Lecture-7

For odd number of elements :

2 3 5 7 8 9 10 12 15

Q_1 = 5, Median, Q_2 = 8, Q_3 = 10

for even number of data, do not include median in either halves.

10 12 14 15 14 16 17 18 10 19 17 17

sorted data :

10 10 12 14 14 15 16 17 17 17 18 19

$Q_1 = \frac{12+14}{2}$
$= 13$

$Q_2 = \frac{15+16}{2}$
$= 15.5$

$Q_3 = \frac{17+17}{2}$
$= 17$

Quartile for grouped data

$$Q_i = L + \frac{\frac{i \times N}{4} - (\Sigma f)_l}{f} \times c \quad ; \; i = 1, 2, 3$$

	x	f	Cumulative freq
	20-30	4	4
	30-40	8	12
Q_1 class	40-50	18	30
Q_2 median class	50-60	30	60
Q_3 class	60-70	15	75
	70-80	10	85
	80-90	8	93
	90-100	~~8~~7	100

$N = 100$

Quartile class is identified by $\frac{i \times N}{4}$ th observation

Q_1 class $= \frac{1 \times 100}{4} = 25$th observation (40-50)

Q_2 class $= \frac{2 \times 100}{4} = 50$th observation (50-60)

Q_3 class $= \frac{3 \times 100}{4} = 75$th observation (60-70)

$$Q_1 = 39.5 + \frac{\frac{1 \times 100}{4} - 12}{18} \times 10$$

$$= 46.72$$

$$Q_2 = 49.5 + \frac{\frac{2 \times 100}{4} - 30}{30} \times 10$$

$$= 56.17.$$

$$Q_3 = 59.5 + \frac{\frac{3 \times 100}{4} - 60}{15} \times 10$$

$$= 69.5$$

Deciles : $D_1, \ldots, D_9$

$$D_i = L + \frac{\frac{i \times N}{10} - (\Sigma f)_l}{f} \times c \quad ; \; i = 1, 2, \ldots 9$$

Decile class is identified by $\frac{i \times N}{10}$ th observation.

Percentiles : $P_1, P_2, \ldots, P_{99}$

$$P_i = L + \frac{\frac{i \times N}{100} - pcf}{f} \times c \quad ; \; i = 1, 2, \ldots, 99$$

Percentile class is identified by $\frac{i \times N}{100}$ th observation.

Box whisker plot → 5 number summary

- → L (minimum)
- → Q_1
- → Q_2
- → Q_3
- → H (maximum)

25 28 29 29 30 34 35 35 37 38

L

$Q_1 = 29$

$$Q_2 = \frac{30+34}{2} = 32$$

$Q_3 = 35$

H

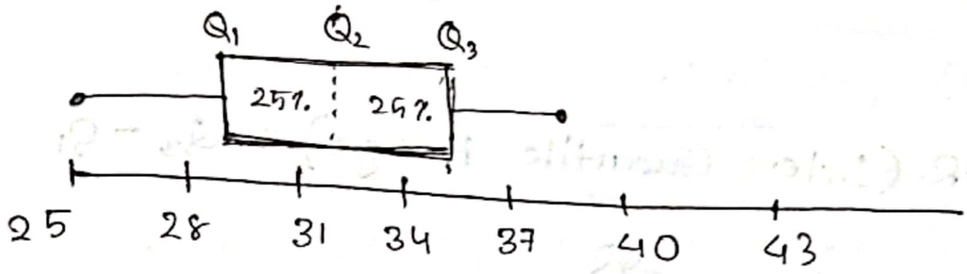
$L = 25$

$Q_1 = 29$

$Q_2 = 32$

$Q_3 = 35$

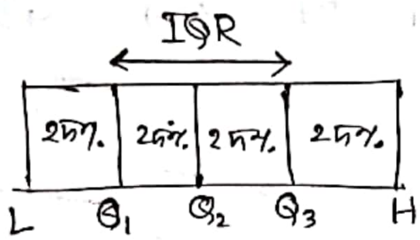
$H = 38$


Q1
Q2
Q3
25%
25%
25
28
31
34
37
40
43

Lecture-8

Outlier detection using IQR

IQR (Inter Quartile Range) $= Q_3 - Q_1$



Compute $x < Q_1 - 1.5 \times IQR$ or

or $x > Q_3 + 1.5 \times IQR$.

x is an outlier

$-15 \quad 15 \quad 15 \quad 15 \quad \underbrace{17 \,\vdots\, 17}_{Q_1 = 17} \quad 19 \quad 20 \quad 21 \quad \underbrace{22 \mid 23}_{Q_2 = \frac{22+23}{2}} \quad 25$

$27 \quad \underbrace{35 \mid 41}_{} \quad 50 \quad 63 \quad 65 \quad 101$

$Q_2 = \frac{22+23}{2} = 22.5$

$Q_3 = \frac{35+41}{2}$

$= 39$.

