Statistics

Dhruv Aggarwal

12/6/2020

#Built-in Datasets in R  
  
data() #List of built-in Datasets in R  
  
data("mtcars") #Loading  
  
head(mtcars,6) #Print the first 6 rows

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

#Statistics  
#Generate a random sample of 100, if Replace is True then numbers can repeat  
x=sample(x=1:100, size=100,replace = TRUE)  
x

## [1] 99 54 10 2 91 73 23 43 45 67 55 66 21 32 47 83 91 39  
## [19] 79 68 17 5 69 2 80 17 3 88 21 19 59 14 50 50 74 5  
## [37] 32 21 17 38 42 1 24 100 72 97 18 59 3 67 60 89 76 14  
## [55] 54 42 18 31 95 26 30 15 58 20 74 87 17 25 80 46 66 49  
## [73] 46 31 73 32 84 54 49 97 96 78 59 98 57 21 82 36 46 37  
## [91] 59 52 75 2 64 47 82 55 71 29

#Simple Arithmetic mean  
mean(x)

## [1] 49.36

y=x  
y[sample(x=1:100, size=20,replace=FALSE)]

## [1] 46 91 32 46 80 5 76 30 57 78 67 71 49 64 66 60 59 47 73 50

y

## [1] 99 54 10 2 91 73 23 43 45 67 55 66 21 32 47 83 91 39  
## [19] 79 68 17 5 69 2 80 17 3 88 21 19 59 14 50 50 74 5  
## [37] 32 21 17 38 42 1 24 100 72 97 18 59 3 67 60 89 76 14  
## [55] 54 42 18 31 95 26 30 15 58 20 74 87 17 25 80 46 66 49  
## [73] 46 31 73 32 84 54 49 97 96 78 59 98 57 21 82 36 46 37  
## [91] 59 52 75 2 64 47 82 55 71 29

y[sample(x=1:100, size=20,replace=FALSE)] = NA  
y

## [1] 99 54 10 NA 91 NA NA 43 45 67 55 66 21 NA 47 83 91 NA 79 68 NA 5 69 2 80  
## [26] 17 3 88 NA 19 59 NA 50 50 74 5 32 21 NA 38 42 NA NA NA 72 97 18 59 NA 67  
## [51] 60 89 76 14 54 NA 18 NA 95 26 NA 15 58 20 74 87 17 25 80 46 NA 49 46 31 73  
## [76] 32 84 54 49 97 96 78 59 98 57 21 82 36 46 37 59 52 NA 2 NA 47 NA 55 71 29

mean(y) #missing values so mean is also NA

## [1] NA

mean(y,na.rm = TRUE) # Remove missing values then find mean

## [1] 52.25

#Weighted Mean  
Grades=c(95,72,87,66)  
Weights= c(1/2,1/4,1/8,1/8)  
mean(Grades)

## [1] 80

weighted.mean(x=Grades,w=Weights)

## [1] 84.625

var(x) #variance

## [1] 798.4954

sd(x)

## [1] 28.25766

sd(y) #contains missing values

## [1] NA

sd(y,na.rm=TRUE) #contains no missing values

## [1] 27.47911

min(x)

## [1] 1

max(x)

## [1] 100

median(x)

## [1] 49.5

summary(x)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 23.75 49.50 49.36 73.00 100.00

summary(y)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 2.00 30.50 54.00 52.25 74.00 99.00 20

#Quantiles  
quantile(x,probs=c(0.25,0.75))

## 25% 75%   
## 23.75 73.00

quantile(x,probs=c(0.1,0.25,0.75,0.99))

## 10% 25% 75% 99%   
## 14.00 23.75 73.00 99.01

# quantile(y,probs=c(0.25,0.75)) #missing values so error  
quantile(y,probs=c(0.1,0.25,0.75,0.99), na.rm = TRUE)

## 10% 25% 75% 99%   
## 16.80 30.50 74.00 98.21

#Correlation and Covariances  
#install.packages("ggplot2")  
library(ggplot2)   
head(economics) #BUild in Dataset of ggplot2 package

## # A tibble: 6 x 6  
## date pce pop psavert uempmed unemploy  
## <date> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1967-07-01 507. 198712 12.6 4.5 2944  
## 2 1967-08-01 510. 198911 12.6 4.7 2945  
## 3 1967-09-01 516. 199113 11.9 4.6 2958  
## 4 1967-10-01 512. 199311 12.9 4.9 3143  
## 5 1967-11-01 517. 199498 12.8 4.7 3066  
## 6 1967-12-01 525. 199657 11.8 4.8 3018

cor(economics$pce,economics$psavert) #Personal Consumption Expenditure & Personal Saving

## [1] -0.7928546

cor(economics[,c(2,4:6)])

## pce psavert uempmed unemploy  
## pce 1.0000000 -0.7928546 0.7269616 0.6145176  
## psavert -0.7928546 1.0000000 -0.3251377 -0.3093769  
## uempmed 0.7269616 -0.3251377 1.0000000 0.8693097  
## unemploy 0.6145176 -0.3093769 0.8693097 1.0000000