Descriptive Statistics

Date 11/1/25

Types of Descriptive Statistics

- **Distribution** → Measures frequency of each value in a data set
- **Central Tendency** → Mean, Median, Mode
- **Variability** → standard deviation, Range, Variance

* Data/ statistical Variable -
A collection of actual observations or scores in a survey or an experiment.
Any statistical analysis is performed on data.

*

Qualitative Data	Quantitative data
1) Deals with descriptions.	1) Deals with numbers
2) Data can be observed but not measured	2) Data which can be measured.
3) colors, textures, smells, tastes, appearance, beauty etc.	3) length, height, area, volume, weight, speed, time, temp., humidity, sound levels, cost, members, ages etc.
4) Qualitative → Quality	4) Quantitative → Quantity.

* frequency Distributions.
~~Represents~~
How many times that observations occurs is called as frequency.

Represents the pattern of how frequently each value of a variable appears in a dataset. It shows the no. of occurrences for each possible value within the dataset.

Range = largest observation − smallest observation

Ungrouped frequency distribution.

Q. Make the frequency distribution table for the ungrouped data given as follows: 10, 20, 15, 25, 30, 10, 15, 10, 25, 20, 15, 10, 30, 25.

⇒ Ungrouped frequency distribution table:

Value	frequency
10	4
15	3
20	2
25	3
30	2

Grouped frequency distribution.

Q. Make the frequency distribution table for the ungrouped data given as follows:

23, 27, 21, 14, 43, 37, 38, 41, 55, 11, 35, 15, 21, 24, 57, 35, 29, 10, 39, 42, 27, 17, 45, 52, 31, 36, 39, 38, 43, 46, 32, 37, 25.

⇒ Min. no is 10, and max. no is 57.

10 include ✓
20 exclude ✗

Class Interval	frequency
10 − 20	5
20 − 30	8
30 − 40	11
40 − 50	6
50 − 60	3

Q. Consider a data set of 26 children of ages 1-6 years.

2, 2, 1, 3, 3, 3, 6, 6, 2, 1, 1, 1, 3, 3, 3, 5, 5, 4, 4, 4, 5, 5, 4, 4.

* ungrouped frequecy distribution:

Age	1	2	3	4	5	6
froquecy	5	3	6	5	4	2

Grouped frequecy distribution:

Age group	1-2	3-4	5-6
frequecy	8	11	6

* Relative frequency Distribution:

Relative frequency = $\frac{\text{frequency of event}}{\text{Total Number of events}}$

This distribution displays the proportion or % of observations in each interval or class.
It is useful for comparing different data sets or for analysing the distribution of data within the set.

Q. find Relative frequency distribution.

Score Range	1-20	21-40	41-60	61-80	81-100
frequency	5	10	20	10	5
Relative frequency	5/50 = 0.10	10/50 = 0.20	20/50 = 0.40	10/50 = 0.20	5/50 = 0.10

0.10 + 0.20 + 0.40 + 0.20 + 0.10 = 1.00

* Cummulative frequency distribution.
It is defined as the sum of all the frequencies in the previous values or intervals up to the current one.

The distributions which represent the frequency distribution using cumulative frequencies are called cumulative frequency distributions.

There are two types of cumulative frequency distributions:
Less than type: We sum all the frequencies before the current interval.
More than type: We sum all the frequencies after the current interval.

Q. The table below gives the values of runs scored by Virat Kohli in the last 25 T-20 matches. Represent the data in the form of less-than-type cumulative frequency distribution.

45	34	50	75	22
56	63	70	49	33
0	8	14	39	86
92	88	70	56	50
57	45	42	12	39

⇒

Runs	frequency.
0-10	2
10-20	2
20-30	1
30-40	4
40-50	4
50-60	5
60-70	1
70-80	3
80-90	2
90-100	1

Runs scored by Virat Kohli	C.F
less than 10	2
less than 20	4
less than 30	5
less than 40	9
less than 50	13
less than 60	18
less than 70	19
less than 80	22
less than 90	24
less than 100	25

Runs scored by Virat Kohli	Cumulative frequency (CF)
more than 0	25
more than 10	23
more than 20	21
more than 30	20
more than 40	16
more than 50	12
more than 60	7
more than 70	6
more than 80	3
more than 90	1

Measures of central Tendency

- → Mean = sum of all obs. / no. of obs. (denoted by $\bar{x}$)
- → Mode
- → Median

NOTE:

If the data is grouped data then each x_i is the class representative of that class i.e $x_i = \frac{x_i^l + x_i^h}{2}$ where, x_i^l = lower bound for i^{th} class.

x_i^h = upper bound

* Mean (Arithmetic Average)

for ungrouped data:

$$\text{Mean} = \frac{\text{Sum of all observations}}{\text{number of observations}}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

for grouped data: $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

where, f_i = frequencies of class i and x_i = midpoint of interval.

case of Median, to find the median class if $\frac{N}{2}$ = CF then the respective class interval of CF is the median class. If $\frac{N}{2}$ is not equally equal to CF then find nearly CF for the value of N/2 and i.e the Median class.

Median = The middle value of the dataset when arranged in ascending order.

for ungrouped data,

firstly, arranged the data in ascending order.

① If n is odd, Median = $\frac{(n+1)^{th}}{2}$ observation.

② If n is even, Median = $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$ observation

for grouped data,

Step 1: Construct the cumulative frequency distribution.

Step 2: find the median class. Median class is the class in which the value of $\frac{N}{2}$ falls in cumulative frequency distribution.

