Lab Session 3

Explanatory Data Analysis Using R

Data Types in R

To make the best of the R language, you'll need a strong understanding of the basic data types. Last week we got to know Data structures and this week let us introduce basic data types. Elements of these data types may be combined to form data structures .

- **character**: `"a"`, `"swc"`
- **numeric**: `2`, `15.5`
- **integer**: `2L` (the `L` tells R to store this as an integer)
- **logical**: `TRUE`, `FALSE`
- **complex**: `1+4i` (complex numbers with real and imaginary parts)

Qualitative Data

A data sample is called qualitative, also known as categorical if its values belong to a collection of known defined non-overlapping classes. Common examples include student letter grade (A, B, C, D or F), gender(Male, female).

Qualitative variables are stored in R as **Factors.**   If qualitative data are stored as characters, the first thing you have to do is to convert them into factors using factor().

```
#Vector colour
> Colour<-c("Blue", "Green", "Green", "Red","Yellow","Blue","Red","Yellow","Blue","Blue")
> #Structure of the data ot type of data
> str(Colour)
 chr [1:10] "Blue" "Green" "Green" "Red" "Yellow" "Blue" "Red" "Yellow" "Blue" "Blue"
> Colour<-factor(Colour)
> str(Colour)
 Factor w/ 4 levels "Blue","Green",..: 1 2 2 3 4 1 3 4 1 1
```

The frequency distribution

The frequency distribution of a data variable is a summary of the data occurs in a collection of non-overlapping categories. Frequency table can be obtained by summary(), and table() commands.

```
#FREQUENCY DISTRIBUTION
> summary(Colour)
  Blue  Green    Red Yellow
```

```
    4      2      2      2
> table(Colour)
Colour
  Blue  Green    Red Yellow
    4      2      2      2
#We apply the cbind function to print the result in column format.
> freqCol<-table(Colour)
> cbind(freqCol)
       freqCol
Blue        4
Green       2
Red         2
Yellow      2
```

To get proportion, you should divide the variable by its size (length(x) for vectors and nrow(x) for data frame).

```
#to get proportions
> proCol<-freqCol/length(Colour)
> proCol
Colour
  Blue  Green    Red Yellow
   0.4    0.2    0.2    0.2
```

Let's see how we can apply these for an R built-in data frame named painters. It is a compilation of technical information of a few eighteenth-century classical painters. The data set belongs to the MASS package and has to be pre-loaded into the R workspace prior to its use.

```
library(MASS)      # load the MASS package
> painters

# print first 10 rows of painters
head(painters, n=10)

# list the structure of painters
str(painters)
```

Bar Graph

A bar graph of a qualitative data sample consists of vertical parallel bars that show the frequency distribution graphically. Create bar plots with the **barplot(***height***)** function, where *height* is a vector or matrix which is a result of table command

```
barplot(freqCol)
```

Type help("barplot") to know how to organize the figure by giving the title, axis titles, colours, etc.

Pie Chart

A pie chart of a qualitative data sample consists of pizza wedges that show the frequency distribution graphically.

```
pie(freqCol)
```

Type help ("pie") to know how to organize the figure by giving the title, axis titles, colours, etc.

Creating ordered Factors

Let X is a vector which represents class received of graduates who passed in last year. X is a ordinal variable which has natural ordering. Hence we can create a ordered factors by specifiying levels and ordered=TRUE in factor function as below.

```
x<-c("F","SU","SU","SL","F","G","SL","F","G","G","F","G","SL","G","G")
> x
 [1] "F"  "SU" "SU" "SL" "F"  "G"  "SL" "F"  "G"  "G"  "F"  "G"  "SL" "G"  "G"
> #class recived
> x<-c("F","SU","SU","SL","F","G","SL","F","G","G","F","G","SL","G","G")
> x
 [1] "F"  "SU" "SU" "SL" "F"  "G"  "SL" "F"  "G"  "G"  "F"  "G"  "SL" "G"  "G"
> Class<-factor(x, levels=c("F","SU","SL","G"), ordered=TRUE)
> Class
 [1] F  SU SU SL F  G  SL F  G  G  F  G  SL G  G
Levels: F < SU < SL < G
>
```

Exercise1

1. Draw bar plot and pie chart to school variable in painter dataset.
2. Draw bar chart for x and Class variable to see the difference in order.

Quantitative Data

Quantitative data, consists of numeric data that support arithmetic operations. This is in contrast with qualitative data, whose values belong to pre-defined classes with no arithmetic operation allowed. I will explain how to apply some of the R tools for quantitative data analysis with examples.

