CM 2062 - Statistical Computing with R
Lab Sheet 6

1 Descriptive Statistics in R

Some R functions for computing descriptive statistics,

Description	R function
Mean	mean()
Standard deviation	sd()
Variance	var()
Minimum	min()
Maximum	maximum()
Median	median()
Range of values (minimum and maximum)	range()
Sample quantiles	quantile()
Generic function	summary()
Interquartile range	IQR()

Recall the "Salary" variable in "EmpFinalData" data frame.

```
Salary <- c(623.3, 515.2, 611, 729, 843.25, 578, 722.50, 632.80)
```

```
> min(Salary)
[1] 515.2
> max(Salary)
[1] 843.25
> mean(Salary)
[1] 656.8813
> median(Salary)
[1] 628.05
```

```
> range(Salary)
[1] 515.20 843.25
> IQR(Salary)
[1] 121.375
> sd(Salary)
[1] 103.0595
> var(Salary)
[1] 10621.25
> quantile(Salary)
     0%      25%      50%      75%     100%
515.200  602.750  628.050  724.125  843.250
> summary(Salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  515.2   602.8   628.0   656.9   724.1   843.2
```

To compute deciles (0.1, 0.2, 0.3, ..., 0.9), use this,

```
> quantile(Salary, seq(0, 1, 0.1))
    0%      10%      20%      30%      40%      50%      60%
515.200  559.160  591.200  612.230  620.840  628.050  650.740
   70%      80%      90%     100%
713.530  726.400  763.275  843.250
```

Let's consider the mtcars data set in R.

```
# To see the first six raws of the mtcars data set,
> head(mtcars)
                   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
```

To get a summary of a single variable in a data frame,

```
> summary(mtcars$mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.40   15.43   19.20   20.09   22.80   33.90
> summary(mtcars$disp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   71.1   120.8   196.3   230.7   326.0   472.0
```

Let's say you want to get the summary of a variable by groups,

```
> by(mtcars$mpg, mtcars$cyl, summary)

mtcars$cyl: 4
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.40   22.80   26.00   26.66   30.40   33.90
------------------------------------------------------------
mtcars$cyl: 6
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.80   18.65   19.70   19.74   21.00   21.40
------------------------------------------------------------
mtcars$cyl: 8
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.40   14.40   15.20   15.10   16.25   19.20
```

Let's import EmpFinalData.csv in to R using data import wizard and get the summury of the data.

```
> summary(EmpFinalData)
      ...1           EmpID              EmpName              Salary       Dept
 Min.   :1.00   Length:8           Length:8           Min.   :615.2   Length:8
 1st Qu.:2.75   Class :character   Class :character   1st Qu.:702.8   Class :charac
 Median :4.50   Mode  :character   Mode  :character   Median :728.0   Mode
:character
 Mean   :4.50                                         Mean   :756.9
 3rd Qu.:6.25                                         3rd Qu.:824.1
 Max.   :8.00                                         Max.   :943.2
```

Let's ty to get the summary statistics for Salary of the employees attached to each department separately.

```
> by(EmpFinalData$Salary, EmpFinalData$Dept, summary)
EmpFinalData$Dept: Finance
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  732.8   785.4   838.0   838.0   890.6   943.2
------------------------------------------------------------
EmpFinalData$Dept: HR
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    829     829     829     829     829     829
------------------------------------------------------------
EmpFinalData$Dept: IT
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  678.0   694.5   711.0   704.1   717.1   723.3
------------------------------------------------------------
EmpFinalData$Dept: Operations
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  615.2   667.0   718.9   718.9   770.7   822.5
```

Exercise
Import the "Students.csv" file in to R and do the followings.

(a) Calculate the Minimum, Maximum, mean, median, range, IQR, standard deviation, variance and quantiles for weight variable.

(b) Get the descriptive statistics for Students data frame.

(c) Calculate the descriptive statistics of height for each gender.

(d) Calculate the descriptive statistics of age for each gender.

Inbuilt Data sets in R

R comes with several built-in data sets, which are generally used as demo data for playing with R functions.

To see the list of pre-loaded data, type the function **data()**.

```
data()
```

It will show you the data sets available in "datasets" package in R.

Some of the most used R demo data sets wich are available in "datasets" package are mtcars, iris, ToothGrowth and PlantGrowth.