Step3: find the median by using the following formula.

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

where, N = Total frequency

L = lower limit of median class

f = frequency of median class

C.f = cumulative frequency of the class before the median class.

h = class width

finding the modal class, if more than one class interval have largest frequency f_i then the modal class will be the class interval which Lower limit

10-20	20
20-30	15
30-40	20

so, 30-40 is our modal class and 30 is Lower

→ If there is one mode, the data is unimodal
→ If there are two modes, the data is bimodal
→ If there are more than two modes, the data is multimodal
→ If no value repeats, there is no mode

* Mode:

Value that occurs the most frequently in data set.

for ungrouped data:

Mode = number that occurs the highest number of times.

for grouped data:

Step 1: find the modal class. Modal class is the class with maximum frequency.

$$\text{Mode} = L + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$$

where, L = lower limit of the modal class

h = class width

f_m = frequency of the modal class

f_2 = frequency of the class succeeding the modal class

f_1 = frequency of the class preceding the modal class.

* Variance:

let $x_1, x_2, \ldots x_n$ are 'n' observations of a variable x.

Then the variance of this variable,

for ungrouped,

$$\text{Var}(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

for grouped data,

$$\text{Var}(X) = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i}$$

standard deviation (σ)

The square root of variance, which gives the spread of data points.

$$S.D = \sqrt{\frac{\sum f_i (x_i - \mu)^2}{n}} \qquad (\mu = \bar{x})$$

Coefficient of variation:

A standardized measure of dispersion, calculated as the ratio of the standard deviation to the mean.

$$C.V = \frac{\text{Standard deviation}}{\mu} \times 100$$

find Variance, standard deviation and coefficient of variation.

class interval	frequency, f_i
0-10	4
10-20	6
20-30	5
30-40	3
40-50	2

find mean, mode, median, variance, standard deviation, coefficient of variation.

class interval	frequency (f_i)
10-20	3
20-30	5
30-40	7
40-50	2

midpoint (x_i) = $\frac{\text{lower limit} + \text{Upper limit}}{2}$

class interval	f_i	x_i	CF	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
0-10	4	5	4	272.25	1089
10-20	6	15	10	42.25	253.5
20-30	5	25	15	12.25	61.25
30-40	3	35	18	182.25	546.75
40-50	2	45	20	552.25	1104.5
	20				3055

for Mean, $\bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{20+90+125+105+90}{20} = \frac{21.5}{}$.

for Median,

$$\text{Median} = L + \left(\frac{\frac{N}{2} - C.F}{f}\right) \times h$$

$L = 10$, $N = 20 \Rightarrow N/2 = 10$, $C.F = 4$, $h = 10$

$f = 6$.

$$\text{Median} = 10 + \left(\frac{10-4}{6}\right) \times 10 = \underline{20}.$$

for Mode, $\text{Mode} = L + \left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right) \times h$

$L = 10$, $f_m = 6$, $f_1 = 4$, $f_2 = 5$, $h = 10$

$$\text{Mode} = 10 + \left(\frac{6-4}{12-4-5}\right) \times 10 = \frac{50}{3} = \underline{16.67}.$$

$$\text{Var}(x) = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i} = \frac{3055}{20} = \underline{152.75}$$

Standard deviation, $S.D = \sqrt{var} = 12.35$

coefficient of variation $= \frac{S.D}{\bar{X}} \times 100 = \frac{12.35}{21.5} \times 100 = 57.44$

class interval	f_i	X_i	CF	$(X_i - \bar{X})^2$	$f_i(X_i - \bar{X})^2$
10-20	3	15	3	216.09	648.27
→ 20-30	5	25	8	22.09	110.45
→ 30-40	7	35	15	28.09	196.63
40-50	2	45	17	234.09	468.18
	17				1423.53.

for mean, $\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{45 + 125 + 245 + 90}{17} = 29.70$

for median, $\text{median} = L + \left(\frac{\frac{N}{2} - CF}{f}\right) \times h$

$L = 20$, $N = 17$, $CF = 3$, $f = 5$, $h = 10$.

$\text{median} = 20 + \left(\frac{\frac{17}{2} - 3}{5}\right) \times 10 = 31$

for mode, $\text{mode} = L + \left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right) \times h$

Modal class = 30-40.

$L = 30$, $f_m = 7$, $f_1 = 5$, $f_2 = 2$, $h = 10$.

$\text{mode} = 30 + \left(\frac{7 - 5}{14 - 5 - 2}\right) \times 10 = 32.85$

for variance, $\text{Var}(x) = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{n} f_i} = \frac{1423.53}{17} = 83.73$

Standard Deviation, $S.D = \sqrt{\text{Var.}} = \sqrt{83.73} = 9.150$

coeff of variation, $C.V = \frac{S.D}{\bar{x}} \times 100 = \frac{9.150}{29.70} \times 100 = 30.80\%$

∴ Mean = 29.70 Variance = 83.73

Median = 31 S.D, σ = 9.150

Mode = 32.85 C.V = 30.80%

Calculate Pearson's skewness coefficient for a dataset of exam scores : 85, 88, 92, 94, 96, 98, 100, 100, 100, 100.

Mean:

$$\text{Mean} = \frac{85+88+92+94+96+98+100+100+100+100}{10}$$

$$= 95.3$$

Median: Write in ascending order first.

85, 88, 92, 94, 96, 98, 100, 100, 100, 100.

