STATS:

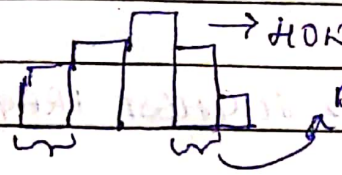
RAW DATA → INFORMATION → KNOWLEDGE → WISDOM

DESCRIPTIVE STATISTICS

FREQUENCY DISTRIBUTION:

RAW DATA → CLASSES & FREQUENCY

HISTOGRAM



ASSUME THERE ARE RANDOM NUMBERS B/W 8 - 23.

THESE NUMBERS ARE CLASSIFIED:

8-11, 11-14, 14-17, 17-20, 20-23 → CLASSES

2, 7, 12, 3, 1 → FREQUENCY

CUMULATIVE DISTRIBUTION = 2+7+12+3+1 = 25

CUMULATIVE DISTRIBUTION CURVE OR OGIVE CURVE:

CENTRAL TENDENCY:

WHERE IS THE DATA AROUND → AVG (LIKE)

1] ARITHMETIC MEAN:

$\bar{x} = \frac{\Sigma x}{n}$ → MEAN

Σx → SUM OF NUMBERS

n → NO OF VALUES / COUNT / SAMPLE SIZE

BUT EXTREME VALUES AFFECT THE MEAN HENCE DO NOT USE IF THE DATA SET HAS EXTREME VALUES.

2] MEDIAN → VALUE @ 50TH PERCENTILE
MIDDLE MOST OBSERVATION AFTER ARRANGING THE DATA SET FROM LOWEST TO HIGHEST.
THIS TO A LARGE EXTENT REDUCES THE EFFECT OF EXTREME VALUE.
EVEN NUMBER OF VALUES → MIDDLE VALUE
ODD NUMBER OF VALUES → AVERAGE OF MIDDLE VALUE

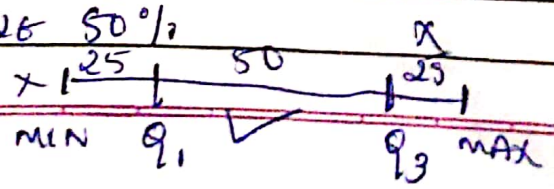
3] MODE
THE VALUE THAT HAS HIGHEST FREQUENCY.
LIKELY OUTPUT.
DRAW BACK → MULTIPLE VALUES OF SAME FREQUENCY

HOW IS DATA DISTRIBUTED AROUND CENTRAL TENDENCY → MEASURE OF DISPERSION

1] RANGE OF DISPERSION
$R = x_{MAX} - x_{MIN}$
IF $x_{MAX} = x_{MIN}$ → ZERO DISPERSION

HOW IS THE SPREAD OF DATA.
RANGE SHOULD NOT BE USED IF DATA SET HAS EXTREME VALUES.

2] INTER-QUARTILE RANGE (IQR)
SOLUTION TO PROBLEM USING RANGE.
HERE LOWEST 25% & HIGHEST 25% ARE IGNORED AND RANGE IS CALCULATED USING ONLY THE MIDDLE 50%



STANDARD DEVIATION:

AVERAGE DEVIATION FROM THE MIDDLE OF THE DATA.

→ CALCULATE MEAN $\bar{x}$
→ FOR EACH VALUE (OBSERVATION) SUBTRACT $\bar{x}$
→ SQUARE THE OUTPUT [EACH VALUE]
→ FIND THE SUM
→ DIVIDE THE SUM BY (NUMBER OF OBSERVATIONS − 1)
→ ↳ VARIANCE
→ $\sqrt{\text{VARIANCE}}$ = STANDARD DEVIATION

$$\text{VARIANCE} = \frac{\sum_{x=1}^{n} (x - \bar{x})^2}{n-1}$$

$$\text{STD DEV} = \sqrt{\text{VARIANCE}}$$

$(n-1)$ → BECAUSE $\bar{x}$ IS ESTIMATE

COEFFICIENT OF VARIATION
↳ RATIO OF S.D TO MEAN

$$CV = \frac{S}{\bar{x}}$$

50 UNITS WITH SD = 5 → SALES MAN 1
75 UNITS WITH SD = 25 → SALES MAN 2
} BETTER??

$$CV_1 = \frac{5}{50} \qquad CV_2 = \frac{25}{75}$$

$$= \frac{1}{10} \qquad = \frac{1}{3}$$

$$= 10\% \qquad = 33\%$$

THE EMPIRICAL RULE

IF $\mu \pm 1\sigma$ HAS 68% OF DATA
$\mu \pm 2\sigma$ HAS 95% OF DATA
$\mu \pm 3\sigma$ HAS 99.7% OF DATA

