SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107
An Autonomous Institution
Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF MANAGEMENT STUDIES

COURSE NAME : 23BAP106– FUNDAMENTALS OF DATA ANALYSIS

I YEAR /I SEMESTER

Unit I – EXPLORING DATA ANALYTICS

Topic 2: CALCULATE AND INTERPRET COMMON DESCRIPTIVE STATISTICS

STATISTICS

Characteristics of the Data Set is explored with various numerical measures namely

- Measures of Central Tendency
- Measures of Dispersion
- Measures of Position
- Measures of Kurtosis

STATISTICS

Characteristics of the Data Set is explored with some numerical measures namely

- Measures of Skewness

STATISTICS

- What is Descriptive Statistics

Are the Various Methods that help collect, summarize ,present and analyze a set of Data

Eg : Mean, Median , Mode, Standard Deviation,Tables, charts etc….,

MEASURES OF CENTRAL TENDENCY

STATISTICS

- In Practical Situations we need one single value to represent the variable/variables in a whole set of data
- Eg the Average heights of students in a class
- Hence it is preferable to characterize each group of observations by a single value .

STATISTICS

- All other values cluster around or vary around that one single value
- This is why it is called a Central Measure of Tendency of that Group
- A measure of Central Tendency is a representative value of the entire group of data

Measures of Central Tendency

. Measure of central tendency provides a very convenient way of describing a set of scores with a single number that describes the PERFORMANCE of the group.

. It is also defined as a single value that is used to describe the “center” of the data.

. There are three commonly used measures of central tendency. These are the following:

. MEAN

. MEDIAN

. MODE

MEAN

The Arithmetic Mean (Mean) is the most common measure of central tendency

Mean is computed by adding together all values in a Dataset and then dividing that sum by the number of values in a data set

$$\overline{X} = \frac{\text{Sum of the Values}}{\text{Number of Values}}$$

MEAN

Example – Given the following data sample:

2 5 8 −3 5 2 6 5 −4

The simple mean of the *sample* of nine measurements is given by:

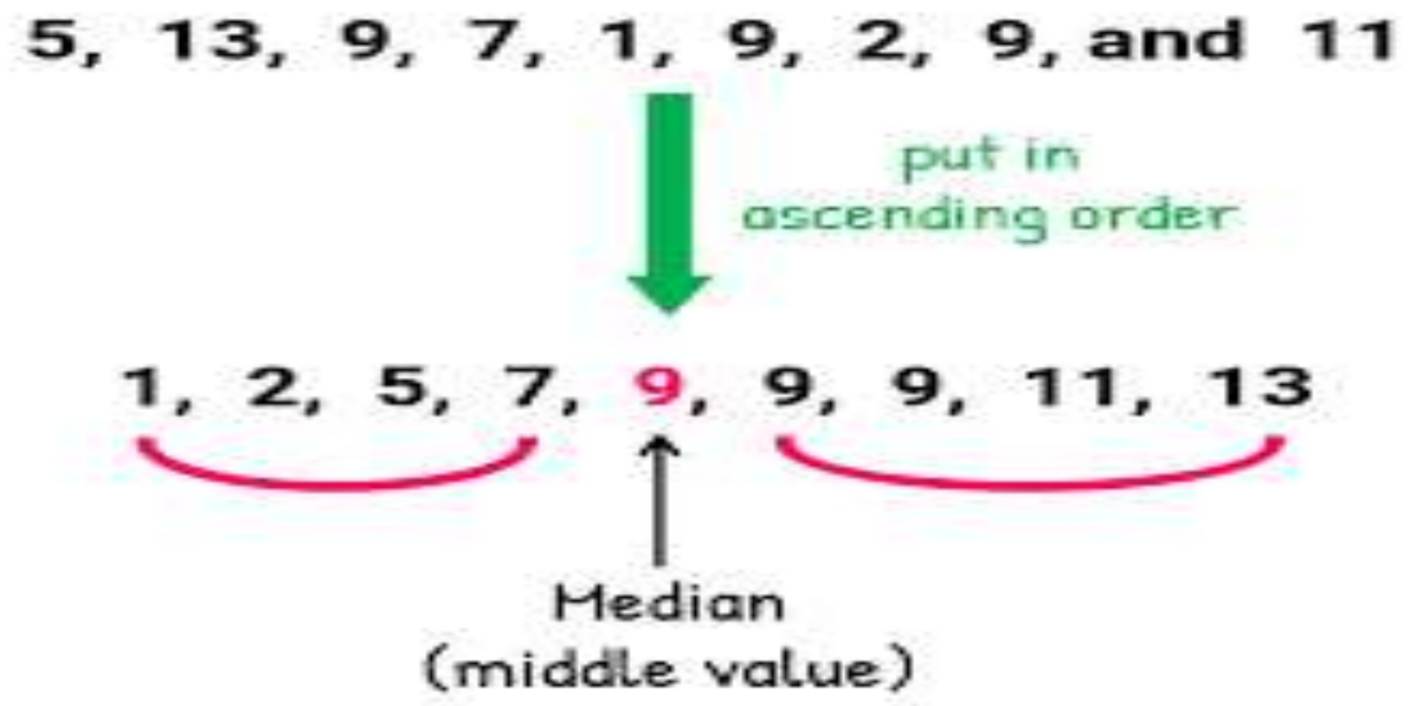
$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{n}$$

$$= \frac{\boxed{2} + \boxed{5} + \boxed{8} + \boxed{-3} + \boxed{5} + \boxed{2} + \boxed{6} + \boxed{5} + \boxed{-4}}{\boxed{9}}$$