Total 10 terms → that are even.

so median is the average of 5th and 6th values when sorted in ascending order:

$$\text{Median} = \frac{96+98}{2} = 97$$

Standard deviation(σ), $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$

$$= \frac{(85-95.3)^2 + (88-95.3)^2 + \ldots + (100-95.3)^2}{10}$$

$\sigma^2 = 26.81$.

S.D $= \sigma = \sqrt{26.81} = 5$

Mode: 100 (100 is most frequently occuring value in the data)

S_k with respect to mean and median.

$S_k = -1.02$ (slight negative skewness in the distribution of exam scores)

S_k with respect to mean and mode

$S_k = -0.94$

This means that the tail of the distribution is slightly longer on the left side, and most of the scores are concentrated on the right side of the mean.

Bowley's Measure:

It is especially useful when dealing with data that may not follow a normal distribution or when a robust measure of skewness is required.

$$B = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

Q_1 is the first quartile (25^{th} %-tile)

Q_2 is the second quartile (50^{th} %-tile or median)

Q_3 is the third quartile (75^{th} %-tile)

* Kurtosis.

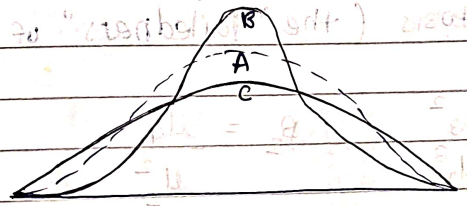
Measure is denoted by β_2

Lepto kurtic - A distribution with heavy tails and a sharp peak ($\beta_2 > 3$). curve is peaked.

Platy kurtic - A distribution with light tails and a flatter peak. ($\beta_2 < 3$). Curve is flat topped.

Meso kurtic - A normal distribution ($\beta_2 = 3$) Curve is Normal.

kurtosis - measures degree of peakedness of the distribution.



A: Meso kurtic

B: Lepto kurtic

C: Platy kurtic

Moments:

Moments are statistical measures that give certain characteristics of the distribution.

* formulae to calculate moments about the Mean:

first Moment (about the Mean)

$\mu_1 = 0$ (since it is always zero)

second moment (about the Mean)

μ_2 (variance) , $\mu_2 = \frac{\Sigma f (x - \bar{x})^2}{N}$

Third Moment (about the Mean)

μ_3 (skewness) , $\mu_3 = \frac{\Sigma f (x - \bar{x})^3}{N}$

fourth Moment (about the Mean)

μ_4 (kurtosis), $\mu_4 = \frac{\sum f(x-\bar{x})^4}{N}$

The coefficient of kurtosis:

To calculate β_1 (Beta 1) and β_2 (Beta 2) for grouped data using a tabular form, we need to first understand what these measures represents.

β_1 used as measures of skewness of distribution

β_2 measures kurtosis (the "tailedness" of the distribution)

formula: $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$

where, μ_2 is the second central moment (variance)

μ_3 is the third central moment (used to measure skewness)

μ_4 is the fourth central moment (used to measure kurtosis)

steps to calculate β_1 and β_2

calculate the Mean.

calculate the deviations $(x-\bar{x})$

calculate the second, third and fourth central moments:

μ_2, μ_3 and μ_4.

calculate β_1 and β_2.

NOTE:

① If $\beta_2 = 3$, the data is normal or mesokurtic

② If $\beta_2 > 3$, the data is peaked or leptokurtic

③ If $\beta_2 < 3$, the data is flat topped or platykurtic

Q.

class Interval	frequency (f)	x_i	$f_i x_i$
0-10	4	5	20
10-20	6	15	90
20-30	8	25	200
30-40	5	35	175
40-50	3	45	135

⇒ calculate the Mean :

$$\bar{X} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{20+90+200+175+135}{26} = \frac{620}{26}$$

$$\bar{X} = 23.8 \qquad N = 26$$

class Interval	f	x_i	$\bar{X} - x_i$	$x_i - \bar{X}$	$(x_i - \bar{X})^2$
0-10	4	5	18.8	-18.8	353.44
10-20	6	15	8.8	-8.8	77.44
20-30	8	25	-1.2	1.2	1.44
30-40	5	35	-11.2	11.2	125.44
40-50	3	45	-21.2	21.2	449.44

$$\mu_2 = \frac{\Sigma f (x - \bar{x})^2}{N}$$

$$\mu_2 = \frac{1413.76 + 465.04 + 11.52 + 627.20 + 1348.32}{26}$$

$$\mu_2 = \frac{3865.84}{26} = 148.68$$

$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
-6644.672	12491g.834
-681.472	5996.9536
1.728	2.0736
1404.928	15735.1936
9528.128	201996.314

$$\mu_2 = \frac{-26578.68 + (-4088.82) + 13.76 + 7024.6 + 28584.36}{26}$$

$$\mu_3 = 190.5853$$

$$\mu_4 = \frac{499679.32 + 35981.7 + 16.58 + 78675.95 + 605988.}{26}$$

$$\mu_4 = 46936.24$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{36320.73}{3286681.78} = 0.01105$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 2.12$$

So, $\beta_2 = 2.12$, indicating that the distribution is platykurtic (flatter than a normal distribution, since β_2 is less than 3).

* Population:

The complete set of all possible observations or measuremen that could be made. It represents the entirety of individuals or instances about which you want to make interferences.

eg, All B.Tech students in a university.

* Sample:
A subset of the population selected for analysis. It is used to draw conclusions about the population.
Ex. A group of 100 randomly selected B.Tech students.

