Linear Statistical Models

Week-1: Graded Assignment

Objective Assignment: (Auto-grading) **Max. Marks: 10**

1. Heights (in inches) of all members of a family are 65, 66, 67, 67, 68, 69, 70 and 72. Based on the given information, answer the following questions.

   (a) Which statement of code in R-software can be used to read the data as a *vector* and find the number of members in the family. [1 Mark]

   (i) $> x \leftarrow$ c(65, 66, 67, 67, 68, 69, 70, 72)
   $> n \leftarrow count(x)$

   (ii) $> x \leftarrow$ c(65, 66, 67, 67, 68, 69, 70, 72)
   $> n \leftarrow total(x)$

   (iii) $> x \leftarrow$ c(65, 66, 67, 67, 68, 69, 70, 72)
   $> n \leftarrow size(x)$

   (iv) $> x \leftarrow$ c(65, 66, 67, 67, 68, 69, 70, 72)
   $> n \leftarrow length(x)$

   Answer: iv

   (b) The output obtained by executing the command to compute the number of members in the family is: [1 Mark]

   (i) 8

   (ii) [1] 8

   (iii) [1] 7

   (iv) 7

   Answer : ii

   (c) Which of the following code can be used in R-software to find the sum of the height (in inches) of family members. [1 Mark]

   (i) $> total \leftarrow add(x)$

   (ii) $> total \leftarrow sum(x)$

   (iii) $> total \leftarrow total(x)$

   (iv) $> total \leftarrow summation(x)$

   Answer: ii

   (d) The output obtained by executing the command to compute the sum of the heights (in inches) of family members is: [1 Mark]

(i) [1] 472
(ii) 544
(iii) [1] 544
(iv) 472

Answer : iii

(e) Which of the following code(s) can be used in R-software to compute the average height (in inches) of family members. [1 Mark]

(i) $> xbar \leftarrow average(x)$
$> \text{print}(xbar)$
(ii) $> xbar \leftarrow avg(x)$
$> \text{print}(xbar)$
(iii) $> xbar \leftarrow mean(x)$
$> \text{print}(xbar)$
(iv) $> xbar \leftarrow sum(x)/length(x)$
$> \text{print}(xbar)$

Answer : iii, iv

(f) The output obtained by executing the command to compute the average heights (in inches) of family members is: (Enter only the numerical value obtained by executing the command correct to 1 decimal place) [1 Mark]
Answer : 68, Range: 67.9 to 68.1

(g) Which of the following code can be used in R-software to compute the sample variance of heights (in inches^2) of family members. [1 Mark]

(i) $> svar \leftarrow var.s(x)$
$> \text{print}(svar)$
(ii) $> svar \leftarrow s.var(x)$
$> \text{print}(svar)$
(iii) $> svar \leftarrow svariance(x)$
$> \text{print}(svar)$
(iv) $> svar \leftarrow var(x)$
$> \text{print}(svar)$

Answer : iv

(h) The output obtained by executing the command to compute the sample variance of heights (in inches^2) of family members is: (Enter only the numerical value obtained by executing the command correct to 2 decimal places) [1 Mark]
Answer : 5.14, Range: 5.11 to 5.17

(i) Which of the following code can be used in R-software, compute the population variance of heights (in inches^2) of family members. [1 Mark]

(i) $> pvar \leftarrow var.p(x)$
$>$ print$(pvar)$

(ii) $> pvar \leftarrow p.var(x)$
$>$ print$(pvar)$

(iii) $> pvar \leftarrow svar * n/(n-1)$
$>$ print$(pvar)$

(iv) $> pvar \leftarrow svar * (n-1)/n$
$>$ print$(pvar)$

Answer : iv

(j) The output obtained by executing the command to compute the population variance of heights (in inches^2) of family members is: (Enter only the numerical value obtained by executing the command correct to 2 decimal places) [1 Mark]
Answer : 4.50, Range: 4.47 to 4.53

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

2. (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]

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

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

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

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

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

4. (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 $\leftarrow$ rdunif$(20, b, a)$

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

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

5. 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]

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

   (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]

   (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]

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

   (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]

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

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