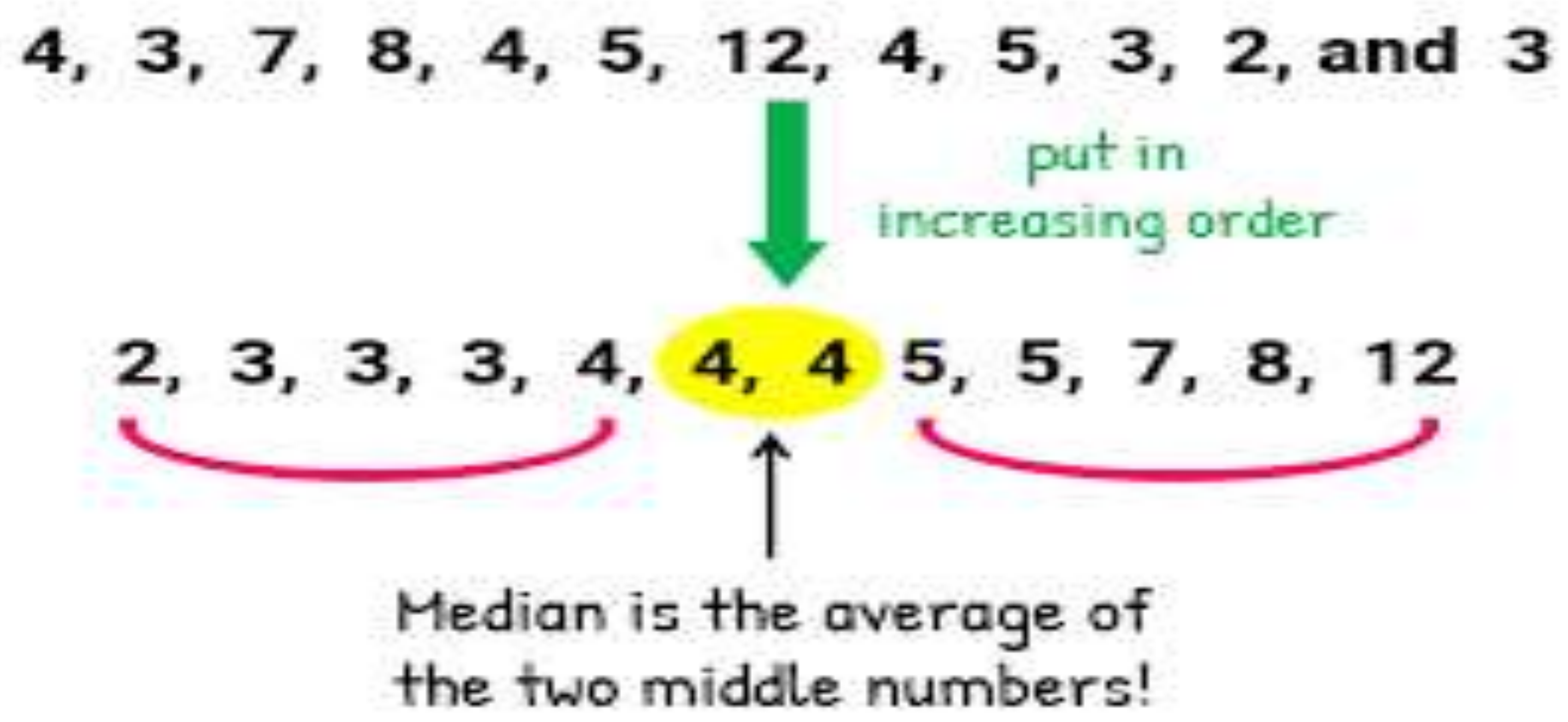
$$= \frac{26}{9} = 2{,}89$$

The ***median*** is the MIDDLE VALUE when all the values are placed in order of size.

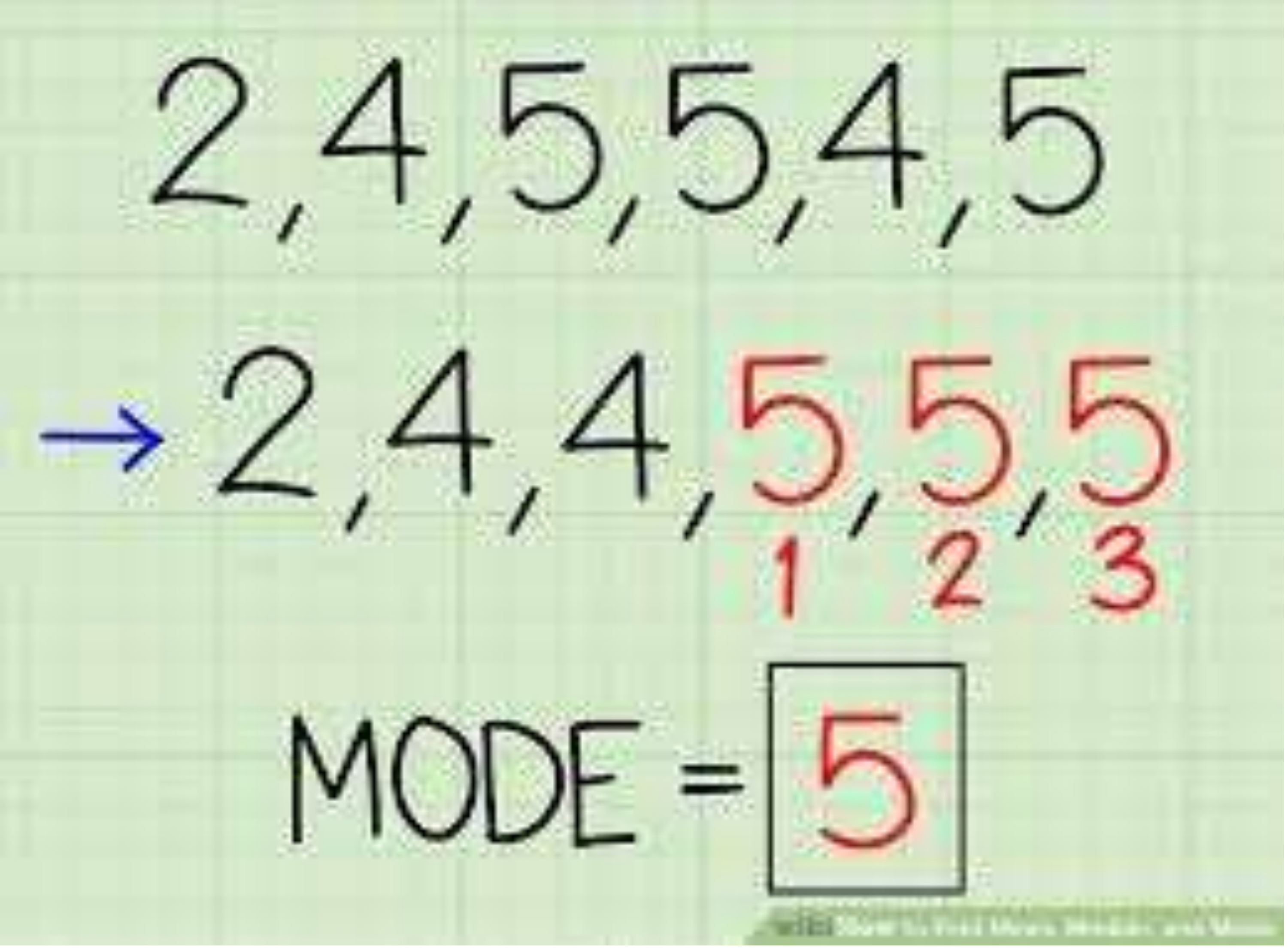
Median calculation When there are Odd Nos



Median calculation When there are Even Nos



The ***mode*** is the value that occurs MOST OFTEN.


2, 4, 5, 5, 4, 5
→ 2, 4, 4, 5, 5, 5
1 2 3
MODE = 5

STATISTICS

What is Ungrouped data ?????

Ungrouped Data

- Data that has not been organized into groups.
- Ungrouped data looks like a big list of numbers.

55	63	44	37	50	57	44	57	42	48	33	44
58	40	54	65	39	27	29	58	38	45	70	60
30	35	56	78	55	27	50	28	44	28	60	61
39	27	65	43								

STATISTICS

What is grouped data ?????

Grouped data are data that has been bundled together in categories

The following table shows the distribution of heights of a group of 40 students.

Height (in cm)	No of students
159-162	1
163-166	4
167-170	11
171-174	12
175-178	6
179-182	4
183-186	2

Population Mean

For ungrouped data, the population mean is the sum of all the population values divided by the total number of population values:

POPULATION MEAN

$$\mu = \frac{\Sigma X}{N} \qquad [3\text{–}1]$$

where:

- μ represents the population mean. It is the Greek lowercase letter "mu."
- N is the number of values in the population.
- X represents any particular value.
- Σ is the Greek capital letter "sigma" and indicates the operation of adding.
- ΣX is the sum of the X values in the population.

EXAMPLE – Population Mean

There are 12 automobile manufacturing companies in the United States. Listed below is the number of patents granted by the United States government to each company in a recent year.

Company	Number of Patents Granted	Company	Number of Patents Granted
General Motors	511	Mazda	210
Nissan	385	Chrysler	97
DaimlerChrysler	275	Porsche	50
Toyota	257	Mitsubishi	36
Honda	249	Volvo	23
Ford	234	BMW	13

Is this information a sample or a population? What is the arithmetic mean number of patents granted?

$$\mu = \frac{\sum X}{N} = \frac{511+385+275+\ldots+36+23+13}{12} = \frac{2340}{12} = 195$$

Sample Mean

- For ungrouped data, the sample mean is the sum of all the sample values divided by the number of sample values:

SAMPLE MEAN $$\bar{X} = \frac{\Sigma X}{n} \qquad [3\text{–}2]$$

where:

$\bar{X}$ is the sample mean. It is read "X bar."

n is the number of values in the sample.

EXAMPLE – Sample Mean

SunCom is studying the number of minutes used monthly by clients in a particular cell phone rate plan. A random sample of 12 clients showed the following number of minutes used last month.

90	77	94	89	119	112
91	110	92	100	113	83

What is the arithmetic mean number of minutes used?

$$\bar{X} = \frac{\Sigma X}{n} = \frac{99+77+94+\ldots+100+113+83}{12} = \frac{1{,}170}{12} = 97.5$$

Mean….