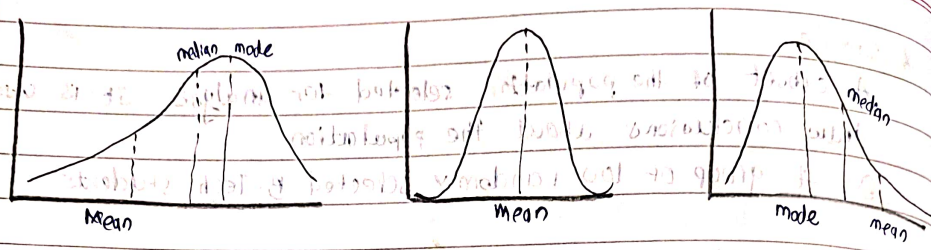
NOTE: Size of sample, denoted by 'n'
Size of population, denoted by 'N'.

VARIABLE

Numerical (quantitative)		Categorical (qualitative)	
Discrete	Continuous	Nominal	Ordinal
Le numerical value	can take any	Unordered, categories	ordered
ically counts	value within a	which are mutually	categories
. of visits to	range	exclusive.	which are
enfist	ex. height in 'cm'	eg, male/female	mutually
			exclusive.
			eg, minimal/severe

* Shape of Data
* Skewness:
Measures asymmetry of the distribution of values in a data set. Skewness helps understand the underlying distribution of data, which is crucial for decision-making, risk assessment, and predicting future trends.

Relative to the mean:
If the plot is skewed to the left, it is negatively skewed.
If the plot is skewed to the right, it is positively skewed.
If the plot is roughly symmetric, it has no skewness.



mean < median < mode

symmetrical data

mode < median < r

mean = median = mode

← negative direction.

Positive direction →

S_K

→ Skewness coefficient. (Pearson's first coefficient of skewness):
This is a numerical measure of skewness, which determines the skewness when mean and mode are not equal.

1) If mean > mode, the skewness will consist positive value.
2) If mean < mode, the skewness will be a negative value.
3) If mean = mode, the skewness will be zero.

→ Coefficient of skewness as per karl pearson's Measure.

1) wrt Mean and Median:

$$S_K = \frac{3(\text{mean} - \text{median})}{\sigma}$$

2) wrt mean and mode:

$$S_K = \frac{\text{Mean} - \text{mode}}{\sigma}$$

1) If $S_K = 0$, it indicates a perfectly symmetric distribution where the data is evenly balanced on both sides of the mean.
2) If $S_K > 0$, it suggests a positively skewed distribution where the tail on the right side is longer or fatter, and most data points are concentrated on the left side of the mean.
3) If $S_K < 0$, it indicates a negatively skewed distribution where the tail on the left side is longer or fatter, and most data points are concentrated on the right side of the mean.

Bowley's Measure:

Bowley's skewness coefficient, is a statistical measure used to assess the skewness or asymmetry in a probability distribution.

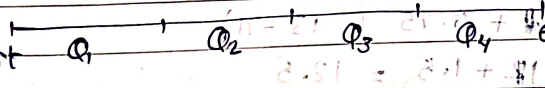
Coefficient of Bowley's Measure.

- If $B=0$, the distribution is perfectly symmetric about the mean (no skewness)
- If $B<0$, the distribution is negatively skewed (left-skewed), meaning the tail on the left side of the distribution is longer or heavier.
- If $B>0$, the distribution is positively skewed (right-skewed), indicating that the tail on the right side of the distribution is longer or heavier.

Quartiles, Deciles, Percentiles.

Partition values are statistical measures that divide a dataset into equal parts.

Quartiles - divide data into equal four parts.



Q_2 = median.

(N: no. of observations)

Quartiles:

Q_1 = first quartile $\quad Q_1 = \left(\frac{N+1}{4}\right)^{th}$ term.

Q_2 = second quartile = median. $\quad Q_2 = \left(\frac{2(N+1)}{4}\right)^{th}$ term

Q_3 = Third quartile $\quad Q_3 = \left(\frac{3(N+1)}{4}\right)^{th}$ term.

for deciles, $D_k = L + \left(\frac{\frac{kN}{10} - cf}{f}\right) \times h$

for centiles, $P_k = L + \left(\frac{\frac{kN}{100} - cf}{f}\right) \times h$

NOTE

Above formulae are for grouped as well as for ungrouped data.

Calculate the lower and upper quartiles of the following weights in the family : 25, 17, 32, 11, 40, 35, 13, 5 and 46.

First, organize the given data in ascending order.

5, 11, 13, 17, 25, 32, 35, 40, 46.

for $N = 9$.

$Q_1 = \frac{(N+1)^{th}}{4}$ term $= \left(\frac{10}{4}\right)^{th}$ term $= 2.5^{th}$ term.

$= \frac{2^{nd} \text{ term} + 3^{rd} \text{ term}}{2} = \frac{11+13}{2} = \frac{12}{}$.

$Q_1 = 12$.

