# Assignment


Name :- Avinash Gautam
Course :- B.Sc. (Hons) Computer Science
Roll No. :- 20201407
Exam Roll No. :- 20020570009

Subject :- Statistical Methods (G.E.)

# Assignment

Q. ①

Ans. (i) Frequency distribution :-

Frequency may be defined as the number of individuals having same measurements or laying in the same measurement group. Frequency distribution is the distribution of frequencies over different measurement.

Frequency distribution has two types.

(a) Discrete frequency distribution

(b) Continuous frequency distribution

We use struges rule for constructing the continuous frequency distribution.

Steps of Struges Rule :-

① First we find maximum and minimum value from the given data-set.

② Then find the range.

$$\text{range} = \text{max} - \text{min}$$

③ Find the $N$ = no. of observations

④ Find the $k$ = no. of classes

$$k = 1 + 3.322 \log_{10} N$$

⑤ Find the class width (size of class interval)

$$\text{width} = \text{range}/k$$

After all the calculations we create the class intervals of equal width and find and write the no. of frequencies of corresponding class intervals' data. Now we get the continuous frequency distribution in exclusive form.

(ii) Histogram :-

It is a graphical representation of an exclusive grouped frequency distribution. It consists of a set of rectangles one over each ~~other~~ class interval having their areas propotional to corresponding class frequency. Class intervals are plotted on x-axis and the class frequencies are plotted against the y-axis.

If the class intervals are of unequal length then we adjust the height (H) of rectangles by using this fact that area is propotional to Frequency.

$$H = \left(\frac{\text{freq.}}{\text{width}}\right) \times k$$

k = minimum class width

Histogram is used to locate the mode a major of central tendancy. Also two or more frequency distribution can be compared with the help of histogram.

* Frequency-polygon :-

It is an another representation of an exclusive grouped frequency distribution. It can be constructed in two ways.

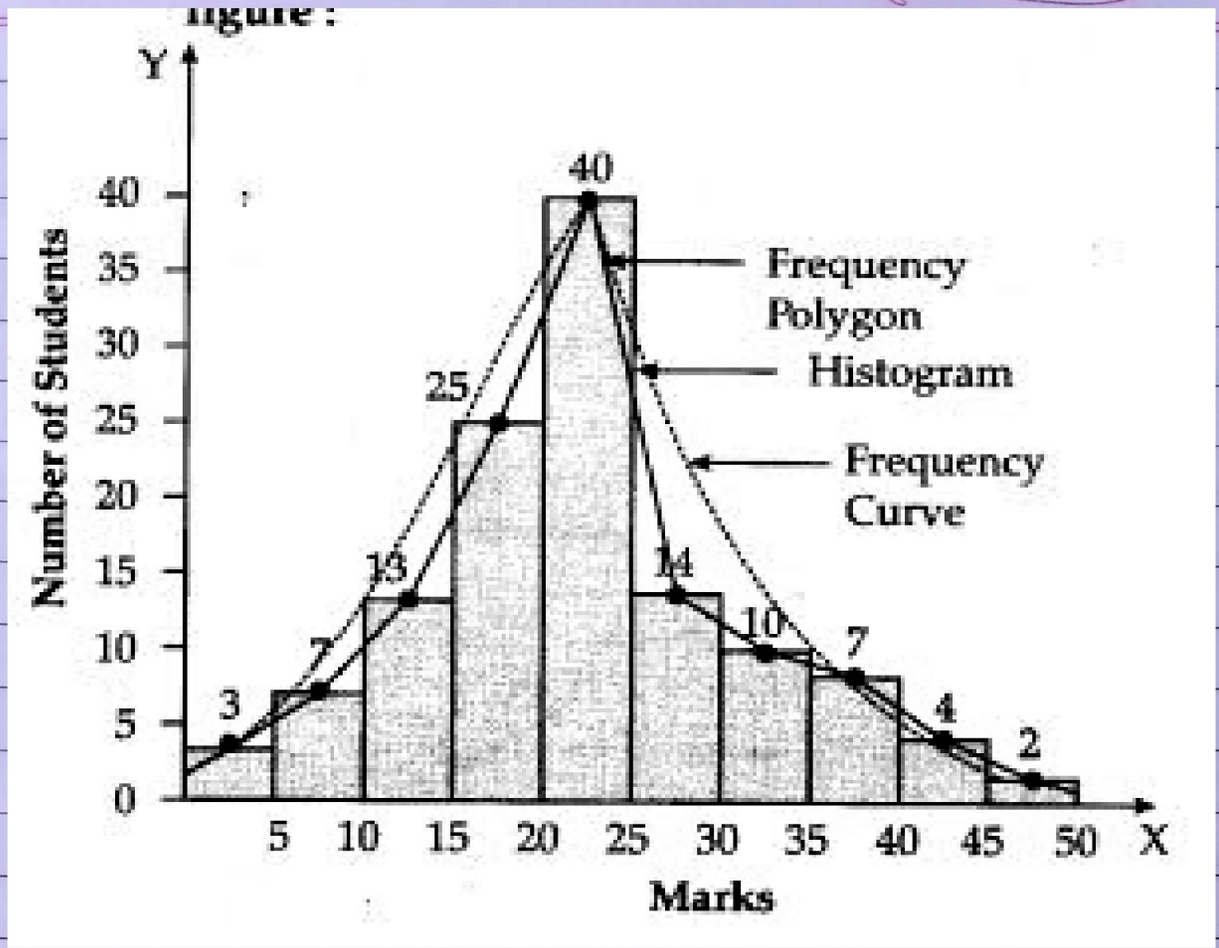
① The class frequency of each class are plotted against the mid values of the class intervals and the plotted points are joined through straight lines.

