Practical - II

Platform- R

1. Suppose we are given 25 observations as follows.
   29.6,28.2,19.6,13.7,13.0,7.8,3.4,2.0,1.9,1.0,0.7,0.4,0.4,0.3, 0.3,0.3,0.3,0.3,0.2, 0.2,0.1,0.1,0.1,0.1,0.1
   Draw a histogram for the given data.

2. Following are the heights of 45 female high school students. Heights are recorded in centimeters. Prepare a frequency distribution and a histogram of the data given below.
   170,151,154,160,158,154,171,156,160,157,148,165,158,159,155,151,152, 161,156,164,156, 163,174,153,170,149,166,154,166,160,160,161,154,163, 164,160,148,162,167,165,158,158, 176

3. Following table shows the frequency distribution of college students according to their pocket money (daily). Draw a histogram and frequency polygon on the same graph paper.

Pocket Money (in Rs.)	No. of students
20-29	10
30-39	24
40-49	18
50-59	12
60-69	8
70-79	5
80-89	3

4. Twenty students, graduates and undergraduates were enrolled in a Statistics course. Their ages were
   18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36

   (a) Find the median age of all students.
   (b) Find the median age of all students under 25 years.
   (c) Find modal age of all students.
   (d) Find mean age of all students.
   (e) Two more students enter the class. Age of both the students is 19. What is the mean, median and mode of all students including the new students.

   Compute mean, median and mode for the following distribution.

5. An entomologist studying morphological variation in species mosquito recorded the following data on body length:
   1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3
   Compute all the measures of dispersion.

Height (in cm)	frequency
145-150	4
150-155	6
155-160	28
160-165	58
165-170	64
170-175	30
175-180	5
180-185	5

6. For the data rivers in the base package of R, obtain all the measures of skewness.

7. Following table shows the chlorophyll content (mg/g) of 50 leaves. Compute the measures of skewness and kurtosis.

Chlorophyll content (mg/g)	No. of leaves
2.1	1
2.4	5
2.7	11
3.0	14
3.3	16
3.6	2
3.9	1

Formulae to be used

Raw Data: $x_1, \cdots, x_n$
Discrete/Ungrouped Frequency Distribution (Discrete Data): $x|f$
Continuouos/Grouped Frequency Distribution (Continuous Data): Class Interval$|f$
where, f denotes the frequency and n is number of observations.

1. Measures of Central Tendency

   (a) Mean : Raw Data: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
   Discrete Data: $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$
   Continuous Data: $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$
   where, x_i is the mid value of the ith class interval

   (b) Median: Raw Data: Arrange the data in ascending or descending order.

      i. If n is odd: Medain $= (\frac{n+1}{2})^{th}$ observation
      ii. if is even: Median= Mean of $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ observation

   Discrete Data:

   Step 1 Find the cumulative frequency (cf).
   Step 2 Find total frequency $N = \sum f$.
   Step 3 Find $\frac{N}{2}$.
   Step 4 Find the cf just greater than $\frac{N}{2}$
   Step 5 The observation corresponding to the cf just greater than $\frac{N}{2}$ will be the median.

   Continuous Data: follow the above 5 steps and obtain the median class interval.Then use the following formula.

$$\text{Median} = l + \left(\frac{N}{2} - C\right)\frac{h}{f}$$

   where, l = lower limit of the median class
   C = cf of the class just preceding the median class
   h = length of class interval
   f = frequency of median class.

   (c) Mode: Raw Data: The most frequent value will be the mode.
   Discrete Data: Mode= X value corresponding to maximum f.
   Continuous Data: Find the modal class (the class with maximum frequency) and use the following formula.

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

   where, l = lower limit of the modal class
   f_1 = frequency of modal class

$h =$ length of class interval
$f_2 =$ frequency of the class just succeeding the modal class.
$f_0 =$frequency of the class just preceding the modal class.

2. Measures of Dispersion:

   (a) Mean Deviation (MD): Raw Data: MD about $a=\frac{1}{n}\sum_{i=1}^{n}|x_i - a|$
   Discrete Data: MD about $a=\frac{1}{n}\sum_{i=1}^{n} f_i|x_i - a|$
   Continuous Data : MD about $a=\frac{1}{n}\sum_{i=1}^{n} f_i|x_i - a|$
   where, x_i is the mid value of i^{th} class interval.

   (b) Standard Deviation (SD)(σ): Raw data: $\sigma = \sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)}$
   Discrete Data: $\sigma = \sqrt{(\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2)}$
   Continuous Data : $\sigma = \sqrt{(\frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2)}$
   where, x_i is the mid value of i^{th} class interval.

3. Measures of Skewness:

   (a) Karl Pearson's coefficient of skewness

   $$S_k = \frac{Mean - Mode}{\sigma}$$

   . If mode is not well defined for the data then

   $$S_k = \frac{3(Mean - Median)}{\sigma}$$

   Here, $-3 \leq S_k \leq 3$. If $-3 \leq S_k < 0$, then the data is negatively skewed. if $0 < S_k \leq 3$ then the data is positively skewed. if $S_k = 0$, the data is not skewed that means it is symmetric.

   (b) Coefficient of skewness based on moments

   $$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

   where μ_2 and μ_3 are the second and the third central moments.
   If $\beta_1 = 0$, the data is not skewed i.e. it is symmetric.
   If $\beta_1 > 0$, the data is positively skewed.
   If $\beta_1 < 0$, the data is negatively skewed.

4. Mesures of Kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$\gamma_2 = \beta_2 - 3$$

where μ_2 and μ_4 are the second and the fourth central moments.
If $\beta_2 = 3(\gamma_2 = 0)$, then the curve is normal(mesokurtic).
If $\beta_2 < 3(\gamma_2 < 0)$ then the curve is platykurtic (flatter than normal curve).
If $\beta_2 > 3(\gamma_2 > 0)$ then the curve is leptokurtic (more peaked than the normal curve).

Remark: Raw Data $\mu_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^i$
Discrete Data $\mu_i = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^i$
Continuous Data $\mu_i = \frac{1}{n}\sum_{i=1}^{n} f_i(x_i - \bar{x})^i$
where, x_i is the mid value of i^{th} class interval.