018-09-21

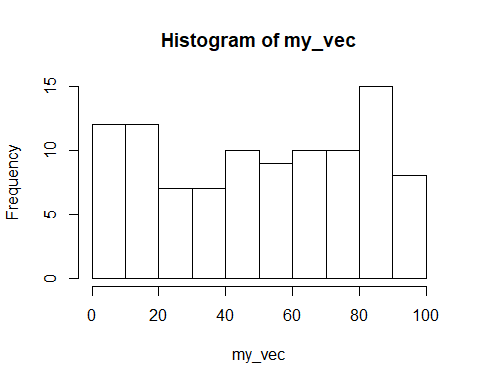
#######  
#Problem 1   
#######  
  
#1a  
my\_vec <- runif(n = 100, min = 0, max = 100)  
str(my\_vec)

## num [1:100] 58.9 17.6 72.8 84.4 62.3 ...

#### str says that the numbers in my\_vec range from 1 to 100, and then lists   
####the first 5 numbers of my vector. It tells a brief structure of my vector.  
summary(my\_vec)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.5382 23.8117 55.5066 50.8032 77.0114 99.1835

####summary my\_vec lists the Min, 1st Quartile, Median, Mean, 3rd Quartile values of my\_vec  
####summary is more descriptive than str. Str only lists that the numbers range between 1:100  
  
#1b  
hist(my\_vec)



####hist(my\_vec) plots of histogram of my\_vec, where it shows the frequency   
####of the values listed in the vector (grouped into ranges of ten,   
####eg 1-10, 11-20, etc)  
  
####the help file for hist(my\_vec) tells you about the description of   
####the function, the Usage, Arguments, Details, and Values of the   
####histogram. Also references and examples.  
  
#1c  
data("mtcars")  
class(mtcars)

## [1] "data.frame"

####Class tells you that mtcars is a data frame  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

####Str(mtcars) tells the structure of the data frame, which has 32 observations of 11 variables,  
####and lists the vectors of each column in the data frame  
summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

####Summary(mtcars) lists the Min, 1st Quartile, Median, Mean, 3rd Quartile,  
####and Max values of each column in the data frame  
  
#1d  
?data.frame

## starting httpd help server ... done

####the help file for Data Frames tells you the that the function data.frame()   
####creates data frames, tightly coupled collections of variables which share   
####many of the properties of matrices and of lists, used as the fundamental   
####data structure by most of R's modeling software.  
  
######  
#Problem 2  
######  
  
#2a  
vole\_vaso <- c(98,96,94,88,86,82,77,74,70,60,  
 59,52,50,47,40,35,29,13,6,5)  
####I messed up 3 times saying Vole vasopressin  
  
#2b  
summary(vole\_vaso)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.00 38.75 59.50 58.05 83.00 98.00

####The mean of the sample is 58.05  
####median is 59.50  
sd(vole\_vaso)

## [1] 29.75244

####sd is 29.75244  
IQR(vole\_vaso)

## [1] 44.25

####interquartile range is 44.25  
  
#2c  
SEmean <- sd(vole\_vaso)/mean(vole\_vaso)%>%  
 sqrt()  
SEmean

## [1] 3.905004

####SE of the mean = 3.90500397  
  
#2d  
####The standard error of the mean tells you how well you can esimate the   
####precision of your mean of your range of values.  
####It's a measure of the dispersion of sample means from the actual population  
####or "true" mean.  
  
#######  
#Problem 3  
#######  
  
#3a  
  
vole\_vaso\_data.frame <- data.frame(vole\_vaso)  
vole\_vaso\_data.frame

## vole\_vaso  
## 1 98  
## 2 96  
## 3 94  
## 4 88  
## 5 86  
## 6 82  
## 7 77  
## 8 74  
## 9 70  
## 10 60  
## 11 59  
## 12 52  
## 13 50  
## 14 47  
## 15 40  
## 16 35  
## 17 29  
## 18 13  
## 19 6  
## 20 5

vole\_vaso\_sample\_size\_10 <- vole\_vaso\_data.frame %>% sample\_n(size = 10, replace = TRUE)  
vole\_vaso\_sample\_size\_10