② First a histogram is made then the mid points of the upper horizental sides of rectangles are joined by straight lines.

At the end, the lines may be joined to the base at the mid point of two class intervals of 0 frequency outside the histogram although it is not necessary. It is preferred for graphical representation of frequency distribution as two or more polygons can be drawn on the same graph.

* Frequency Curve :-

To make a frequency curve for a given frequency distribution, the mid points of the class intervals are joined smoothly in such a way that the area included is just the same as that of the histogram and polygon. If the class intervals are smaller and smaller, the original class frequency remains constant then the histogram and polygon approach more and more closely to become a smooth curve in the limit. The frequency curve represents more truely the distribution of continuous measurements.



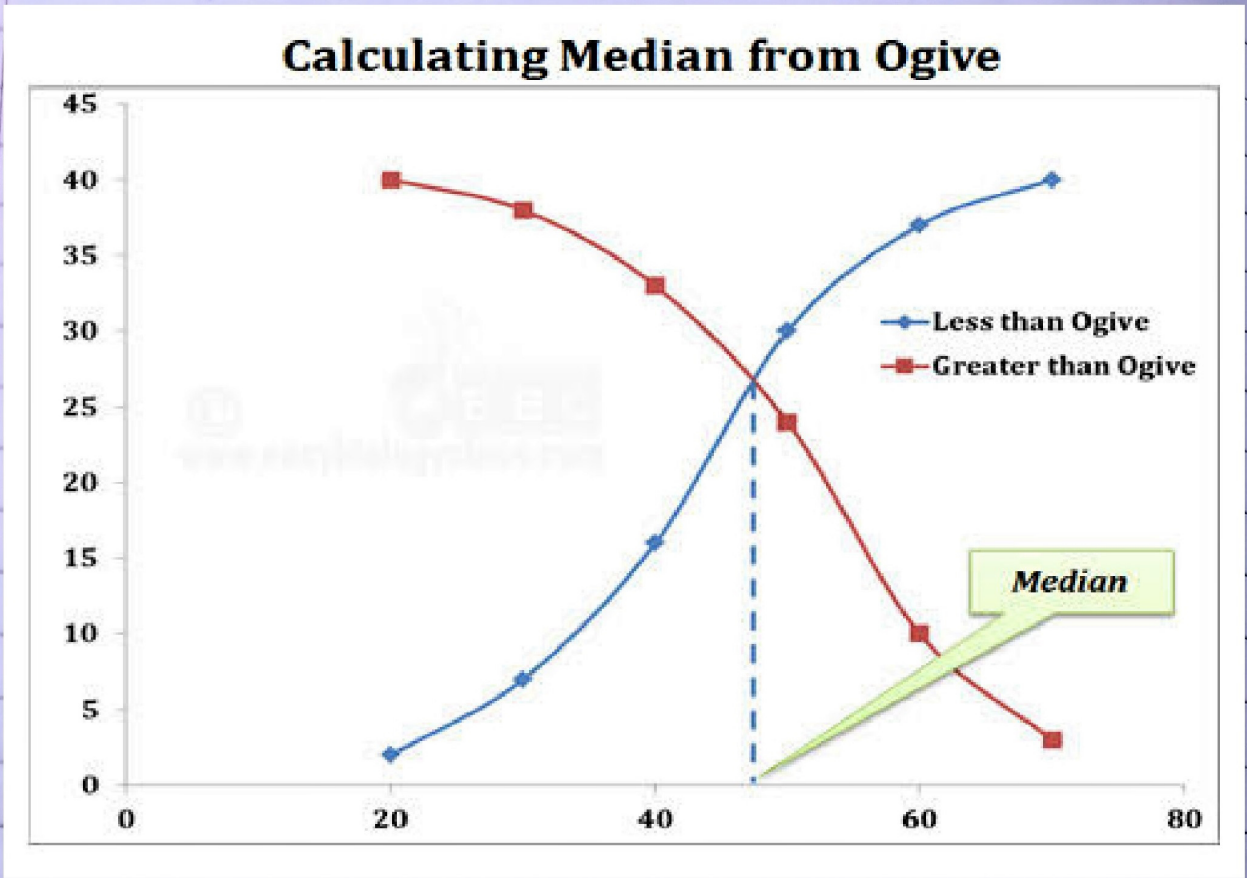

(iii) Ogive :-

It is a graphical representation of cumulative frequency distribution. It is drawn with the exclusive class intervals. It has two types :-

① Less than type ogive :- When the less than type cumulative frequencies are plotted against the corresponding ~~to~~ upper class limit and points are joined by ~~straight~~ smooth curve then we get the less than type ogive.

② More than type ogive :- When the more than cumulative frequencies are plotted against the corresponding lower class limit and points are joined by smooth curve then we get more than type ogive.

In the graph where these two types of ogive are intersect, that the point on x-axis corresponding intersection point is the median of frequency distribution.




* Properties of a good average :-

① It should be simple to calculate and easy to understand.
② It should be rigidly defined.
③ Its computation be based on all the observations.
④ It should be capable of further algebraic treatment.
⑤ It should not be affected by extreme items.
⑥ It should be least affected by sampling fluctuations.

* Examine the properties with A.M., G.M., H.M. :-

| Arithmetic Mean | Geometric Mean | Harmonic Mean |
|---|---|---|
| ① It's regidily defined. | ① It's regidily defined. | ① It's regidily defined. |
| ② It's easy to understand & calculate. | ② It's not easy to calculate. | ② It's not easy to calculate & understand. |
| ③ It's based on all values. | ③ It's also based on all values. | ③ It's also based on all values. |
| ④ It's affected by extreme values. | ④ It is very much affected by extreme values. | ④ It is also very much affected by extreme values. |

All of three can be used for ~~alger~~ algebraic treatment.

* Suitable situations :-

① A.M. should be used for the average of a data set or the raw values such as stock prices.

② G.M. should be used when dealing with a set of percentages which are derived from raw values. Eg. - percent change in stock prices.