Sample mean is fundamentally different from population mean

because

samples from a population can have different values for their sample mean,

that is,

they can vary from sample to sample within the population.

Mean….

The population mean, however, is constant for a given population.

MEAN

Properties of the Mean

- It measures stability. Mean is the most stable among other measures of central tendency because every score contributes to the value of the mean.

- The sum of each score’s distance from the mean is zero.
- It may easily affected by the extreme scores.
- It can be applied to interval level of measurement.
- It may not be an actual score in the distribution.
- It is very easy to compute.

MEAN

Mean for Grouped Data

Grouped data are the data or scores that are arranged in a frequency distribution.

Frequency distribution is the arrangement of scores according to category of classes including the frequency.

Frequency is the number of observations falling in a category.

MEAN

The only one formula in solving the mean for grouped data is called midpoint method. The formula is:

$$\bar{X} = \frac{\Sigma f x_m}{n}$$

$\bar{X}$ = mean value

x_m = midpoint of each class or category

f = frequency in each class or category

$\Sigma f x_m$ = summation of the product of $f x_m$

MEAN

Steps in Solving Mean for Grouped Data

1. Find the midpoint or class mark (X_m) **of each class or category using the formula** $X_m = \frac{LL + LU}{2}$.

2. Multiply the frequency and the corresponding class mark $f x_m$.

3. Find the sum of the results in step 2.

4. Solve the mean using the formula

$$\bar{X} = \frac{\Sigma f x_m}{n}$$

MEAN

Example:

Scores of 40 students in a science class consist of 60 items and they are tabulated below.

X	f	Xm	fXm
10 - 14	5	12	60
15 - 19	2	17	34
20 - 24	3	22	66
25 - 29	5	27	135
30 - 34	2	32	64
35 - 39	9	37	333
40 - 44	6	42	252
45 - 49	3	47	141
50 - 54	5	52	260
	n = 40		Σ f Xm = 1 345

$$\bar{X} = \frac{\Sigma f x_m}{n}$$

$$= \frac{1345}{40}$$

$$=33.63$$

MEAN

Analysis:

The mean performance of 40 students in science quiz is 33.63. Those students who got scores below 33.63 did not perform well in the said examination while those students who got scores above 33.63 performed well.

Arithmetic Mean for Grouped Data

The Frequency distribution below represents the weights in pounds of a sample of packages carried last month by a small airfreight company.

CLASS	FREQUENCY
10.0 -10.9	1
11.0 -11.9	4
12.0 -12.9	6
13.0 -13.9	8
14.0 -14.9	12
15.0-15.9	11
16.0 -16.9	8
17.0 -17.9	7
18.0 -18.9	6
19.0 -19.9	2

Arithmetic Mean for Grouped Data

Compute the sample mean

Arithmetic Mean for Grouped Data

David Furniture Company has a revolving credit agreement with the first National Bank.The Loan showed the following ending monthly balances last year.

Month	Ending Monthly Balance($)
Jan	121,300
Feb	112,300
March	72,800
April	72,800
May	72,800
June	57,300
July	58,700

Arithmetic Mean for Grouped Data

David Furniture Company (contd)

Month	Ending Monthly Balance($)
August	61,100
September	50,400
October	52,800
November	49,200
December	46,100

The Company is eligible for a reduced rate of interest if it's average monthly balance is over $ 65,000.Does it qualify?

MEDIAN

- Median is what divides the scores in the distribution into two equal parts.

- Fifty percent (50%) lies below the median value and 50% lies above the median value.

- It is also known as the middle score or the 50th percentile.

MEDIAN

Median of Ungrouped Data

1. Arrange the scores (from lowest to highest or highest to lowest).

2. Determine the middle most score in a distribution if n is an odd number and get the average of the two middle most scores if n is an even number.

Example 1: Find the median score of 7 students in an English class.

x (score)
19
17
16
15
10
5
2

MEDIAN

Example: Find the median score of 8 students in an English class.

x (score)
30
19
17
16
15
10
5
2

$$\tilde{x} = \frac{16+15}{2}$$

$\tilde{x}$ = 15.5

MEDIAN FOR GROUPED DATA

$$X = X_{LB} + \left(\frac{\frac{N}{2} - cf_b}{f_m} \right) i$$

where:

X = median
X_{LB} = lower boundary of the median class
N = total frequency
cf_b = cumulative frequency before the median class
f_m = frequency of the median class
i = size of the class interval

MEDIAN FOR GROUPED DATA

- What is Median Class

- Let us see how it is calculated

0 – 15	5	5
15 – 30	20	25
30 – 45	40	65
45 – 60	50	115
60 – 75	25	140

Here, $N = 140 \Rightarrow \frac{N}{2} = 70$

The cumulative frequency just greater than 70 is 115.

Hence, median class is 45 – 60.

Find the Median for the grouped data

Class Interval	Class Frequency (f)	< Cumulative Frequency (<CF)
43-53	7	33
54-64	23	56
65-75	55	111
76-86	7	118

1. Lower class boundary of *M*

$$\tilde{x}_{LB} = LL - 0.5$$

$$\tilde{x}_{LB} = 65 - 0.5$$

$$= 64.5$$

2. Class size (*i*)

2. Class size (*i*)

$$i = (UL - LL) + 1$$

$$i = 75 - 65 + 1$$

$$= 11$$

Find the Median for the grouped data

Class Interval	Class Frequency (f)	< Cumulative Frequency (<CF)
43-53	7	33
54-64	23	56
65-75	55	111
76-86	7	118

3. Less than cumulative frequency before the median class

$$< cfb = 56$$

4. Median class frequency

$$fm = 55$$

$$\tilde{x} = \tilde{x}_{LB} + i\left(\frac{\frac{N}{2} - < cfb}{fm}\right) = 64.5 + 11\left(\frac{\frac{130}{2} - 56}{55}\right)$$

$$\tilde{x} = 64.5 + 11\left(\frac{65 - 56}{55}\right) = 64.5 + 11\left(\frac{9}{55}\right)$$

$$= 64.5 + 11(0.16363)$$

$$\tilde{x} = 64.5 + 1.8 = 66.3$$

MEDIAN PROBLEM

The below is the data on the account balance of 600 customers .Find the median account balance .

Class in Dollars	Frequency
0-49.99	78
50.00 – 99.99	123
100.00 – 149.99	187
150.00 – 199.99	82
200.00 – 249.99	51
250.00 – 299.99	47
300.00 – 349.99	13
350 .00 – 399.99	9
400.00 – 449.99	6
450.00 – 499.99	4
	600

MEDIAN PROBLEM

The Tamilnadu Road Transport Authority in Chennai feels that excessive speed on its buses increases maintenance cost .It believes that a reasonable medium time from Meenambakkam Airport to Vadapalani is 30 mts .From the following sample data in minutes can you help them determine whether the buses have been driven at excessive speeds ? If you conclude from these data that they have, what explanation might you get from the bus drivers?

MEDIAN PROBLEM

17	32	21	22
29	19	29	34
33	22	28	33
52	29	43	39
44	34	30	41

MEDIAN

Properties of the Median

- It may not be an actual observation in the data set.
- It can be applied in ordinal level.
- It is not affected by extreme values because median is a positional measure.

When to Use the Median

- The exact midpoint of the score distribution is desired.
- There are extreme scores in the distribution.

MODE

The mode or the modal score is a score or scores that occurred most in the distribution.

It is classified as unimodal, bimodal, trimodal or mulitimodal.

Unimodal is a distribution of scores that consists of only one mode.

Bimodal **is a distribution of scores that consists of two modes.**

Trimodal **is a distribution of scores that consists of three modes** or multimodal **is a distribution of scores that consists of more than two modes.**

MODE

Example: Scores of 10 students in Section A, Section B and Section C.

Scores of Section A	Scores of Section B	Scores of Section C
25	25	25
24	24	25
24	24	25
20	20	22
20	18	21
20	18	21
16	17	21
12	10	18
10	9	18
7	7	18

MODE

The score that appeared most in Section A is 20, hence, the mode of Section A is 20. There is only one mode, therefore, score distribution is called unimodal.

