Linear Statistical Models

Week-1: Graded Assignment

Subjective Assignment: (Manual-grading) **Max. Marks: 25**

1. (a) Simulate 100 samples from *Binomial* distribution with parameters $n = 20$ and $p = 0.5$ by using command rbinom(100, n, p) in R-software. [1 Mark]

Solution:

Output of 100 samples from Binomial distribution on executing the command rbinom(100, 20, 0.5) is as follows:

```
> binom_sample=rbinom(100,20,0.5)
> binom_sample
  [1] 13 13  7 11 11 10 10 11  7 17  9  8 12 10  8 10 15  7 15  9
 [21]  7 10 15 12 10 11  7  8  8 10 16  7  9 12  9 11 14  9 12 10
 [41] 11 12 12  6  9 10 10 13  9  8 10  9 10  7 10 11  8  7 10 10
 [61] 12  8 11 11  9 12  9 12 13 11 12 11 10  5  9  8 12 10 11 12
 [81]  8 11 12  9 11 10 10  6  9  9 10 11 12  8  9  9 13 12  9 12
```

Since, these are 100 random values. Therefore, by executing the same command, one can get 100 different values.

(b) Find the summary of the generated dataset by using command summary() in R-software. [1 Mark]

Solution:

summary() function produce the summary of the data set and it gives descriptive statistics such as the minimum, first quantile, median, mean, third quantile and the maximum value of the input data set.

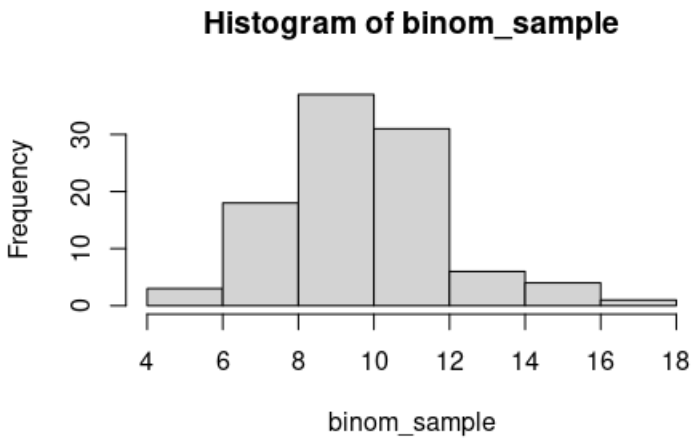
Output after executing the summary() command for the dataset of part (a) is as follows:

```
> summary(binom_sample)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0     9.0    10.0    10.2    12.0    17.0
>
```

(c) Plot the histogram of the generated dataset the by using the command hist() in R-software. [1 Mark]

Solution:

In R- software to create a histogram for a dataset, we use hist() command. Thus, the command > hist(binom_sample) will give the histogram for the data set of part (a) which is as follows:



2. (a) Simulate 100 samples from *Normal* distribution with parameters $\mu = 10$ and $\sigma^2 = 25$ by using command rnorm(100, μ, σ) in R-software. [1 Mark]

Solution:

Output of 100 samples from Normal distribution on executing the command rnorm(100, 10, 5) is as follows:

```
> normal_sample=rnorm(100,10,5)
> normal_sample
  [1] 10.5516418  8.5124764  2.2017297 14.3412290  4.2994468
  [6] 10.5431863  8.5966843 15.8781124  7.4545942 11.8526876
 [11] 21.2254033 17.1822402 10.8704506  8.8597536 -0.3419151
 [16] 13.3912287 12.6343479 21.2364221 15.3608992  3.2079983
 [21] 11.3659468  6.2104410 11.6904613 17.3762170  9.6578485
 [26] 14.8653587  5.7505307  9.7160966  4.4511153  5.2400745
 [31] 11.5075006 12.2587347 10.1463574  3.4744324  8.9670124
 [36]  9.9703367  8.7447970  9.6794987  8.0899694 12.9817192
 [41] 15.9532955 12.2931784 10.7475788 13.9440372  9.0645101
 [46] 12.3239344 10.2436957 18.6525233 15.4982639  9.6619770
 [51]  5.5102587  7.1523106  7.9874200  3.4468574  6.0483347
 [56] 13.5259238 14.1242730 16.3193928 15.4122063 14.3962876
 [61] -0.7380503  9.3412827 11.7184359 14.1342602  9.4402690
 [66] 12.3395542 13.1874202 10.1595702 18.6887458 11.5103025
 [71]  6.5520434  4.0102639  9.9778456 13.9630132 10.6417787
 [76]  4.7693031 10.9286093  6.2550806  8.0494823  3.6733662
 [81]  7.7816012  6.2676123 18.5154345  9.7830752  7.8212280
 [86]  1.7542951  4.6504336 11.7460702 16.4404561 13.0017764
 [91] 19.0526284 18.1240349  8.2279156  7.9792224 11.7234654
 [96]  7.9073576 14.4871071  4.3062911  1.8385870 17.7407394
```