Histogram

One of the simplest ways to get a feel for the distribution of quantitative data is to generate a histogram. This is done using the `hist()` command. By setting the value of arguments of `hist()` we can alter the appearance of the histogram; setting `probability = TRUE` will give relative frequencies, `nclass` allows us to suggest the number of classes to use and `breaks` allow the precise breakpoints in the histogram to be specified.

```
#To load mtcars load dpylr package
library(dplyr)
data("mtcars")
hist(mtcars$mpg)
hist(mtcars$mpg, probability = TRUE)
hist(mtcars$mpg, nclass=10)
hist(mtcars$mpg, breaks=c(10,20,30,40))
# You change bin size using these options
```

Stem-and-Leaf Plot

A stem-and-leaf plot of a quantitative variable is a textual graph that classifies data items according to their most significant numeric digits. In addition, we often merge each alternating row with its next row in order to simplify the graph for readability.

```
stem(mtcars$mpg)
```

Box plot

Boxplots provide another mechanism for getting a feel for the distribution of data. Boxplots can be created for individual variables or for variables by group. The format is **boxplot(*x*, data=)**, where *x* is a formula and **data=** denotes the data frame providing the data. An example of a **formula** is y~group where a separate boxplot for numeric variable y is generated for each value of group. Add **varwidth=TRUE** to make boxplot widths proportional to the square root of the samples sizes. Add **horizontal=TRUE** to reverse the axis orientation.

```
boxplot(mtcars$mpg)
#Boxplot for each group cyl
boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

Graphical Parameters

You can customize many features of your graphs (fonts, colors, axes, titles) through graphic options.

You might want to make a permanent change to a graphical parameter. This can be done using the `par(...)` command.

```
stem(mtcars$mpg)
par() # list the graphics parameters and defaults
par(c("pch", "col"))
```

Refer https://www.statmethods.net/advgraphs/parameters.html for more information.

Summary statistics for a single group

It is easy to calculate simple summary statistics with R. Here is how to calculate the mean, standard deviation, variance, and median.

```
mean(mtcars$mpg)
[1] 20.09062
> sd(mtcars$mpg)
[1] 6.026948
> var(mtcars$mpg)
[1] 36.3241
> median(mtcars$mpg)
[1] 19.2
># Five number summary
>quantile(mtcars$mpg)
    0%    25%    50%    75%   100%
10.400 15.425 19.200 22.800 33.900
```

A nice summary display of a numeric variable is obtained from the summary function:

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

Lab Exercise

In this lab, you will determine which factors are associated with low birth weight, among other explorations.

You will prepare a word-processed document answering the questions below, integrating appropriate graphs and output. Please save your work frequently.

- The dataset is named birthweight.xls
- Start Word:

Data:

Column	Abbreviation	Variable
C1	ID	Identification Code
C2	LOW	Low Birth Weight (no = Birth Weight >= 2500g,yes = Birth Weight < 2500g)
C3	AGE	Age of the Mother in Years
C4	LWT	Weight in Pounds at the Last Menstrual Period
C5	RACE	Race (1 = White, 2 = Black, 3 = Other)
C6	SMOKE	Smoking Status During Pregnancy (Yes, No)
C7	PTL	History of Premature Labor (0 = None 1 = One, etc.)
C8	HT	History of Hypertension (Yes, No)
C9	UI	Presence of Uterine Irritability (Yes, No)
C10	FTV	Number of Physician Visits During the First Trimester
C11	BWT	Birth Weight in Grams

1) How many individuals are in the dataset?

2) Identify all the quantitative variables.

3) Construct a simple bar graph for the variable LOW.

4) Draw a pie chart for the variable RACE.What percent of mothers were white?

5) Examine the distribution of the AGE graphically using stem and leaf plot.

6) From the stemplot, what is the age of the oldest mother :_________

7) Explain where you would place an observation (by hand) to the stem-and-leaf plot that would represent a new mother who is 37 years old.

8) Describe the distribution of AGE using histogram. (identify correct bin sizes for the histogram)

9) Describe the *shape* of this distribution

10) What information about the data values could you see in a stem-and-leaf plot that you are not able to see in a histogram?

11) How does this change the information conveyed by your histogram? Which graph do you feel provides the better overall depiction and summary of the distribution? Explain.

12) To construct a *box plot* of the distribution of LWT by RACE

13) Find the mean and the median for LWT.

What conclusions can you arrive at about the shape of the distribution by looking at the two values?

Draw a histogram to support your conclusions.

14) Write a paragraph summarizing in your own words how this variable behaves. (Focus on center (e.g., mean or median), spread (e.g., standard deviation or IQR), and shape of the distribution, supporting your statements with graphical and (relevant) numerical summaries, and make sure your comments are in context.)