The modes of Section B are 18 and 24, since both 18 and 24 appeared twice. There are two modes in Section B, hence, the distribution is a bimodal distribution.

The modes for Section C are 18, 21, and 25. There are three modes for Section C, therefore, it is called a trimodal or multimodal distribution.

MODE

Properties of the Mode

- It can be used when the data are qualitative as well as quantitative.
- It may not be unique.
- It is affected by extreme values.
- It may not exist.

When to Use the Mode

- When the "typical" value is desired.
- When the data set is measured on a nominal scale.

Estimating Mode

- The following are the marks scored by 20 students in a class. Find the Mode

90,70,50,30,40,86,65,73,68,90,90,10,73,25,35,88,
67,80,74,46

Answer is ???

Estimating Mode

- A doctor who checked 9 patients sugar levels is given below. Find the modal value of the sugar levels 80,112,110,115,124,130,100,90,150, 180

Estimating Mode

- Compute Modal value for the following observations

2,7,10,12,10,19,2,11,3,12

MODE for grouped Data:

Modal Class is the class interval with the heights class frequency.

Class Interval	Class Frequency (f)

$$\hat{x} = x_{LB} + i\left(\frac{fm - fmb}{2fm - fma - fmb}\right)$$

Find the Mode for the

Class Interval	Class Frequency (f)
43-53	7
54-64	23
65-75	55
76-86	7

Modal class

1. Lower class boundary of the modal class

$$x_{LB} = 65 - 0.5 = 64.5$$

2. Class size $i = 75 - 65 + 1 = 11$

3. Class frequency of the modal class

$$fm = 55$$

4. Class frequency of the class after the modal class $fma = 7$

5. Class frequency of the class before the modal class $fmb = 23$

$$\hat{x} = x_{LB} + i\left(\frac{fm - fmb}{2fm - fma - fmb}\right)$$

Estimating Mode

- The Number of solar heating systems available to the public is quite large , and their heat storage capacities are quite varied here is a distribution of heat –storage capacity (in days) of 28 systems that were tested recently by Universal Laboratories , Inc

Universal Laboratories Inc knows that its report on the tests will be widely circulated and used as the basis for tax legislation on solar heat allowances .It therefore wants the measures it uses to be as the

Estimating Mode

Basis for tax legislation on solar heat allowances .It therefore wants the measures it uses to be reflective of the data as possible.

Days	Frequency
0 – 0.99	2
1-1.99	4
2-2.99	6
3-3.99	7
4-4.99	5
5-5.99	3
6-6.99	1

Estimating Mode

- Ed Grant is the director of the Student Financial Aid Office at Wilderness College. He has used available data on the summer earnings of all students who have applied to his office for financial aid to develop the following frequency distribution.

Find the Mode .

Also if student aid is restricted to those whose

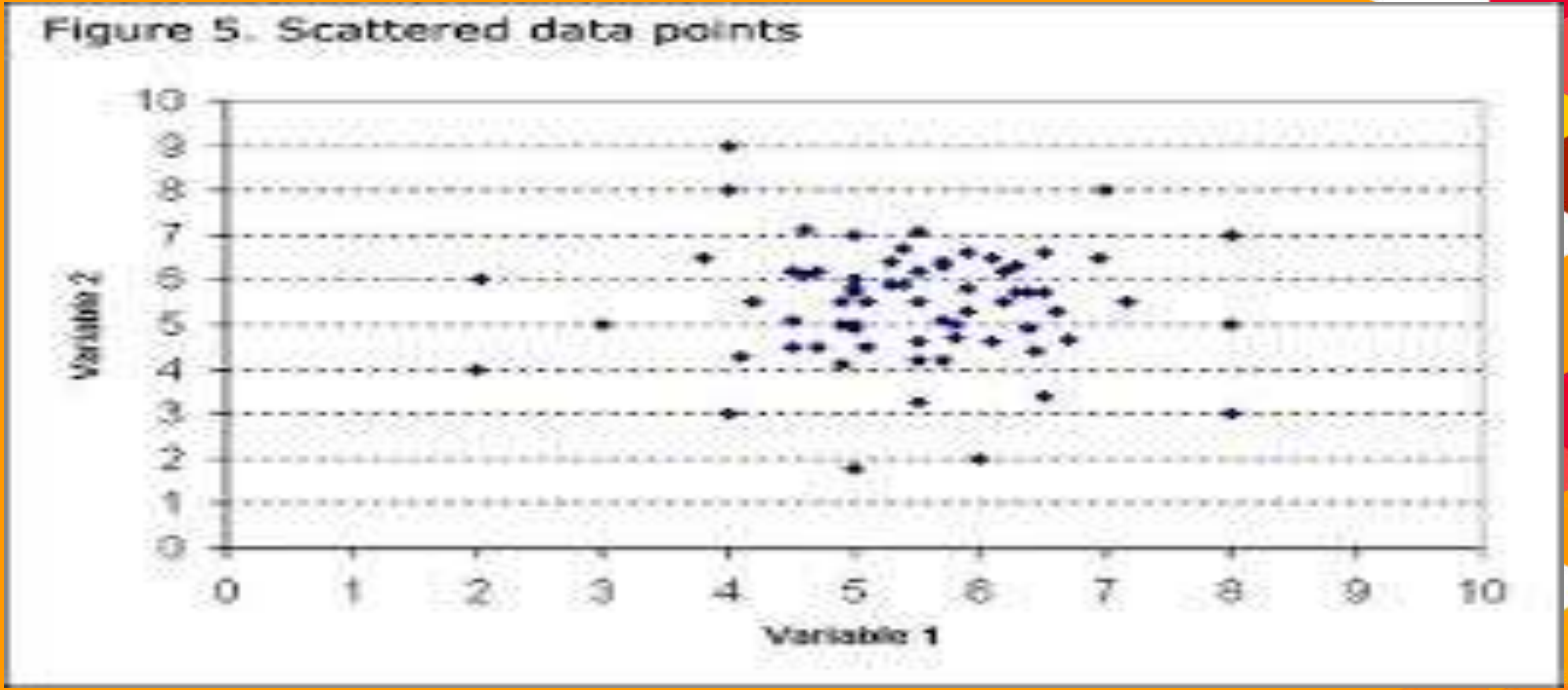
Estimating Mode

Whose summer earnings were at least 10 percent lower than the modal summer earnings, how many of the applicants qualify.

Summer Earnings ($)	Number of Students
0 - 499	231
500 - 999	304
1000 - 1499	400
1500 - 1999	296
2000 -2499	123
2500 - 2999	68
3000 or more	23

Dispersion or Variability

- It is the spread of the data in a distribution or the extent to which observations are scattered.



Measures of Dispersion

Measures of Dispersion

- The numerical values by which we measure the dispersion or variability of a set of data or a frequency distribution are called measures of dispersion.

- There are two kinds of Measures of Dispersion:

1. Absolute measures of dispersion
2. Relative measures of dispersion

Absolute Measure of Dispersion

- Absolute Measures on Dispersion gives you a pure number without a unit of measure like kg, cm ,Rs etc.,

Eg Range (a very commonly used Absolute Measure of Dispersion)

Relative Measure of Dispersion

- Relatives Measures of Dispersion is a ratio

- Relative Measures of Dispersion helps us to compare between two or more groups or sets of data

Eg Percentage, Coefficient of Variation

The Absolute measures of dispersion are:

1. Range
2. Quartile deviation
3. Mean deviation
4. Variance & standard deviation

The Relative measures of dispersion are:

1. Coefficient of range
2. Coefficient of quartile deviation
3. Coefficient of mean deviation
4. Coefficient of variation

1. Range

- Range is the difference between the largest & smallest observation in set of data.
  - In symbols, Range = L – S.

  Where,

  L = Largest value.

  S = Smallest value.