Since, these are 100 random values. Therefore, by executing the same command, one can get 100 different values.

(b) Find the summary of the generated dataset by using command summary() in R-software. [1 Mark]

Solution:

summary() function produce the summary of the data set and it gives descriptive statistics such as the minimum, first quantile, median, mean, third quantile and the maximum value of the input data set.

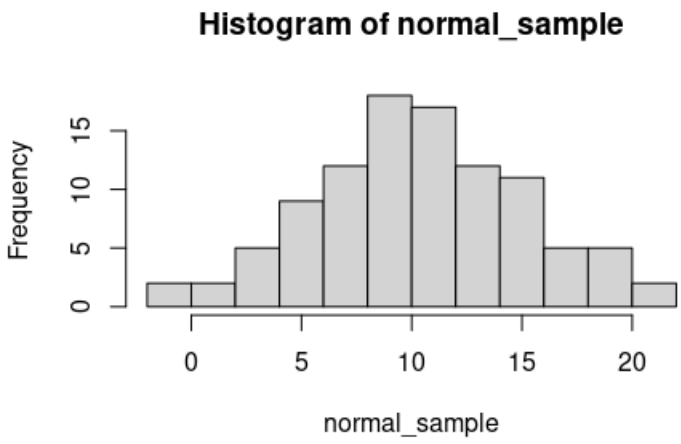
Output after executing the summary() command for the dataset of part (a) is as follows:

```
> summary(normal_sample)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.738   7.700  10.202  10.401  13.630  21.236
```

(c) Plot the histogram of the generated dataset the by using the command hist() in R-software. [1 Mark]

Solution:

In R- software to create a histogram for a data set, we use hist() command. Thus, the command > hist(normal_sample) will give the histogram for the data set of part (a) which is as follows:



3. (a) Simulate 20 samples from discrete uniform samples with parameters $b = 50$ and $a = 21$ by using the following commands in R-software: [1 Mark]

> library(purrr)
> data ← rdunif(20, b, a)

Solution:

Output of 20 samples from discrete uniform distribution on executing the command rdunif(20, 50, 21) is as follows:

```
> library(purrr)
> data=rdunif(20,50,21)
> data
 [1] 50 44 29 21 49 31 26 29 23 23 23 41 32 39 47 25 40 32 23 43
```

Since, these are 20 random values. Therefore, by executing the same command, one can get 20 different values.

(b) Find the summary of the generated dataset by using command summary() in R-software. [1 Mark]

Solution:

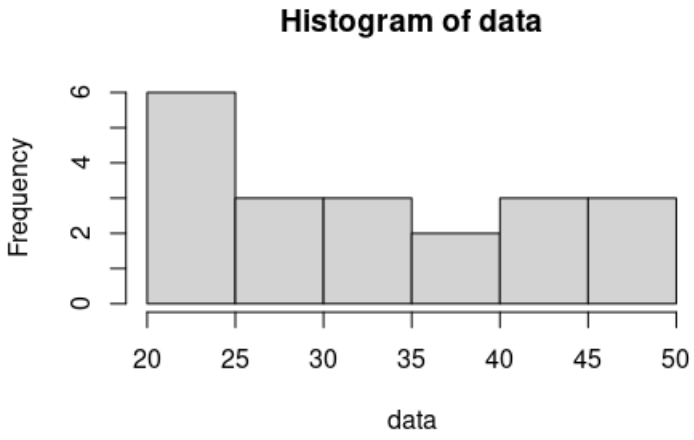
summary() function produce the summary of the data set and it gives descriptive statistics such as the minimum, first quantile, median, mean, third quantile and the maximum value of the input data set.
Output after executing the summary() command for the dataset of part (a) is as follows:

```
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21.0    24.5    31.5    33.5    41.5    50.0
```