Let's consider the "mtcars" data set that we used earlier for visualizing data.

```
# To load the "mtcars" data set in to R
data(mtcars)
# To see the first six raws of the "mtcars" data set
head(mtcars)
# To see the description of the data set including the full variable names
?mtcars
```

Let's consider the "ToothGrowth" data set in R "datasets" package.

```
> data(ToothGrowth)
> head(ToothGrowth)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5
```

```
> ?ToothGrowth


A data frame with 60 observations on 3 variables.

[,1]     len      numeric Tooth length
[,2]     supp     factor  Supplement type (VC or OJ).
[,3]     dose     numeric Dose in milligrams/day
```

Exercise
Load the "PlantGrowth" and "iris" data sets and check for the variables.

To list the data sets in all *available* packages use,

```
data(package = .packages(all.available = TRUE))
```

Let's say you want to load the "Animals" data set in "MASS" package in to R,

```
> library(MASS)
> data(Animals)
> head(Animals)
                    body  brain
Mountain beaver     1.35    8.1
Cow               465.00  423.0
Grey wolf          36.33  119.5
Goat               27.66  115.0
Guinea pig          1.04    5.5
Dipliodocus     11700.00   50.0
> ?Animals
```

Let's load "Accident" data set in "Ecdat" package in R.

```
> library(Ecdat)
> data(Accident)
> head(Accident)
  type constr operate months acc
1    A  C6064   O6074    127   0
2    A  C6064   O7579     63   0
3    A  C6569   O6074   1095   3
4    A  C6569   O7579   1095   4
5    A  C7074   O6074   1512   6
6    A  C7074   O7579   3353  18
>  ?Accident
```

Exercise
Load some data sets in different packages in to R and check their structures.

Descriptive statistics using inbuilt data sets in R

Let's consider the "ToothGrowth" data set in R.
(a) Get the descriptive statistics of the "length(len)" variable
(b) Get the descriptive statistics of the "length(len)" variable for two different Supplement types (supp).

```
data(ToothGrowth)
head(ToothGrowth)
?ToothGrowth

summary(ToothGrowth$len)
> summary(ToothGrowth$len)
   Min. 1st Qu.  Median    Mean 3rd Qu.
   4.20   13.07   19.25   18.81   25.27
   Max.
  33.90

>  by(ToothGrowth$len, ToothGrowth$supp, summary)

ToothGrowth$supp: OJ
   Min. 1st Qu.  Median    Mean 3rd Qu.
   8.20   15.53   22.70   20.66   25.73
   Max.
  30.90
------------------------------------------------------------
ToothGrowth$supp: VC
   Min. 1st Qu.  Median    Mean 3rd Qu.
   4.20   11.20   16.50   16.96   23.10
   Max.
  33.90
```

Exercise
Consider the "iris" data set in R.
(a) Check the structure of the data set.
(b) Get the summary statistics for "iris" data set.
(c) Get the Species wise summary statistics for each numerical variable.

Home Work
Try to find any function in R to get the Species wise summary statistics for all numerical variables at once.

It's also possible to use the function **sapply()** to apply a particular function over a list or vector. For instance, we can use it, to compute for each column in a data frame, the mean, sd, var, min, quantile, ...

Example
Let's consider the "iris" data set in R.

```
# Compute the mean of each column
 >  sapply(iris[, -5], mean)

Sepal.Length  Sepal.Width Petal.Length
    5.843333     3.057333     3.758000
 Petal.Width
    1.199333

# Compute quartiles
>  sapply(iris[, -5], quantile)

     Sepal.Length Sepal.Width Petal.Length
0%            4.3         2.0          1.00
25%           5.1         2.8          1.60
50%           5.8         3.0          4.35
75%           6.4         3.3          5.10
100%          7.9         4.4          6.90
     Petal.Width
0%           0.1
25%          0.3
50%          1.3
75%          1.8
100%         2.5
```

Other functions available in different R packages to calculate descriptive statistics

stat.desc() function in in **pastecs** package

```
> install.packages("pastecs")
> library(pastecs)
> stat.desc(iris[, -5])
> results <- stat.desc(iris[, -5])
> round(results, 2)
```

```
> round(results, 2)
              Sepal.Length Sepal.Width
nbr.val             150.00      150.00
nbr.null              0.00        0.00
nbr.na                0.00        0.00
min                   4.30        2.00
max                   7.90        4.40
range                 3.60        2.40
sum                 876.50      458.60
median                5.80        3.00
mean                  5.84        3.06
SE.mean               0.07        0.04
CI.mean.0.95          0.13        0.07
var                   0.69        0.19
std.dev               0.83        0.44
coef.var              0.14        0.14
              Petal.Length Petal.Width
nbr.val             150.00      150.00
nbr.null              0.00        0.00
nbr.na                0.00        0.00
min                   1.00        0.10
max                   6.90        2.50
range                 5.90        2.40
sum                 563.70      179.90
median                4.35        1.30
mean                  3.76        1.20
SE.mean               0.14        0.06
CI.mean.0.95          0.28        0.12
var                   3.12        0.58
std.dev               1.77        0.76
coef.var              0.47        0.64
```