- The monthly incomes in rupee of seven employees of a firm are 5500,5750,6500,6750,7000 & 8500. Compute Range

- **Solution**

The range of the income of the employees is

Range = 8500- 5500

= 3000

Calculating Range for Grouped Data

Range = (Upper class boundary of the Highest Interval - Lower class Boundary of the Lowest Interval)

Illustrative Example: solve for the range

Scores in the Second Periodical Test of 7 – Faith in Mathematics 7

Scores	Frequency
46 – 50	1
41 – 45	10
36 – 40	10
31 – 35	16
26 – 30	9
21 - 25	4

Solutions:
Upper Class Limit of the highest Interval = 50
Upper Class Boundary of the Highest Interval = 50 + 0.5 = 50.5

Lower Class Limit of the lowest Interval = 21
Lower Class Boundary of the Lowest Interval = 21 - 0.5 = 20.5

Range = Upper Class Boundary of the Highest Interval - Lower Class Boundary of the Lowest Interval

Range = 50.5 – 20.5

Range = 30

Therefore, the range of the given data set is 30.

Calculate Range for this data

Marks	60 -63	63 -66	66 -69	69 -72	72- 75
No of Students	5	18	42	27	8

When To Use the Range

- The range is used when you have ordinal data or you are presenting your results to people with little or no knowledge of statistics.

- The range is rarely used in scientific work as it is fairly insensitive.

When To Use the Range

- It depends on only two scores in the set of data, X and X

  Two very different sets of data can have the same range:
  1 1 1 1 9 vs 1 3 5 7 9

Merits and Demerits of Range

Merits

1. The range measure the total spread in the set of data.
2. It is rigidly defined.
3. It is the simplest measure of dispersion.
4. It is easiest to compute.
5. It takes the minimum time to compute.
6. It is based on only maximum and minimum values.

Demerits

1. It is not based on all the observations of a set of data.
2. It is affected by sampling fluctuation.
3. It cannot be computed in case of open-end distribution.
4. It is highly affected by extreme values (outliers).

Coefficient of Range

- The coefficient of range is a relative measure corresponding to range and is obtained by the following formula:

$$\boxed{\text{Coefficient of range} = \frac{L-S}{L+S} \times 100}$$

- where, "L" and "S" are respectively the largest and the smallest observations in the data set.

Concept of Coefficient of Range

- Let us take two set of observations.

SET A 10,15,18,20,20 Five Marks of Students in Maths out of 25 marks

SET B 30,35,40,45,50 Five Marks of Students in English out of 100 Marks

The values of the ranges and coefficient of range are calculated as

Concept of Coefficient of Range

	RANGE	COEFFICIENT OF RANGE
SET A (MATHEMATICS)	20-10=10	(20-10)/(20+10)=0.33
SET B (ENGLISH)	50-30 =20	(50-30)/(50+30)= 0.25

Cannot compare A & B as their base is different

Hence we use the concept of coefficient of Range

Thus there is more variations in Set A when compared to SET B

- The monthly incomes in rupee of seven employees of a firm are 5500,5750,6500,6750,7000 & 8500. Compute Coefficient of Range

The range of the income of the employees is

$$\text{Coefficient of Range} = \frac{(8500-5500)}{(8500+5500)} = \frac{3000}{14000}$$

$$= 21.42\ \%$$

Concept of Standard Deviation

https://www.mathsisfun.com/data/standard-deviation.html

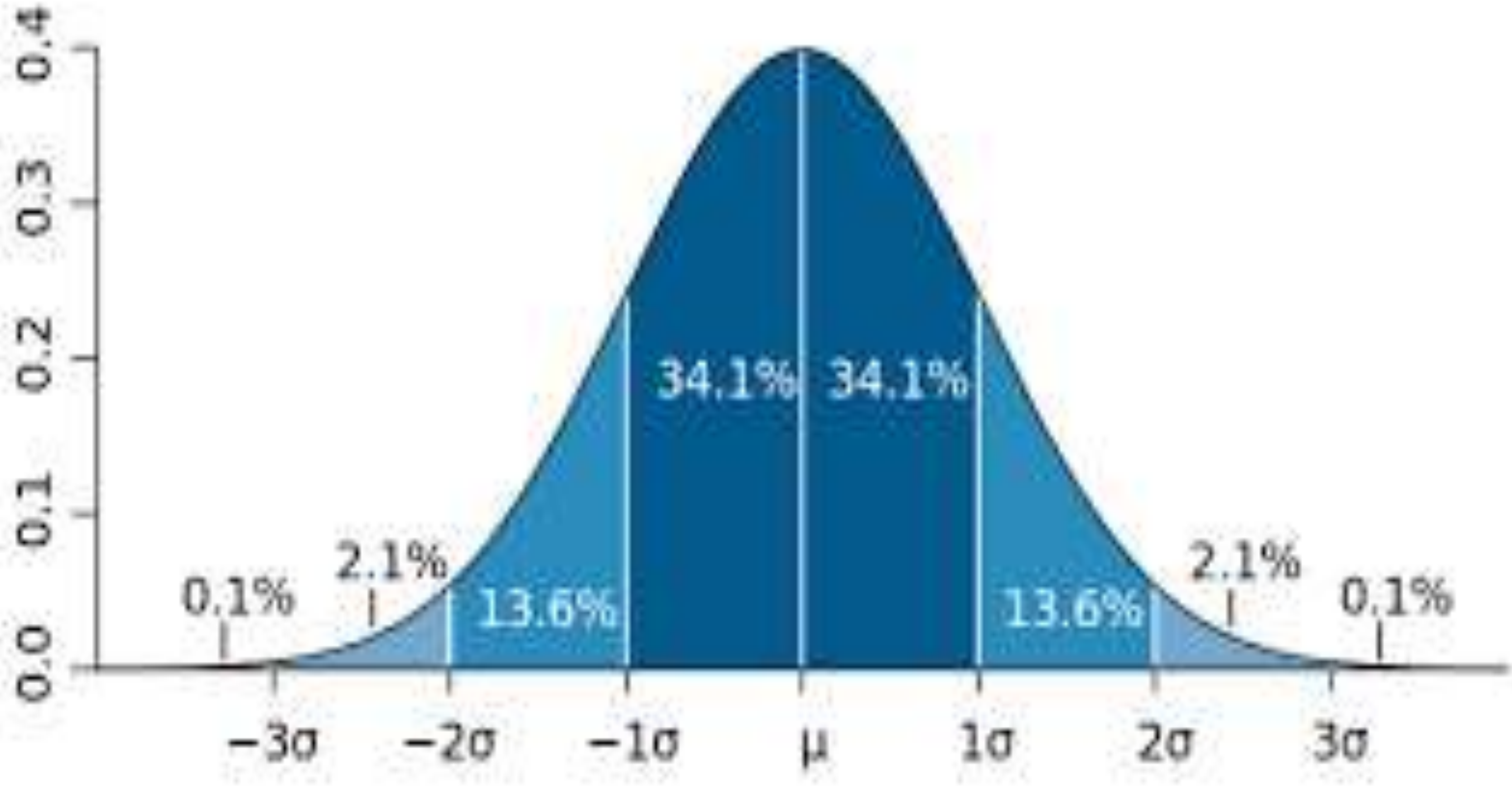
https://365datascience.com/explainer-video/distribution-in-statistics/