## vole\_vaso  
## 8 74  
## 19 6  
## 19.1 6  
## 16 35  
## 2 96  
## 2.1 96  
## 6 82  
## 19.2 6  
## 3 94  
## 13 50

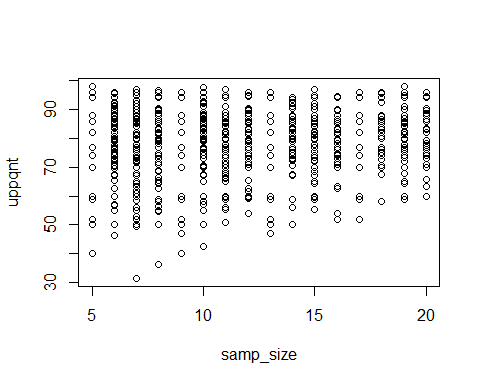
quantile(vole\_vaso, probs = 0.75)

## 75%   
## 83

vole\_vaso\_sample\_size\_10\_vector <- unlist(vole\_vaso\_sample\_size\_10)  
quantile(vole\_vaso\_sample\_size\_10\_vector, probs = 0.75)

## 75%   
## 91

####The upper quartile of the resample with a sample size of 10 is 74.   
  
#3b Build an initial data frame for simulations with the sample sizes 5 through 20.   
#Have 100 simulations per sample size.  
  
samp\_sim <- data.frame(samp\_size = rep(5:20, time = 100))   
  
#3c Use this data frame to get simulated upper quartiles for each sample size.  
  
vole\_samp\_sin\_uppqnt <- samp\_sim %>%   
 rowwise() %>%  
 mutate(uppqnt= quantile(sample(vole\_vaso, size = samp\_size, replace = TRUE), probs = 0.75))  
  
#3d With a plot, make a guesstimate as to the best sample size for   
####estimating the upper quartile of the population.  
  
plot(uppqnt ~ samp\_size, data = vole\_samp\_sin\_uppqnt)



####the best estimate of the best sample size to estimate the upper quartile is 10  
  
#4a With the upper quantile simulations, calculate the SD for each sample size using dplyr  
  
sd\_uppqnt\_sample <- vole\_samp\_sin\_uppqnt %>%  
 group\_by(samp\_size) %>%  
 summarize(sd\_up\_qnt = sd(uppqnt), ave\_upp\_quant = mean(uppqnt)) %>%  
 ungroup()

## Warning: Grouping rowwise data frame strips rowwise nature

sd\_uppqnt\_sample

## # A tibble: 16 x 3  
## samp\_size sd\_up\_qnt ave\_upp\_quant  
## <int> <dbl> <dbl>  
## 1 5 14.6 75.8  
## 2 6 11.1 78.6  
## 3 7 11.8 78.6  
## 4 8 11.7 78.7  
## 5 9 12.6 77.6  
## 6 10 11.0 79.5  
## 7 11 9.89 78.4  
## 8 12 9.20 79.2  
## 9 13 11.0 79.7  
## 10 14 8.32 81.6  
## 11 15 9.05 80.3  
## 12 16 8.48 80.0  
## 13 17 8.74 81.6  
## 14 18 6.73 82.8  
## 15 19 8.04 80.9  
## 16 20 7.43 82.5

#4b. What does this value, the standard error of the upper quartile, mean?  
  
####The standard error of the upper quartile is the standard deviation of the mean of the   
####top 25 percent of the data  
  
#4c. What is the CI of the upper quartile with a sample size of 10. What does this mean?  
  
SE\_sample\_10 <- sd\_uppqnt\_sample[6,2]  
SE\_sample\_10

## # A tibble: 1 x 1  
## sd\_up\_qnt  
## <dbl>  
## 1 11.0

Confidence\_Int\_sample\_10 <- SE\_sample\_10\*1.96  
Confidence\_Int\_sample\_10

## sd\_up\_qnt  
## 1 21.57804

#The CI of the upper quartile is sample 10 is 23.21368  
  
####the confidence interval means that there is a 95% chance the value lies within that range.