NOTE ① suppose $Q_1 = 2^{nd}$ term $\Rightarrow Q_1 = 11$

NOTE ② suppose $Q_1 = 2.75$

$= 2^{nd} \text{ term} + 0.75 (3^{rd} \text{ term} - 2^{nd} \text{ term})$

$= 11 + 0.75 (13 - 11)$

$= 11 + 1.5 = 12.5$

In the worst case, you take $Q_1 = 13$.

as it is near to 3rd term, as it is greater than 2nd term. so, ans will be 13 in the worst case.

for

Q_2 = median = 25

for Q_3, $Q_3 = \left[\frac{3(N+1)}{4}\right]^{th}$ term $= \left[\frac{3(9+1)}{4}\right]^{th}$ term

$= \left(\frac{3 \times 10}{4}\right)^{th} = (7.5)^{th}$ term

If given data is grouped and in intervals then use formula ⇒ $Q_k = L + \left(\frac{\frac{kN}{4} - Cf}{f}\right) \times h$ for Quartile

$Q_3 = (7.5)^{th}$ term

$= \frac{8^{th} \text{ term} + 7^{th} \text{ term}}{2} = \frac{35+40}{2} = 37.5$

$Q_3 = 37.5$

Q2) Calculate Q_1 and Q_3 for the data related to the age in years of 99 members in a housing society.

Age (in years)	no. of members.	cumulative freq.
10	20	20
18	5	25
25	10	35
35	30	65
40	20	85
45	14	99.

→ first find cumulative frequency.

N (all add the frequency) = 99.

$Q_1 = \left(\frac{N+1}{4}\right)^{th}$ term $= \left(\frac{100}{4}\right)^{th}$ term $= 25^{th}$ term

Now, the 25^{th} term falls under CF of 25 and the age against this CF value is 18.

NOTE: If we have to find exact value of Q_1 for grouped data. Then in median formula N/2 replaced by N/4.

$\text{Median} = L + \frac{(N/2 - CF)}{f} \times h$

$Q_1 = L + \frac{(N/4 - CF)}{f} \times h$

for $Q_3 = \left[\frac{3(N+1)}{4}\right]^{th}$ term. $N = 99$.

$Q_3 = \left(\frac{3 \times 100}{4}\right)^{th}$ term $= 75^{th}$ term

$\therefore$ 75^{th} term falls under cf of 85 and the age against this cf value is 40

Q. calculate Bowley's Measure of skewness for the following dataset representing the ages of a group of people in a sample: 20, 24, 28, 32, 35, 40, 42, 45, 50.

$\Rightarrow$ arrange in ascending order:

20, 24, 28, 32, 35, 40, 42, 45, 50. $N = 9$

Q_2 (median)

$\Rightarrow Q_2 = 35$. (the middle value)

for $Q_1 = \frac{(N+1)^{th}}{4}$ term $= \left(\frac{10}{4}\right)^{th}$ term $= 2.5^{th}$ term

$Q_1 = \frac{2^{nd} \text{ term} + 3^{rd} \text{ term}}{2} = \frac{24 + 28}{2} = \frac{52}{2}$

$Q_1 = 26$

for $Q_3 = \left(\frac{3(N+1)}{4}\right)^{th}$ term $= \left[\frac{3(10)}{4}\right]^{th}$ term $= 7.5^{th}$ term

$Q_3 = \frac{7^{th} \text{ term} + 8^{th} \text{ term}}{2} = \frac{42 + 45}{2} = \frac{87}{2}$

$Q_3 = 43.5$

$B = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1} = \frac{26 + 43.5 - 2(35)}{43.5 - 26}$

$$B = \frac{26 + 43.5 - 70}{43.5 - 26} = \frac{69.5 - 70}{17.5} = \frac{-0.5}{17.5} = \frac{-5}{175}$$

$B = -0.028$

since, B is negative ($B < 0$) the distribution is negatively skewed (left-skewed). This means that the tail of the distribution is longer on the left side, indicating that there may be outliers or high values on the right side of the data.

→ Deciles:

The deciles involve dividing a dataset into ten equal parts based on numerical values. There are therefore nine deciles altogether. Deciles are represented as follows:

$D_1, D_2, D_3, D_4 \ldots$

A decile is used to group big data sets in descriptive statistics either from highest to lowest values or vice versa.

$$D_1 = \left(\frac{N+1}{10}\right)^{th} \text{ term}, \quad D_2 = \left[\frac{2(N+1)}{10}\right]^{th} \text{ term} \ldots \text{ so on.}$$

$$D_9 = \left[\frac{9(N+1)}{10}\right]^{th} \text{ term.}$$

Q Calculate the D_1, D_5 from the following weights in a family: 25, 17, 32, 11, 40, 35, 13, 5 and 46

→ In ascending order: 5, 11, 13, 17, 25, 32, 35, 40, 46.

Here, $N = 9$

$$D_1 = \left(\frac{N+1}{10}\right)^{th} \text{ term} = \left(\frac{10}{10}\right)^{th} \text{ term} = 5.$$

$$D_5 = \left(\frac{5(N+1)}{10}\right)^{th} \text{ term} = \left(\frac{5 \times 10}{10}\right)^{th} \text{ term} = 25$$

Percentiles.

Centiles is another form of percentiles.

Any given observation is essentially divided into a total of 100 equal parts by a centile or percentile.

represents as $P_1, P_2, P_3, P_4 \ldots P_{99}$.

P_1 is a typical value of peaks for which 1/100 of any given data is either less than P_1 or equal to P_1.

$$P_1 = \left[\frac{N+1}{100}\right]^{th} \text{ term}, \quad P_2 = \left[\frac{2(N+1)}{100}\right]^{th} \text{ term} \ldots \text{ so on.}$$

$$P_{99} = \left[\frac{99(N+1)}{100}\right]^{th} \text{ term}$$

N: total no. of observations.

P_1 is first percentile, P_2 is second percentile and so on.

calculate P_{10} and P_{75} for the data related to the age (in years) of 99 members in a housing society.