Concept of Standard Deviation

- Standard Deviation shows the how Data is spread out relative to the mean

- If the Data is close together the standard deviation is small

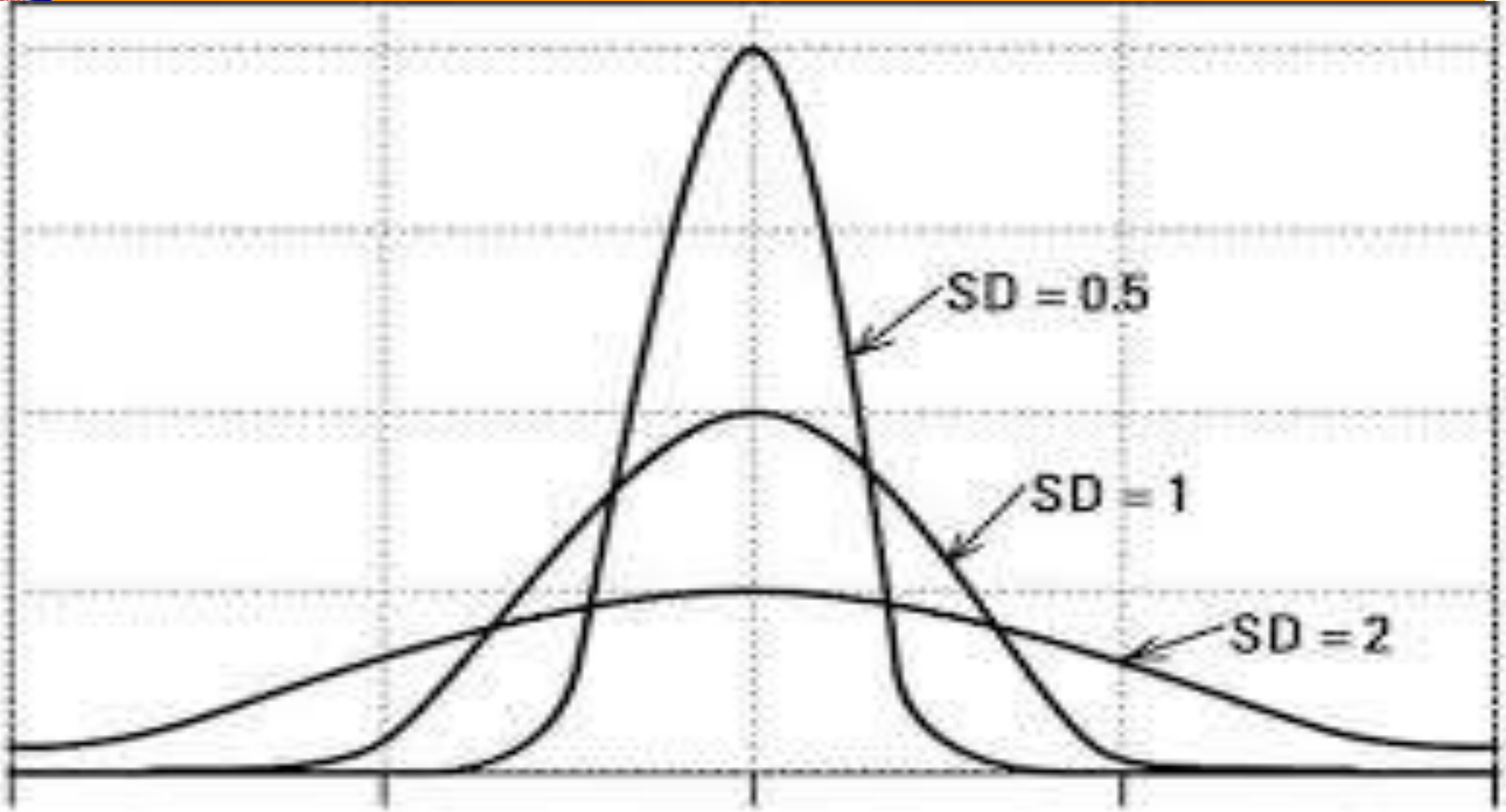
- If the data is spread out Standard deviation will be large


0.4
0.3
0.2
0.1
0.0
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ

Concept of Standard Deviation

- A standard deviation is a unit of measurement which helps you to figure out where the data items are likely to fall
- Now let us Interpret the bell curve
- 68.2 % of all the measurements in the data set fall within one standard deviation on either side of the mean
- If we take 2 standard deviations then 95.2 % of data falls within two standard deviation on either side of the mean
- If we take 3 standard deviations then 99.6 % of data falls within three standard deviation on either side of the mean

Bell curves with Various standard deviation



Standard Deviation for Ungrouped Data

- The following is the data gives the number of books taken in a school library in 7 days .Find the standard deviation

  7,9, 12,15,5,4,11

first we find out the average $= \dfrac{7+9+12+15+5+4+11}{7}$

$$= \frac{63}{7}$$

$$\bar{x} = 9$$

Standard Deviation for Ungrouped Data

- Next we calculate by using the following

x	d = x - $\overline{x}$	d	d^2
7	= 7- 9	-2	4
9	= 9 -9	0	0
12	= 12-9	3	9
15	= 15-9	6	36
5	= 5 -9	-4	16
4	= 4 - 9	-5	25
11	= 11 -9	2	4
			94

Standard Deviation for Ungrouped Data

$$\sqrt{\frac{d^2}{n}}$$

$$= \sqrt{\frac{94}{7}}$$

$$= \sqrt{13.43}$$

$$= 3.66$$

Standard Deviation for Ungrouped Data

- The below is the data of the results of Purity test on Compounds

Observed Percentage Impurity				
0.04	0.14	0.17	0.19	0.22
0.06	0.14	0.17	0.21	0.24
0.12	0.15	0.18	0.21	0.25

- Calculate the Standard Deviation.

Standard Deviation for Grouped Data

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

$$\sigma = \sqrt{\sigma^2}$$

$$\bar{x} = \frac{\Sigma (f X x)}{N}$$

Standard Deviation

Profit In Crores	No of Companies
0- 10	8
10-20	12
20-30	20
30-40	30
40 -50	20
50 -60	10

Standard Deviation

Profit (Crores)	f	Mid Value (x)	fx			
0 -10	8	5	40			
10-20	12	15	180			
20-30	20	25	500			
30-40	30	35	1050			
40-50	20	45	900			
50-60	10	55	550			
	100		3220			

$$\overline{x} = \frac{\Sigma (f X x)}{N} = \frac{3220}{100} = 32.20$$

Standard Deviation

mean is 32.20 A is

Profit (Crores)	f	Mid Value (x)	fx	Mean (μ)	x - μ	$(x - \mu)^2$
0 -10	8	5	40	32.20	- 27.20	739.84
10-20	12	15	180	32.20	- 17.20	295.84
20-30	20	25	500	32.20	- 7.20	51.84
30-40	30	35	1050	32.20	2.80	7.84
40-50	20	45	900	32.20	12.80	163.84
50-60	10	55	550	32.20	22.80	519.84
	100			32.20		1779.04

Standard Deviation

mean is 32.20 A is

Profit (Crores)	f	Mid Value (x)	fx	Mean (μ)	x - μ	(x - μ)^2	f ((x - μ)^2
0 -10	8	5	40	32.20	- 27.20	739.84	5918.72
10-20	12	15	180	32.20	- 17.20	295.84	3550.08
20-30	20	25	500	32.20	- 7.20	51.84	1036.8
30-40	30	35	1050	32.20	2.80	7.84	235.2
40-50	20	45	900	32.20	12.80	163.84	3276.8
50-60	10	55	550	32.20	22.80	519.84	5198.4
	100			32.20			19216

Standard Deviation for Grouped Data

- $\sigma^2 = \dfrac{\Sigma f(\bar{x} - \mu)^2}{N}$

$$\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$$

$$= \sqrt{\frac{19216}{100}} = \sqrt{192.16} = 13.86$$

Calculate Standard Deviation

Length of Life of the Bulb (In Hours)	No of Bulbs
550-650	10
650 -750	22
750 - 850	52
850 - 950	20
950 - 1050	16
	120

Merits of Standard Deviation

1. It is rigidly defined.
2. It is based on all observations of the distribution.
3. It is amenable to algebraic treatment.
4. It is less affected by the sampling fluctuation.
5. It is possible to calculate the combined standard deviation

Demerits of Standard Deviation

1. As compared to other measures it is difficult to compute.
2. It is affected by the extreme values.
3. It is not useful to compare two sets of data when the observations are measured in different ways.