$IQR = Q_3 - Q_1 = 39 - 17 = 22$

$Q_1 - 1.5 \times IQR = 22.5 - 1.5 \times 22$

$= -16$

$Q_3 + 1.5 \times IQR = 72$.

Outlier $= 101$

Variable

A variable is a characteristic, often but ~~not always~~ quantitatively measured, containing two or more values or categories that can vary from person to person, object to object or from phenomenon to phenomenon.

Variable

Observation/ Record

ID	Gender	Age	Education level	Annual Income
120	F	24	Bachelor's	$ 32000
137	M	45	Diploma	$ 55000
138	M	31	Master's	$ 47000

Variable এর value এক observation থেকে আরেক observation এ change হতে পারে।

Variable

1. Qualitative (categorical) → not numerical

Ex: color of a ball, breed of a dog etc.

2. Quantitative (Numerical)

Ex: population of a city.

Level of measurement

defines the amount of information contained in the data.

1. Nominal
2. Ordinal
3. Interval scales
4. Ratio scales

Level of measurement যত higher, তত data নিয়ে বেশি কাজ করা যাবে।

* Nominal variable – Mode calculate করা যায়

a categorical variable without an intrinsic (general) order.

Ex: Gender (Male, Female)

Nationality (Indian, American).

* Ordinal variable – Median calculate করা যায়.

a categorical variable with some intrinsic order

Ex: Frequency (always, often, sometimes, never)

Rating (good, fair, poor).

Lecture-9

Interval Scales – Numeric value (সুষম ভাগ করা যাবে না)

* No absolute zero.

* Interval data are measured and have constant, equal distances between values, but the zero point is arbitrary.

* No meaningful zero.

Ex: Temperature difference.

Ratio Scales – Numeric value.

* Meaningful zero.
* An absolute zero.

Ex: height, weight.

Continuous variable - numeric variable. Observations can take any value between a certain set of real numbers.

Discrete variable - numeric variable. Obsearvations can take a value based on a count from a set of distinct whole values.

Univariate vs Multivariate data.

Pie charts

- Summarize categorical variable.

Bar graph

- Summarize categorical variable.
- vertical bars for each category.

Lecture-10

Ogive :

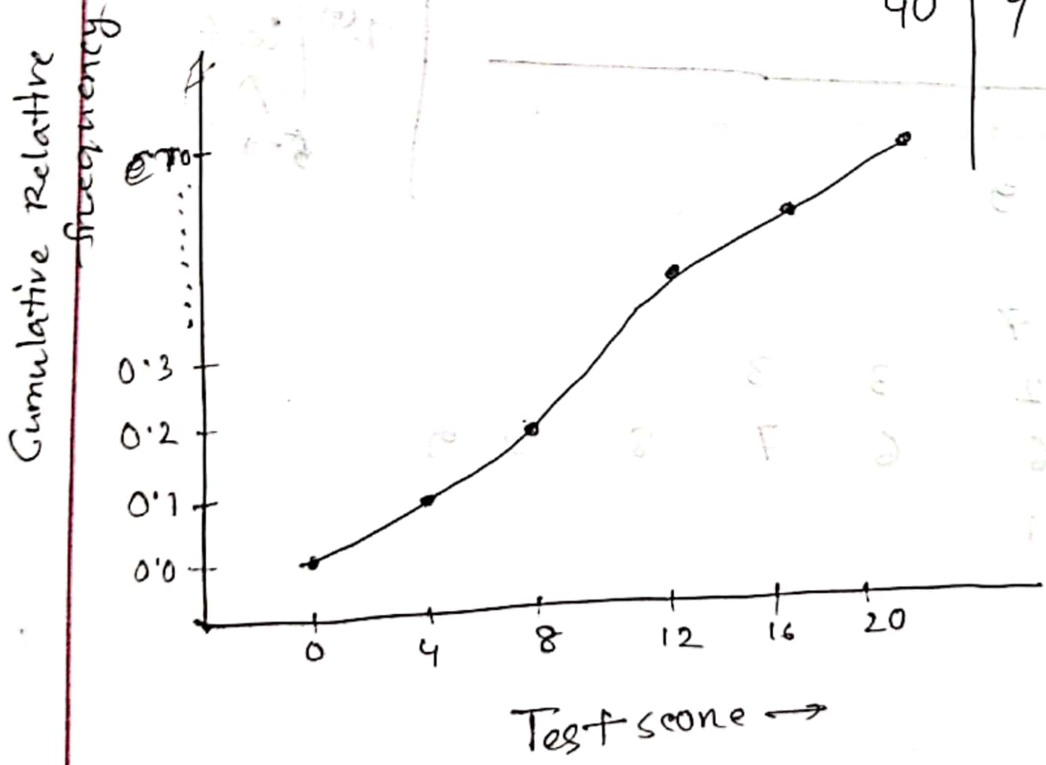
A line graph of cumulative frequency or cumulative relative frequency distribution.