Age (in years)	No. of members	CF
10	20	20
18	5	25
25	10	35
35	30	65
40	20	85
45	14	99

first find cumulative frequency. $N = 99$

$$P_{10} = \left(\frac{10(N+1)}{100}\right)^{th} \text{ term.}$$

$$= \left[\frac{10(99+1)}{100}\right]^{th} = 10^{th} \text{ term}$$

Now, the 10th term falls under the cumulative frequency of 20 and the age against this CF value is 10.

P_{10} = 10 years.

P_{75} = 75th term.

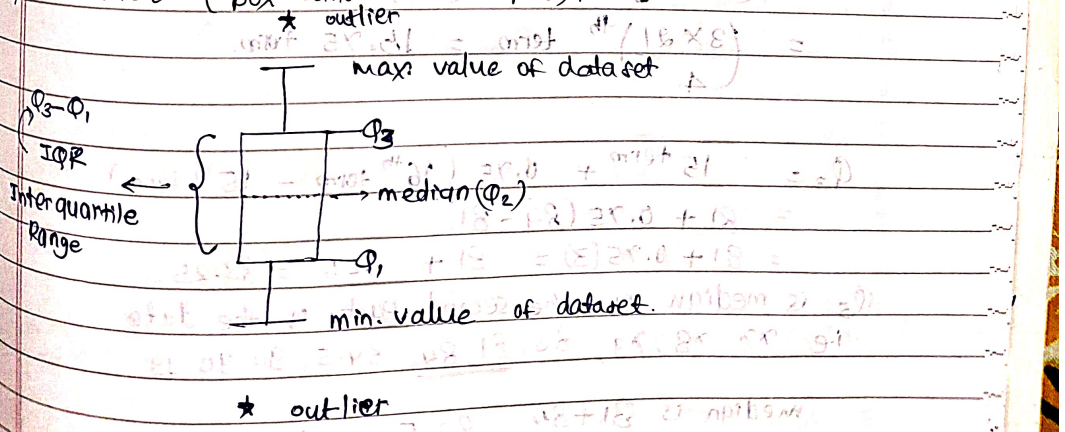
Now, the 75th term falls under the CF of 85 and the age against CF value is 40.

Data Visualization:

1) Histogram.

A histogram is a graphical representation of the distribution of numerical data. It is similar to a bar chart but specifically used for quantitative data that is divided into ranges (also called bins or intervals). Histograms are essential for visualizing the frequency of data points in each range, helping to understand the shape and spread of the data distribution.

2) Box Plot (Box-and-Whisker Plot).



of data set has the wider spread for
IQR for the first data set is greater than the IQR of the second set
- This means that there more variability in the middle 50% of the first data set

Test scores for a college statistics class held during the day are: 99, 56, 78, 55.5, 32, 90, 80, 81, 56, 59, 45, 77, 84.5, 84, 70, 72, 68, 32, 79, 90

Find the smallest and largest values, the median, and the first and third quartile for the day class.

Arrange the data in ascending order.

32, 32, 45, 55.5, 56, 56, 59, 68, 70, 72, 77, 78, 79, 80, 81, 84, 84.5, 90, 90, 99.

min. value = 32.

max. value = 99. N = 20.

median = $\frac{72+77}{2} = \frac{149}{2} = 74.5$

$Q_1 = \left(\frac{N+1}{4}\right)^{th}$ observation $= \left(\frac{21}{4}\right)^{th}$ observation.

$= (5.25)^{th}$ term.

$= \frac{5^{th}\text{ term} + 6^{th}\text{ term}}{2} = \frac{56+56}{2} = \frac{112}{2} = 56$

$Q_3 = \left[\frac{3(N+1)}{4}\right]^{th}$ term $= \left[\frac{3(20+1)}{4}\right]^{th}$ term

$= \left(\frac{3\times 21}{4}\right)^{th}$ term $= 15.75$ term.

$Q_3 = 15^{th}\text{ term} + 0.75(16^{th}\text{ term} - 15^{th}\text{ term})$

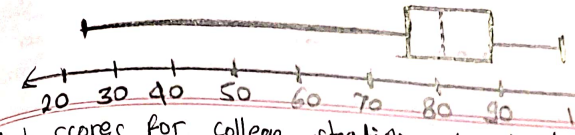
$= 81 + 0.75(84 - 81)$

$= 81 + 0.75(3) = 81 + 2.25 = 83.25.$

Q_3 is median of the second part of the data.

i.e 77, 78, 79, 80, 81, 84, 84.5, 90, 90, 99

median is $\frac{81+84}{2} = 82.5 \rightarrow Q_3.$



Test scores for college statistics class held during the evening are:
98, 78, 68, 83, 81, 89, 88, 76, 65, 45, 98, 90, 80, 84.5,
85, 79, 78, 98, 90, 79, 81, 25.5

Find the smallest and largest values, the median, and the first and third quartile for the night class.

Arrange the data in ascending order.

25.5, 45, 65, 68, 76, 78, 78, 79, 79, 80, 81, 81, 83, 84.5,
85, 88, 89, 90, 90, 98, 98, 98

min value = 25.5.

max. value = 98.