Concept of Variance

- Mathematically It is defined as the Average of the Squared difference from the Mean

Concept of Variance

Height of Person (cms)	Mean	Difference from Mean	Squared Difference
200	150	50	2500
100	150	-50	2500
150	150	0	0
175	150	25	625
125	150	-25	625

Concept of Variance

- Interpreting from the table

Sum of the Squared Difference is 6250

Variance = 6250/n , n –no of data points

= 6250/5

= 1250

Now standard deviation is the Square root of the Variance

$$SD = \sqrt{1250} = 35.35$$

6. Coefficient of Variation

- It is the percentage ratio of Standard deviation and mean

For grouped & Ungrouped data

Formula for Calculating Coefficient for Variance

- Mean = $\frac{(\sum fx)}{n}$

$$\text{Coefficient of Variation} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

Concept of Coefficient of Variation

- Comparing two unequal sets of data

- Eg Income of White collared worker in USA vs Income of White collared worker in India

Coefficient of Variation

- Mean Score of two batsmen A & B in ten innings during a Season as Under

Find out which of the Batsmen is more consistent in scoring

	A	B
Mean Score	50	75
Standard Deviation	5	25

$CV_A = (5/50) * 100 = 10\,\%$, $CV_B = (25/75) * 100 = 33.33\,\%$

The Batsman who has lesser Coefficient of Variation is more consistent

Let us work out a Problem

- Two Brands of Tyres are Tested with the following results

  Find out which Brand is more consistent

Life (in 1000 Miles	20-25	25-30	30-35	35-40	40-45	45-50
Brand A	8	15	12	18	13	9
Brand B	6	20	32	30	12	0

Coefficient of Variation for Grouped Data

- find the coefficient of Variance of the following data for the marks obtained in a test by 80 students

Marks (x)	0-10	10-20	20-30	30-40	40-50
Frequency (f)	6	16	24	25	17

$$\bullet = \sqrt{\frac{133}{88} - \left(\frac{31}{88}\right)^2} \quad 10$$

$$= 11.77$$

$$\text{Mean} = \frac{\sum f_x}{n}$$

$$= \frac{2510}{88} = 28.52$$

Coefficient of Variation = (Standard Deviation/Mean)

= (11.77/28.52)* 100 = 41.26 % Variation

Concept of Coefficient of Variation

- The following frequency Distribution summarize the price changes on May 24, 1993, of all companies traded on the New York Stock Exchange whose names begin with L or R.

- Use their coefficients of variation to determine which distribution has less relative variability

Concept of Coefficient of Variation

Change in Price	Number of L Companies	Number of R Companies
- 1.25 to – 1.01	1	1
- 1.00 to - 0.76	1	1
- 0.75 to - 0.51	1	0
- 0.50 to -0.26	7	5
-0.25 to – 0.01	19	20
0.00	14	20
0.01 to 0.25	21	14
0.26 to 0.50	5	8
0.51 to 0.75	3	1
0.76 to 1.00	2	4
1.01 to 1.25	1	0

CONCEPT OF MEASURES OF POSITION

- These tell where a specific data value falls within the data Set or it's Relative Position in comparison with other data values

THE DIFFERENT MEASURES OF POSITION ARE:

- QUARTILES
- DECILES
- PERCENTILES

PERCENTILE

- The value below which a percentage of Data falls

eg There are 100 people and assume that you are the fourth tallest person . What does this mean is ????

Percentile = (number of people behind you/total no of people) *100

= (96/100) * 100 = 96 %

so you are in the 96 th percentile

Formula for Calculating Percentile for Ungrouped Data

- Formula is

$$P_k = \left\{ \frac{k(n+1)}{100} \right\} \text{th item}$$

k is the percentile which we want to calculate

n is the sample size

Calculating Percentile for Ungrouped Data

The following is the monthly income (in 1000) of 8 persons working in a factory. Find the 30 th percentile income

Data : 10,14,36,25,15,21,29,17

Arrange the data in ascending order

10,14,15,17,21,25,29,36

Calculating Percentile for Ungrouped Data

- $n = 8$
- Apply the formula

$$P_k = \left\{ \frac{k(n+1)}{100} \right\} \text{th item}$$

$$P_{30} = \frac{30(8+1)}{100} \text{ th item}$$

$= 2.7$ th item

$= 2$ nd item $+ 0.7$ (3 rd item – 2^{nd} Item)

Calculating Percentile for Ungrouped Data

= 14+ 0.7(15-14)

= 14 + 0.7

= 14.7

The 30 th percentile income is 14.7 (in thousands)

Final answer is = 14.7 * 1000

= 14700

Calculating Percentile for Ungrouped Data

- Let us solve a problem to find out the percentile in ungrouped data following is the height data collected from students

91,89,88,87,89,91,87,92,90,98,95,97 ,
96,100,101,96,98,99,98,100,102,99,
101,105,103,107,105,106,107,112

Find out the 10 th and 95 th percentile

Calculating Percentile for grouped Data

- Formula is

$$P_i = l + \frac{h}{f}\left\{\frac{i(n+1)}{100} - C\right]$$

P_i denotes percentile value which we want to find

l – Lower Limit of the Percentile group

Calculating Percentile for grouped Data

h – width of the Percentile Group

f – frequency of the Percentile Group

C – Cumulative Frequency before the percentile group

i – is the percentile value

n - is the number of observations

Calculating Percentile for grouped Data

- First step is calculating the percentile group
- How to we find out percentile group ???

Calculate $\frac{i(n+1)}{100}$ = Say this value as A

Compare this value with the Cumulative frequency distribution.

Calculating Percentile for grouped Data

- Compare this value with the Cumulative frequency distribution.

The cumulative frequency interval which is just highest above the value A contains the percentile group

Calculating Percentile for grouped Data

- We will work out a sum

Height (in cms)	0-5	5-10	10-15	15-20	20-25	25-30
No of Plants	18	20	36	40	26	16

Rearrange this data in the form of Frequency distribution table

Calculating Percentile for grouped Data

Class	f	cf
0-5	18	18
5-10	20	38
10-15	36	74
15-20	40	114
20-25	26	140
25-30	16	156

Calculating Percentile for grouped Data

- First find out percentile group
- $\frac{i(n+1)}{100}$ th item

Now we want to find out 61 st percentile

$$= \frac{61\ (156+1)}{100}$$

= 95.77 th item

Go to frequency and table and refer

- 15- 20 is the percentile group
- Now derive values using this group
- l = 15 , h = 5 ,f = 40 ,c = 74
- Formula is

$$P_i = l + \frac{h}{f}\left[\frac{i(n+1)}{100} - C\right]$$

$$= 15 + \frac{5}{40}\left[\frac{61(156+1)}{100} - 74\right] = \_\_\_\_ ???$$

Percentile for Grouped Data

find out the 78 th percentile

CLASS INTERVAL	FREQUENCY	CUMMULATIVE FREQUENCY
14- 16	9	9
16 - 18	13	22
18 -20	24	46
20 -22	38	84
22 - 24	16	100

QUARTILE

QUARTILES

- A ***quartile*** divides a *sorted* data set into 4 equal parts, so that each part represents ¼ of the data set