(c) Plot the histogram of the generated dataset the by using the command hist() in R-software. [1 Mark]

Solution:

In R- software to create a histogram for a data set, we use hist() command. Thus, the command > hist(data) will give the histogram for the data set of part (a) which is as follows:



4. An analyst wishes to study inheritance of traits from generation to generation. For this, he collected the data (given in the file heights.txt) of mothers' height (Mheight) and the height of one of their adult daughter (Dheight). Based on the given information, answer the following:

   Note: *Explore and try out different commands in R for computation. Also, the plots should be properly labelled.*

   (a) Read the data as a data frame in R. [1 Mark]

   Solution:

   We can read the data as a data frame in R as follows:

```
> height_data=read.table(file = "heights.txt",header=TRUE)
> head(height_data)
  Mheight Dheight
1    59.7    55.1
2    58.2    56.5
3    60.6    56.0
4    60.7    56.8
5    61.8    56.0
6    55.5    57.9
>
```

   Where, the command read.table() is used to read the data in data frame and the command head() is used to display the first observations of first 6 rows of the data set.

   (b) Find the dimension of the dataframe using R. [1 Mark]

   Solution:

   We use dim() command to find the dimension of a dataframe. And, dimension for the data set of part(a) is given as:

```
> dim(height_data)
[1] 1375    2
>
```

Where, first element of the output represents the number of rows which is 1375 and second element represents the number of columns which is 2.

(c) Create a new column named '**Category**' in the existing dataframe which categorizes the data into two categories based on the height of daughters, i.e. 'Dheight'. The categories are defined as follows:
'**Short**': If daughter's height is less than its mean value.
'**Tall**': If daughter's height is greater than or equal its mean value. [2 Marks]

Solution:
To add a column based on the values in another columns of the data frame we can work with "dplyr". Thus, first we need to install the package by using command install.packages("dplyr"). And, we can add a new column based on the given condition by using the following codes:

```
> library("dplyr")
> height_data_1 = height_data %>% mutate(Category = if_else(Dheight<me
an(height_data$Dheight),"Short","Tall"))
> head(height_data_1)
  Mheight Dheight Category
1    59.7    55.1    Short
2    58.2    56.5    Short
3    60.6    56.0    Short
4    60.7    56.8    Short
5    61.8    56.0    Short
6    55.5    57.9    Short
```

(d) Using R, find the summary of each of the columns of the above dataframe. Comment on the dataset based on output obtained. [3 Marks]

Solution:
summary() function produce the summary of the data set and it gives descriptive statistics such as the minimum, first quantile, median, mean, third quantile and the maximum value of the input data set. And, output is as follows:

```
> summary(height_data_1)
    Mheight         Dheight        Category
 Min.   :55.40   Min.   :55.10   Length:1375
 1st Qu.:60.80   1st Qu.:62.00   Class :character
 Median :62.40   Median :63.60   Mode  :character
 Mean   :62.45   Mean   :63.75
 3rd Qu.:63.90   3rd Qu.:65.60
 Max.   :70.80   Max.   :73.10
```

As we can clearly observe from the above output that minimum height of mothers and their daughters are almost same while maximum height of daughters is more than heights of mothers. Also, we can infer that average daughter's height is more or less the same as that of average height of mothers.
Also, class of the new column "Category" is character as we have coded the values of "Dheight" column with two characters, i.e., "Short" and "Tall".