③ H.M. can deal with fraction denominators. eg. - the P/E, EV/EBITDA ratios.

Q ②

Ans. **Moment :-**

Moments are a set of statistical parameters to measure a distribution. Generally, in any frequency distribution four moments are obtained. These four moments describe the information about mean, variance, skewness and kurtosis of a frequency distribution that gives us some important statistical features.

Three types of moments are :-

① Moments about arbitary point
② Moments about mean / Central moments
③ Moments about origin

* Relation between moments about mean in terms of moments about origin :-

We have

$$\mu_r = \frac{1}{N} \sum_i (x_i - \bar{x})^r f_i$$

$$= \frac{1}{N} \sum f_i (x_i + A - A - \bar{x})^r$$

$$\mu_r = \frac{1}{N} \sum f_i (d_i + A - \bar{x})^r \quad \text{—— ①}$$

Where $d_i = x_i - A$

We know that

$$\mu_r' = \frac{1}{N} \Sigma f_i d_i^r$$

where $d_i = x_i - A$

then $$\bar{x} = A + \frac{1}{N} \Sigma f_i d_i = A + \mu_1' \quad -(2)$$

putting (2) in (1)

$$\mu_r = \frac{1}{N} \Sigma f_i (d_i - \mu_1')^r$$

$$= \frac{1}{N} \Sigma f_i \left[ d_i^r - {}^rC_1 d_i^{r-1} \mu_1' + {}^rC_2 d_i^{r-2} \mu_1'^2 + \dots + (-1)^r \mu_1'^r \right]$$

$$\mu_r = \mu_r' - {}^rC_1 \mu_{r-1}' \mu_1' + \dots + (-1)^r \mu_1'^r \quad -(3)$$

On putting $r = 2, 3, 4$ in eq. (3)

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2' \mu_1' + 2\mu_1'^3$$

$$\mu_4 = \mu_4' - 4\mu_3' \mu_1' + 6\mu_2' \mu_1'^2 - 3\mu_1'^4$$

* Effect of change of origin and scale on moments :-

Let $u = \frac{x-A}{h}$ so that $x = A + hu$,

$$\bar{x} = A + h\bar{u} \quad \text{and} \quad (x - \bar{x}) = h(u - \bar{u})$$

Now the $r^{th}$ moment of $x$ about any point $x = A$ is

$$\mu'_r = \frac{1}{N} \Sigma f_i (x_i - A)^r$$

$$= \frac{1}{N} \Sigma f_i (hu_i)^r$$

$$\mu'_r = \frac{1}{N} \cdot h^r \cdot \Sigma f_i u_i^r \quad \text{— (1)}$$

and the $r^{th}$ moment about mean is

$$\mu_r = \frac{1}{N} \Sigma f_i (x_i - \bar{x})^r$$

$$\mu_r = \frac{1}{N} \cdot h^r \cdot \Sigma f_i (u_i - \bar{u})^r \quad \text{— (2)}$$

Thus you can say that moments are not affected by change of origin but they are affected by change of scale.

Q.③

Ans. Given no. of ~~ob~~ students ($n$) = 10

rank correlation coefficient ($\rho_i$) = 0.5

Find corrected rank correlation coefficient ($\rho_c$) = ?

So

$$\rho_i = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$0.5 = 1 - \frac{6\sum d_i^2}{10(100-1)}$$

$$\sum d_i^2 = \frac{495}{6} = 82.5$$

Now corrected square of deviations

$$\sum d_c^2 = \sum d_i^2 - (3)^2 + (7)^2$$
$$= 82.5 - 9 + 49$$
$$\sum d_c^2 = 122.5$$

So the corrected rank correlation coefficient

$$\rho_c = 1 - \frac{6\sum d_c^2}{n(n^2-1)}$$

$$= 1 - \frac{6(122.5)}{990} = 0.258$$

$$\boxed{\rho_c = 0.258}$$

Q.④

Ans.