SP1	SP2
AVG 50 UNITS	AVG 75 UNITS
SD 5 UNITS	SD 25 UNITS
(40, 60)	(25, 125)
$\mu \mp 2\sigma$	$\mu \pm 2\sigma$
SP1 SELLS ATLEAST 40 UNITS	SP2 SELL ATLEAST 25 UNITS

CHEBYSHEV RULE

IF DATA NOT BELL SHAPED OR BELL SHAPED
ATLEAST $\left(1 - \frac{1}{K^2}\right)$% WILL FALL WITHIN

K STANDARD DEVIATIONS.

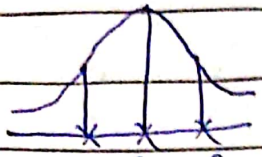
IF K = 2
75% OF DATA WILL FALL BETWEEN 2 SD of MEAN

FIVE NUMBER SUMMARY:

$x_{smallest}$
FIRST QUARTILE (Q_1)
MEDIAN (Q_2)
THIRD QUARTILE (Q_3)
$x_{largest}$

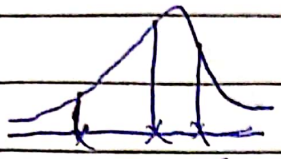
} DIFF B/N EACH IS MEASURE OF DISPERSION

DISTRIBUTION SHAPE



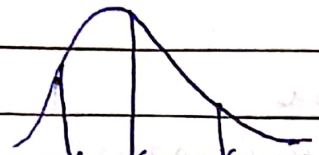
Q_2 → middle of Q_1 & Q_3

SYMMETRIC



Q_2 & Q_3 closer
Q_1 larger observations

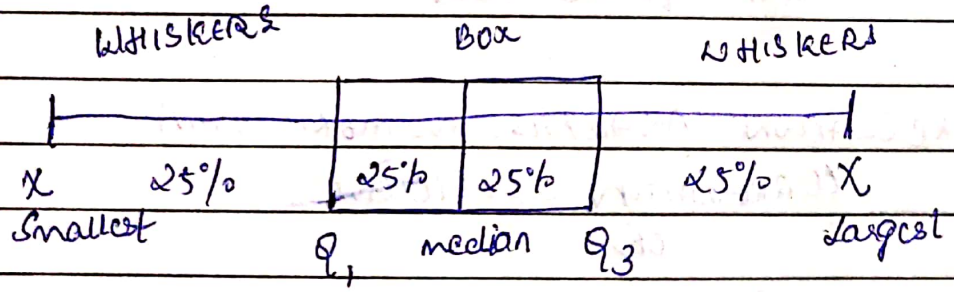
LEFT SKEW



Q_1 & Q_2 closer
Q_3 larger observations

RIGHT SKEW

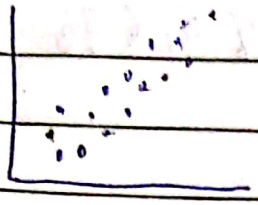
BOX PLOT : GRAPHICAL REP OF FIVE NUMBER SUMMARY.



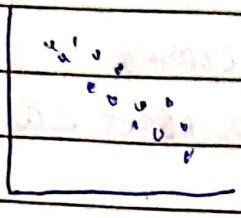
BOX → Q_1 to Q_3

BOX PLOT → HORIZONTAL OR VERTICAL
→ SHOWS THE SKEW
→ GOES TILL 1.5 × IQR

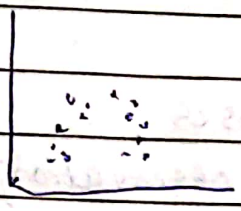
CORRELATION: VARIABLE 1 VARIABLE 2
V1 V2



+VE CORRED
$V_1 \propto V_2$



-VE CORRED
$V_1 \propto \frac{1}{V_2}$



NO CORRED
(MULTIPLE REGRESSION
OR NO PATTERN)

CORRELATION ANALYSIS: NOMINAL DATA
CHI SQUARE TEST

CORRELATION ANALYSIS: NUMERIC DATA
CORRELATION COEFFICIENT
OR
PEARSON'S COEFFICIENT.

SUMMARY:
HISTOGRAM
CENTRAL TENDENCY
DISPERSION
CORRELATION ANALYSIS.