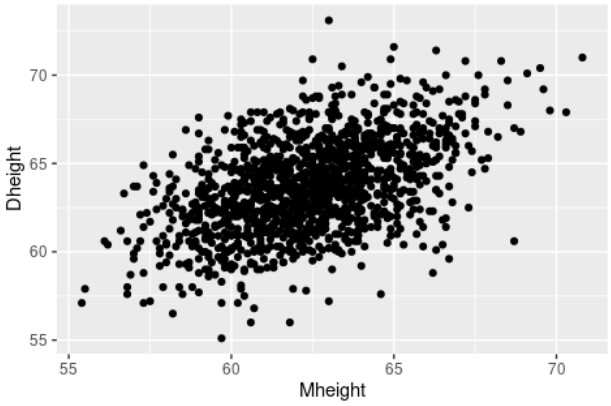
(e) Using the 'ggplot' in R, plot the scatter plot between the mother and daughter heights. [2 Mark]

Solution:

To use 'ggplot' in R, we need to install a package called "tidyverse" by using the command install.packages("tidyverse") and the scatter plot between the mother's height and daughter's heights can be plotted by using following commands

```
> library("tidyverse")
> ggplot(data=hdata)+geom_point(mapping=aes(x=Mheight,y=Dheight))
>
```

Scatter plot after executing the above commands is as follows:

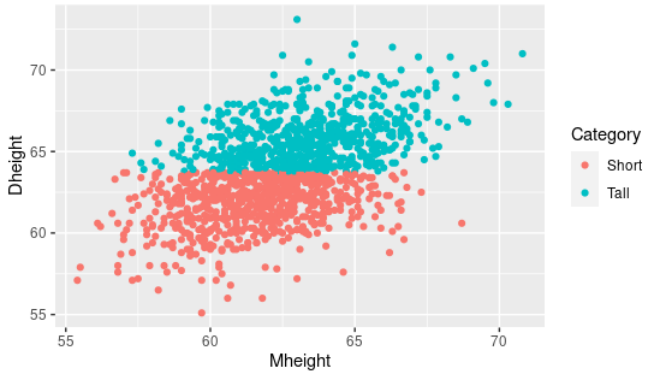


(f) Color map the plotted points in the scatter plot based on the category created in part(c). Comment on the obtained scatter plot. [2 Marks]

Solution

We can use "color" argument in ggplot2 while making graphical representation. And, code and output is as follows:

```
> colored_plot = ggplot(data = height_data_1) + geom_point(mapping = aes(x =
Mheight, y = Dheight, color=Category))
> colored_plot
>
```



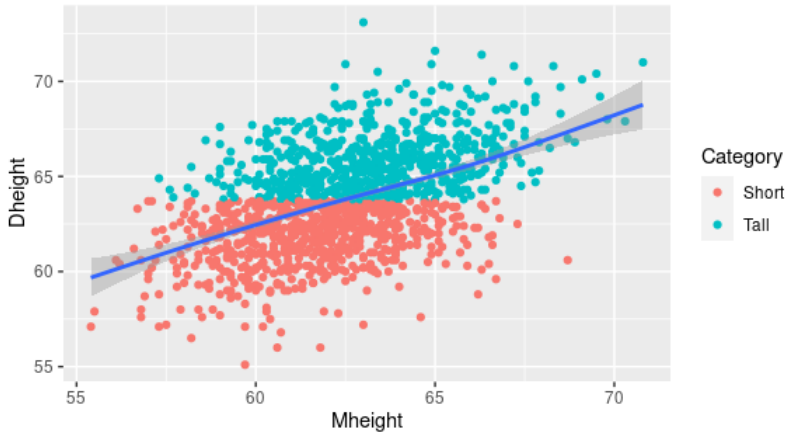
As we have categorised the heights of daughters in two different category "Short" and "Tall". In the above scatter plot, observations are displayed in two different colours and it can be easily observe that there is a clear distinction around the mean value between two categories of daughter's height ,i.e., short and tall daughters.

(g) Using 'ggplot()-geoms' in R, draw the function as a continuous curve in the above plotted scatter plot. [2 Marks]

Solution:

We can add "geom_smooth()" function to plot a continuous curve in the above plot and code and output is as follows:

```
> curve_plot = ggplot(data = height_data_1) + geom_point(mappi
ng = aes(x = Mheight, y = Dheight, color=Category))+ geom_smoo
th(mapping = aes(x = Mheight, y = Dheight))
> curve_plot
```



(h) On visualizing the scatter plot, comment on the independence of the two variables (Dheight and Mheight). [3 Marks]

Solution:

On visualizing the scatter plot, we can clearly see that there is a positive linear relationship between mother's height and daughter's height as mother's height is increasing (decreasing) then daughter's height is also increasing (decreasing).