Q. (4)

Ans. Given $\gamma_{12} = 0.59$, $\gamma_{13} = 0.46$, $\gamma_{23} = 0.77$

Find $\gamma_{12.3} = ?$, $\gamma_{23.1} = ?$, $\gamma_{31.2} = ?$, $R_{1.23} = ?$

Calculations $\gamma_{12}^2 = 0.3481$, $\gamma_{13}^2 = 0.2116$, $\gamma_{23}^2 = 0.5929$

$$\gamma_{12.3} = \frac{\gamma_{12} - \gamma_{23}\gamma_{13}}{\sqrt{(1-\gamma_{23}^2)(1-\gamma_{13}^2)}}$$

$$= \frac{0.59 - (0.77)(0.46)}{\sqrt{(1-0.5929)(1-0.2116)}}$$

$$= \frac{0.2358}{0.5665}$$

$$\boxed{\gamma_{12.3} = 0.4162}$$

$$\gamma_{23.1} = \frac{\gamma_{23} - \gamma_{12}\gamma_{13}}{\sqrt{(1-\gamma_{13}^2)(1-\gamma_{12}^2)}}$$

$$= \frac{0.77 - (0.59)(0.46)}{\sqrt{(1-0.2116)(1-0.3481)}}$$

$$= \frac{0.4986}{0.7169}$$

$$\boxed{\gamma_{23.1} = 0.6955}$$

$$r_{31.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.46 - (0.59)(0.77)}{\sqrt{(1 - 0.3481)(1 - 0.5929)}}$$

$$= \frac{0.0057}{0.5152}$$

$$\boxed{r_{31.2} = 0.01106}$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2 r_{12} r_{13} r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.3481 + 0.2116 - 2(0.59)(0.46)(0.77)}{1 - 0.5929}$$

$$= \frac{0.5597 - 0.4180}{0.4071}$$

$$= \frac{0.1417}{0.4071}$$

$$R_{1.23}^2 = 0.34807$$

$$\boxed{R_{1.23} = 0.58997}$$

or

$$\boxed{R_{1.23} \approx 0.59}$$

Q.⑤

Ans. (i) Line of regression :-

It's a line that best describes the behaviour of a data-set. It's a line that best fits the trend of given data. It describes the relationship of a dependent variable (Y) with an independent variable (X).

(ii) Regression Coefficient :-

They are estimates of the unknown population parameters and describe the relationship between a predictor (independent) variable and the response variable.

The sign of each coefficient indicates the direction of the relationship between a predictor and a response variable.

A positive sign indicates that as predictor variable increases, response variable also increases.

A negative sign indicates that as predictor variable increases, response variable decreases.

Let regression equation

$$y = 3x + 5$$

3 is the coefficient of regression
$x$ is predictor variable
5 is constant

* Regression Equations :-

The standard form of the regression equation of variable X on Y is

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

where $b_{xy}$ is the regression coefficient

$$b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right)$$

$r$ = correlation coefficient of X and Y
$\sigma_x, \sigma_y$ = standard deviation of X and Y respectively

regression equation of Y on X —

$$(y - \bar{y}) = b_{yx}(x - \bar{x})$$

where $b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right)$

* Coefficient of correlation is the geometric mean of regression coefficients :-

We know that ~~bxy~~ $b_{xy} = r\left(\frac{\sigma_x}{\sigma_y}\right)$ and $b_{yx} = r\left(\frac{\sigma_y}{\sigma_x}\right)$

$$\therefore \quad b_{xy} \cdot b_{yx} = r\left(\frac{\sigma_x}{\sigma_y}\right) \cdot r\left(\frac{\sigma_y}{\sigma_x}\right)$$

$$\Rightarrow \quad b_{xy} \cdot b_{yx} = r^2$$

$$\Rightarrow \quad r = \sqrt{b_{xy} \cdot b_{yx}}$$

Hence, proved.