describe function in **psych** package

```
install.packages("psych")
 library(psych)
 describe(iris[, -5])
```

```
> describe(iris[, -5])
             vars   n mean   sd median
Sepal.Length    1 150 5.84 0.83   5.80
Sepal.Width     2 150 3.06 0.44   3.00
Petal.Length    3 150 3.76 1.77   4.35
Petal.Width     4 150 1.20 0.76   1.30
```

```
             trimmed  mad min max range
Sepal.Length    5.81 1.04 4.3 7.9   3.6
Sepal.Width     3.04 0.44 2.0 4.4   2.4
Petal.Length    3.76 1.85 1.0 6.9   5.9
Petal.Width     1.18 1.04 0.1 2.5   2.4
              skew kurtosis   se
Sepal.Length  0.31    -0.61 0.07
Sepal.Width   0.31     0.14 0.04
Petal.Length -0.27    -1.42 0.14
Petal.Width  -0.10    -1.36 0.06
```

A simple way of generating summary statistics by grouping variable is available in the psych package.

```
> describe.by(iris[, -5], iris$Species)

 Descriptive statistics by group
group: setosa
             vars  n mean   sd median
Sepal.Length    1 50 5.01 0.35    5.0
Sepal.Width     2 50 3.43 0.38    3.4
Petal.Length    3 50 1.46 0.17    1.5
Petal.Width     4 50 0.25 0.11    0.2
             trimmed  mad min max range skew
Sepal.Length    5.00 0.30 4.3 5.8   1.5 0.11
Sepal.Width     3.42 0.37 2.3 4.4   2.1 0.04
Petal.Length    1.46 0.15 1.0 1.9   0.9 0.10
Petal.Width     0.24 0.00 0.1 0.6   0.5 1.18
             kurtosis   se
Sepal.Length    -0.45 0.05
Sepal.Width      0.60 0.05
Petal.Length     0.65 0.02
Petal.Width      1.26 0.01
------------------------------------------------------------
group: versicolor
             vars  n mean   sd median
Sepal.Length    1 50 5.94 0.52   5.90
Sepal.Width     2 50 2.77 0.31   2.80
Petal.Length    3 50 4.26 0.47   4.35
Petal.Width     4 50 1.33 0.20   1.30
             trimmed  mad min max range
Sepal.Length    5.94 0.52 4.9 7.0   2.1
Sepal.Width     2.78 0.30 2.0 3.4   1.4
Petal.Length    4.29 0.52 3.0 5.1   2.1
Petal.Width     1.32 0.22 1.0 1.8   0.8
```

```
              skew kurtosis   se
Sepal.Length  0.10    -0.69 0.07
Sepal.Width  -0.34    -0.55 0.04
Petal.Length -0.57    -0.19 0.07
Petal.Width  -0.03    -0.59 0.03
```

```
group: virginica
             vars  n mean   sd median
Sepal.Length    1 50 6.59 0.64   6.50
Sepal.Width     2 50 2.97 0.32   3.00
Petal.Length    3 50 5.55 0.55   5.55
Petal.Width     4 50 2.03 0.27   2.00
             trimmed  mad min max range
Sepal.Length    6.57 0.59 4.9 7.9   3.0
Sepal.Width     2.96 0.30 2.2 3.8   1.6
Petal.Length    5.51 0.67 4.5 6.9   2.4
Petal.Width     2.03 0.30 1.4 2.5   1.1
              skew kurtosis   se
Sepal.Length  0.11    -0.20 0.09
Sepal.Width   0.34     0.38 0.05
Petal.Length  0.52    -0.37 0.08
Petal.Width  -0.12    -0.75 0.04
```

Home work
Check for other functions available in different packages in R to get different summary statistics.

Graphical representation of R inbuilt data

Exercise 1
Consider the "ToothGrowth" data set in R.

(a) Draw a suitable plot to represent "len" variable.
(b) Draw a suitable plot to represent "len" variable for each "supp" type.
(c) Draw a suitable plot to represent "len" variable for each dose.

Exercise 2
Consider the "iris" data set in R.

(a) Draw "Sepal.Length", "Sepal.Width", "Petal.Length" and Petal.Width in a same plot.
(b) Draw a suitable plot to represent "Sepal.Length" for each "Species" type.
(c) Draw a suitable plot to check whether ther any relationship among Sepal.Length", "Sepal.Width", "Petal.Length" and Petal.Width variables.