Components !

1. Title → population or sample
2. Vertical scale – cumu-freq. or rel. cumu. freq
3. Horizontal scale – upper class boundary

x_i	f_i	Cumu. freq	Relative cumu. freq $= \frac{cf_i}{n}$
10	1	1	1/10
20	2	3	3/10
30	3	6	6/10
40	4	10	10/10 = 1

সর্বোচ্চ value 1 হয়



$x_i = 30$ হলে cumu freq = 6, মানে total 60% data এই range এর মধ্যে আছে.

Stem & Leaf Display:

Stem - Multi digit / single digit.

Leaf - single digit.

Stem ↑ Leaf

7 | 6 9

76, 79

Stem ↑ Leaf

14 | 2

$1^{(0)}$ | Leaf Leaf ↑ 0-4 0-4

$1^{(5)}$ | leaf ↑ 5-9

$7^{(0)}$							
$7^{(5)}$	6	9					
$2^{(0)}$	4						
$2^{(5)}$	6	7					
$3^{(0)}$	1	2	3	3			
$3^{(5)}$	5	6	6	7	8	9	9
$4^{(0)}$	0	1					
$4^{(5)}$	9						
$5^{(0)}$							
$5^{(5)}$	5						

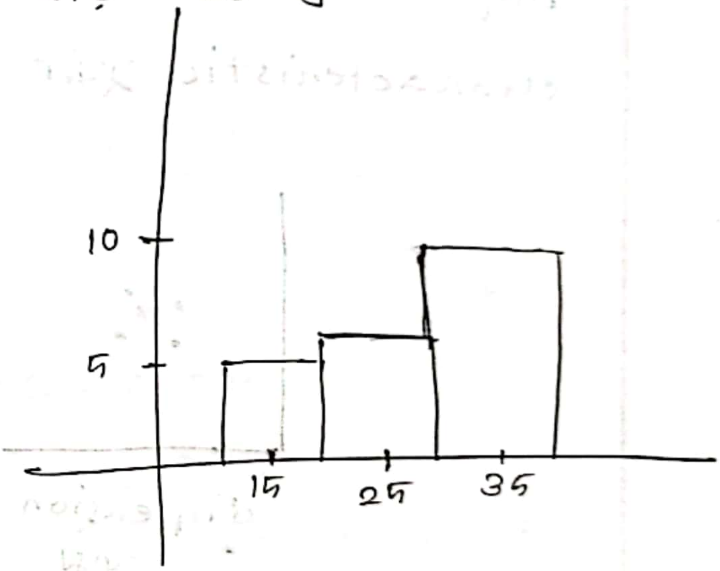
Histogram

A bar graph representing a frequency distribution of quantitative variable.

$$\text{Relative frequency} = \frac{f_i}{\sum f_i} = \frac{f_i}{n}.$$

overall class এর frequency কোনটা কত বেশি সেটা জানা যায় histogram দিয়ে।

10-20	৫
20-30	6
30-40	10



▥ → 0

□ → 2৫৫

Lecture-11

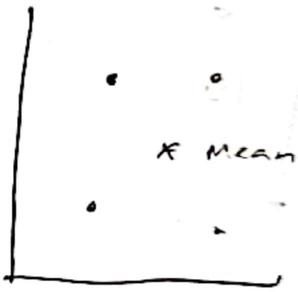
Measurement of dispersion

central value থেকে data গুলোর কত বা কতটা ছড়িয়ে বা বিস্তৃত।

central tendency থেকে central value জানা যায়,

Dispersion থেকে overall dataset এর characteristics জানা যায়,

dispersion কম



বেশি.

i) Range : simple and sensitive to outlier. It can be misleading

Range = H - L.