$N = 22$

$$\text{median} = \frac{81+81}{2} = \frac{162}{2} = 81$$

$$Q_1 = \left(\frac{N+1}{4}\right)^{th} \text{term} = \left(\frac{22+1}{4}\right)^{th} \text{term} = \left(\frac{23}{4}\right)^{th} \text{term}$$

$$= 5.75^{th} \text{ term} = 5^{th} \text{ term} + 0.75\left(6^{th} - 5^{th}\right) \text{term}$$

$$= 76 + 0.75(78 - 76)$$

$$= 76 + 0.75(2)$$

$$= 76 + 1.50 = 77.5$$

$Q_1 = 77.5$

$$Q_3 = \left[3\left(\frac{N+1}{4}\right)\right]^{th} \text{term} = \left[3\left(\frac{22+1}{4}\right)\right]^{th} \text{term}$$

$$= \left(\frac{3 \times 23}{4}\right)^{th} \text{term} = 17.25^{th} \text{ term}$$

Q_1 is the median of first half data & Q_3 is the median of second half data.

For Q_1, 25.5, 45, 65, 68, 76, 78, 78, 79, 79, 80, 81 $Q_1 = 78$

For Q_2, 81, 83, 84.5, 85, 88, 89, 90, 90, 98, 98, 98 $Q_3 = 89$.

Data Visualization.

1) Histogram:

A histogram is a graphical representation of the distributio of numerical data. It is similar to a bar chart but specifically used for quantitative data that is divided into ranges (also called bins or intervals). Histograms are essential for visua the frequency of data points in each range, helping to understand the shape and spread of the data distribution.

Bins (intervals) – A bin is a continuous range of values within which data points are grouped together. Each bin represen a specific interval of values, and the height of the bar for each bin shows the no. of data points (or frequency) that fall within that range.

Ex. If you're analyzing the test scores of students (0-100), bins could be set at intervals of 10 (0-10, 11-20, 21-30 etc)

Frequency distribution: Choosing the no. of bins approx. equal to the square root of the no. of observ. often works well in practi

→ Equal units along the horizontal axis (x-axis, abscissa) reflect the various class intervals of the frequency distribution.

→ Equal units along the vertical axis (the y-axis, ordinate) reflect increases in frequency. (the units along the vertical axis do not have to be the same width as those along the horizont axis)

→ use wiggly lines to highlights breaks in scale

→ Choosing Bin width:

Too few Bins – If the ~~few bins~~ bin width is too large, the histogram may not show imp. details about the distribution

Too many Bins = If the bin width is too small, the histogram can become overly detailed and noisy.

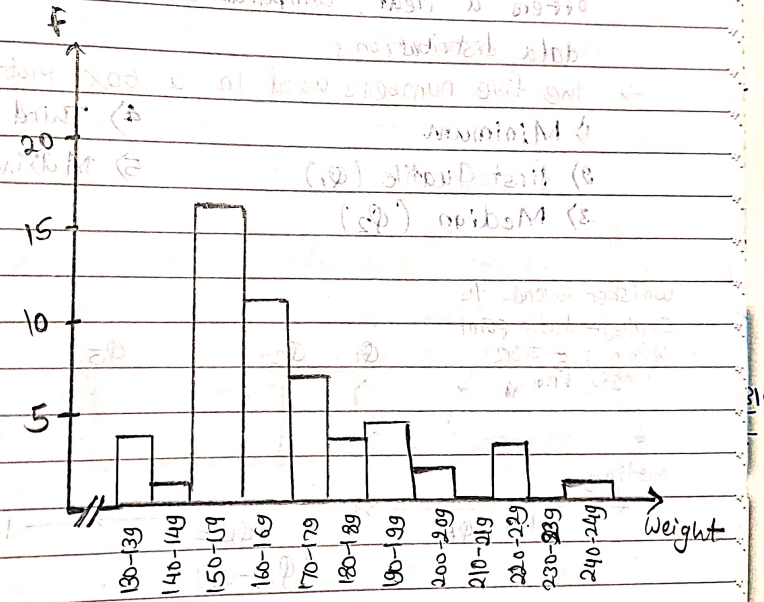
→ Label the bin (class interval) boundaries on a horizontal scale.

→ Mark and label the vertical scale with the frequencies or the relative frequencies.

Q. Draw the histograms for following details.

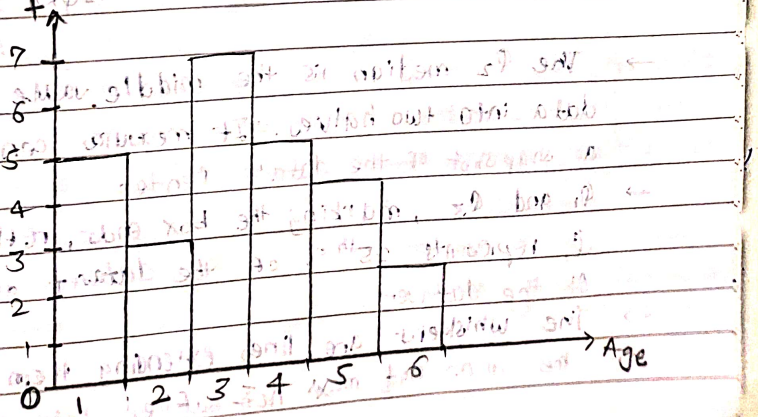
1)

weight	f
130-139	3
140-149	1
150-159	17
160-169	12
170-179	7
180-189	3
190-199	4
200-209	2
210-219	0
220-229	3
230-239	0
240-249	1



2)

Age	f
1	5
2	3
3	7
4	5
5	4
6	2

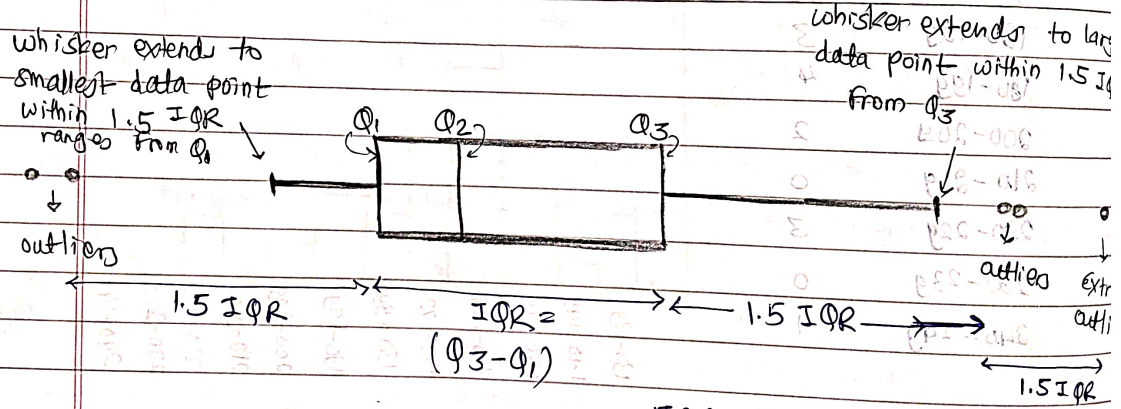


2) Box Plot (Box-and-whisker plot)

The Box Plot is a graphical representation of a dataset's five-number summary: minimum, first quartile (25^{th} %), median (50^{th} %), third quartile (75^{th} %), and maximum. It's a powerful tool in data analysis b'coz it can clearly highlight the dataset's central tendency, dispersion and skewness. It effectively visualizes outliers. This is particularly useful when comparing multiple datasets, as it offers a clear, comparative visualization of the different data distributions.

→ The five numbers used in a box plot are:

1) Minimum
2) first Quartile (Q_1)
3) Median (Q_2)
4) Third quartile (Q_3)
5) Maximum.



IQR: Interquartile ranges.

→ The Q_2 median is the middle value that separate the data into two halves. It measures central tendency, providing a snapshot of the data's center.

→ Q_1 and Q_3, marking the box ends, reflect the data's dispersion. Q_1 represents 25^{th} % of the dataset and Q_3 represents 75^{th} % of the datasets.

→ The whiskers are lines extending from the box, reaching the min. and max. non-outlier data points

Q_1 → median of first half of the data.
Q_3 → median of second half.

usually, the lower whisker extends from Q_1 to the smallest non-outlier data point and the upper whisker extents from Q_3 to the largest non-outlier data point.

→ The length of the box is the IQR.

$$IQR = Q_3 - Q_1$$

→ The IQR measures the middle 50% of the data, measuring dispersion or spread.

→ Outliers are typically calculated as data points that fall below $(Q_1 - 1.5\,IQR)$ or above $(Q_3 + 1.5\,IQR)$.

These outliers are represented as individually points outside the whiskers in the box plot.

→ A point more than 3 IQR from the box edge is called an extreme outlier.

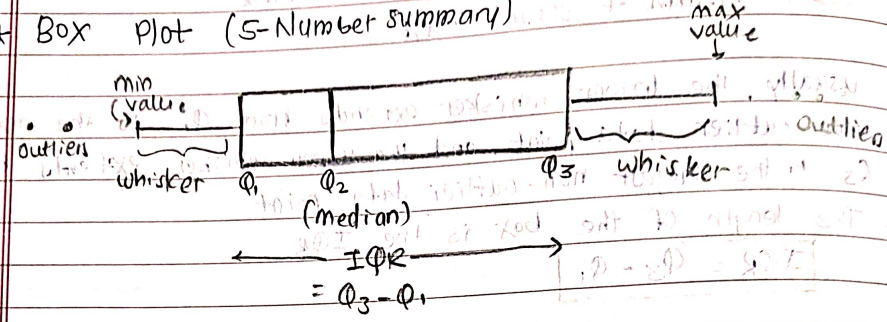
* How to Interpret a Box Plot:

1) length of the Box: The length of the box (betn Q_1 and Q_3) represents the IQR, showing the spread of the middle 50% of the data.

2) Median line: If the median line is closer to Q_1, the data is left skewed (longer whisker on the right). If it's closer to Q_3 the data is right skewed (longer whisker on the left)

3) whiskers: The whiskers represent the range of the data, excluding outliers. longer whiskers indicate more variability.

✱ Box Plot (5-Number summary)



Q. 22, 25, 23, 17, 19, 33, 64, 17, 20, 18 Plot Box Plot.

→ firstly arrange in ascending order.

17, 17, (18), 19, 20 | 22, 23, (25), 33, 64

first part → Q_1 ; second part → Q_3

$Q_2 = \frac{20+22}{2} = 21 \rightarrow Q_2$

$18 \rightarrow Q_1$ and $25 \rightarrow Q_3$.

✱ If two terms are present at center, then it would take average of those two terms for finding Q_1 and Q_3.

$IQR = Q_3 - Q_1 = 25 - 18 = 7$

Now check for outliers

① Higher outlier
$= Q_3 + (1.5 \times IQR)$
$= 25 + (1.5 \times 7)$
$= 35.5$
Any data > 35.5 is higher outlier

② lower outlier
$= Q_1 - (1.5 \times IQR)$
$= 18 - 10.5$
$= 7.5$
Any data < 7.5 is lower outlier.

$Q_1 = 18$; $Q_2 = 21$, $Q_3 = 25$

min = 17; max = 33 outlier = 64

(not 64 bcoz, 64 is outlier
∴ 33 is max.