6 46 46 46 46 46

Range = 46 - 6 = 40

6 6 6 6 6 46

Range = 46 - 6 = 40

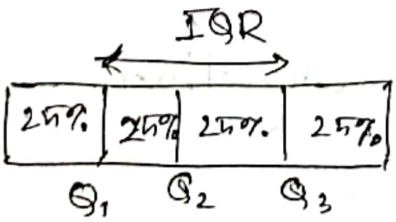
6 20 25 35 39 46

Range = 46 - 6 = 40

ii) Quartile Deviation → semi Quartile Range

not sensitive to outlier.

$$Q.D = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$



Limitation - 1st 25% and last 25% data consider করে না।

#

Profits	No. of company	c f
20-30	4	4
30-40	8	12
Q_1 class→40-50	18	30
50-60	30	60
Q_3 class→60-70	15	75
70-80	10	85
80-90	8	93
90-100	7	100
		N=100.

$$Q_i = L + \frac{\frac{i \times N}{4} - p.c.f}{f} \times h$$

Quantile class is given by $\frac{i \times N}{4}$ th observation.

$$Q_1 = \frac{1 \times 100}{4} = 25\text{th observation.}$$

$= 40-50$ class

$$Q_3 = \frac{3 \times 100}{4} = 75\text{th observation.}$$

$$Q_1 = 40 + \frac{\frac{1 \times 100}{4} - 12}{18} \times 10$$

$$= 47.22 \text{ lakhs}$$

$$Q_3 = 60 + \frac{\frac{3 \times 100}{4} - 60}{30} \times 10$$

$$= 70 \text{ lakhs}$$

$$Q_3 - Q_1 = 70 - 47.22 \text{ Lakhs}.$$

$$= 22.78 \text{ Lakhs}$$

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{22.78}{2} = 11.39 \text{ lakhs}.$$

iii) Mean Absolute Deviation:

Average distance from Average data

for un grouped data,

$$M.D = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$\sum_{i=1}^{n} x_i - \bar{x} = 0$$

Proof-self study

For grouped data,

$$M.D = \frac{1}{n}\sum_{i=1}^{n} f_i|x_i - \bar{x}|$$

$$\bar{x} = \frac{\sum f_i x_i}{N}$$

৭ 10 68 90 40

$$\bar{x} = \frac{208}{4} = 52$$

$$M.D = \frac{|10-52| + |68-52| + |90-52| + |40-52|}{4}$$

$$= \frac{108}{4} = 27.$$

Variance : outlier ও consider করে,

For ungrouped data,

Population variance, $\sigma^2 = \dfrac{\sum (x_i - \bar{x})^2}{N}$

for grouped data.

$$\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N}$$

Sample variance,

For ungrouped data,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

For grouped data,

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N-1}$$

For ungrouped data,

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

$$= \frac{1}{n} \sum \left(x_i^2 + \bar{x}^2 - 2 x_i \bar{x} \right)$$

$$= \frac{1}{n} \sum x_i^2 + \frac{1}{n} \sum \bar{x}^2 - \frac{2\bar{x}}{n} \sum x_i$$

$$= \frac{1}{n} \sum x_i^2 + \frac{\bar{x}^2}{n} \sum_{i=1}^{n} 1 - \frac{2\bar{x}}{n} \sum x_i$$

$$= \frac{1}{n} \sum x_i^2 + \frac{\bar{x}^2}{n} . n - 2\bar{x} . \bar{x}$$

$$= \frac{1}{n} \sum x_i^2 + \bar{x}^2 - 2\bar{x}^2$$

$$= \frac{1}{n} \sum x_i^2 - \bar{x}^2 .$$

$$\cancel{= \frac{1}{n} \sum x_i^2}$$

$$= \frac{x_i^2}{n} - \left(\frac{\sum x_i}{n} \right)^2$$

for grouped data,

$$\sigma^2 = \frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2$$