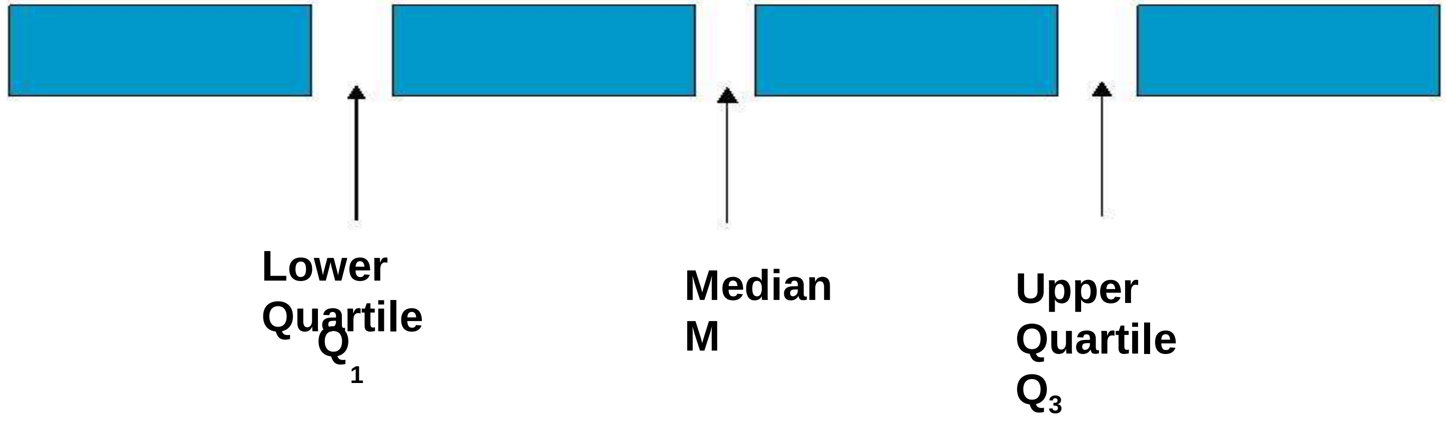
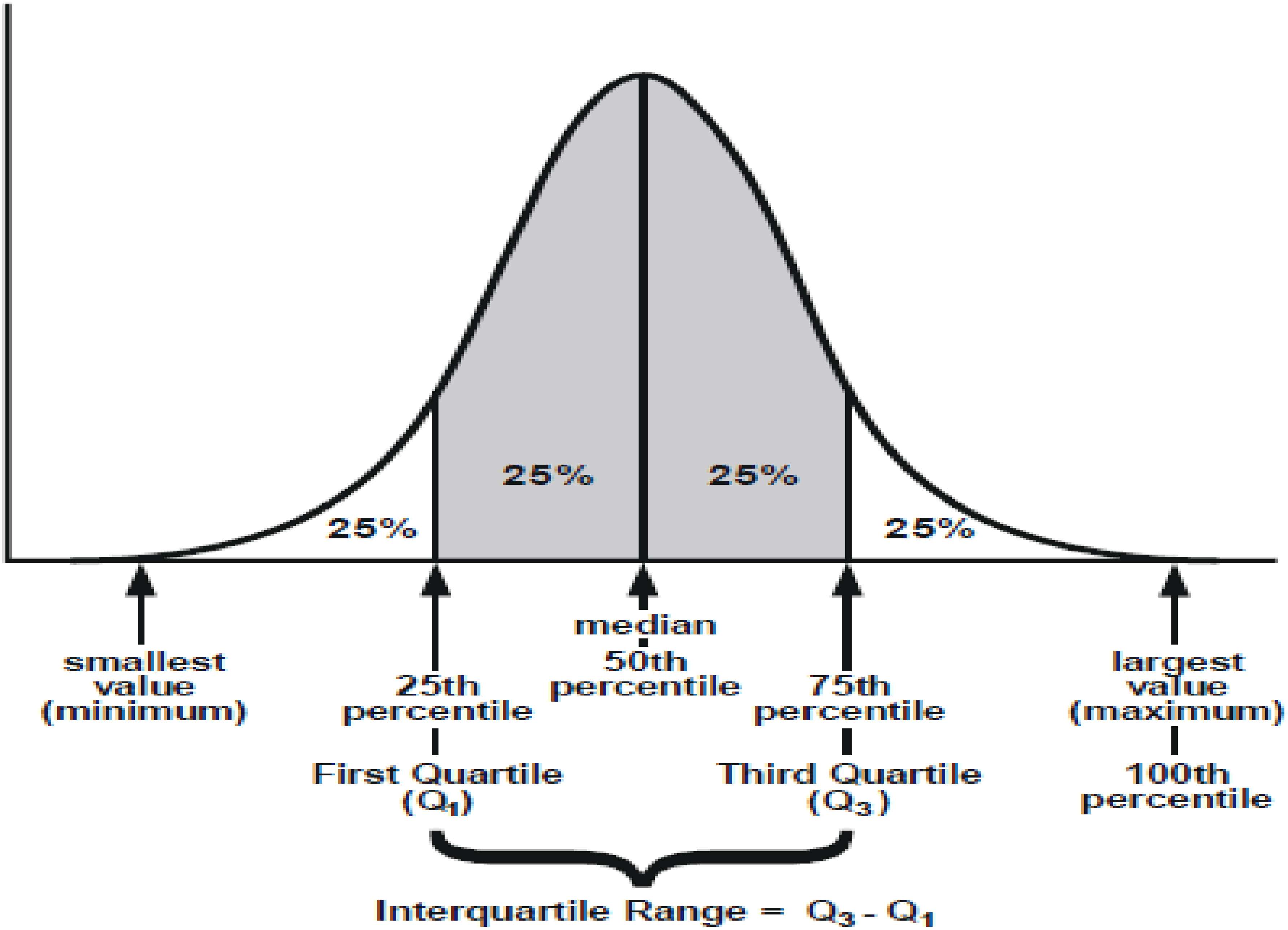


Figure 3.8
The middle half of the observations in a frequency distribution lie within the interquartile range



Quartiles

- First Quartile (Q1) separates the bottom 25 % of the sorted values from the top 75 %

- Second Quartile (Q2) separates the bottom 50 % of the sorted values from the top 50 %

- Third Quartile (Q3) separates the bottom 75 % of the sorted values from the top 25 %

Calculating Quartile for Ungrouped Data

- Considering data set having n items
- Arrange data set in ascending order

$$Q1 = \left|\frac{(n+1)}{4}\right| \text{th item}$$

$$Q2 = \left|\frac{(n+1)}{2}\right| \text{th item}$$

$$Q3 = \left|\frac{3\,(n+1)}{4}\right| \text{th item}$$

Calculating Quartile for Ungrouped Data

- Calculated quartile for the marks of 8 students in an examination given below

25,48,32,52,21,64,29,57

Calculating Quartile for Ungrouped Data

- Rearrange this data in ascending Order

  21,25,29,32,48,52,57,64

- $Q1 = \left|\frac{N+1}{4}\right|$ th Item

  $= \left|\frac{(8+1)}{4}\right|$ th Item =2.25 th Item

  $= 2^{nd}$ Item + 0.25(3 rd Item – 2 nd Item)

  = 25 + 0.25(29-25) = 26

Calculating Quartile for Ungrouped Data

- $Q2 = \left|\frac{N+1}{2}\right|$ th Item

  $= \left|\frac{(8+1)}{2}\right|$ th Item

  = 4.5 th Item

  = 4 th Item + 0.5(5 th Item – 4 th Item)

  = 32 + 0.5 (48-32)

  = 40

Calculating Quartile for Ungrouped Data

- $Q3 = 3\left|\dfrac{N+1}{4}\right|$ th Item

  $= 3\left|\dfrac{(8+1)}{2}\right|$ th Item

  $= 3(2.25)$ th Item

  $= 6.75$ th item

  $=$ 6 th Item + 0.75 (7 th Item – 6 th Item)

  $= 52 + 0.75\ (57 - 52)$

  $= 52 + 3.75$

  $= 55.75$

Calculating Quartile for Ungrouped Data

- Find out the 3 Quartiles Q1, Q2 & Q3 in this ungrouped data following is the height data collected from students .

  91,89,88,87,89,91,87,92,90,98,95,97 ,
  96,100,101,96,98,99,98,100,102,99,
  101,105,103,107,105,106,107,112

Quartiles for Discrete Series (Grouped Data)

- Below is the data of age in years of 543 members belonging to a city

AGE IN YEARS	20	30	40	50	60	70	80
NO OF MEMBERS	3	61	132	153	140	51	3

Quartiles for Discrete Series (Grouped Data)

- Rearrange this data in the form of a table

x	f	cf
20	3	3
30	61	64
40	132	196
50	153	349
60	140	489
70	51	540
80	3	543

Quartiles for Discrete Series (Grouped Data)

- $$Q1 = \left[\frac{N+1}{4}\right] \text{ th Item}$$

$$= \left[\frac{(543+1)}{4}\right] \text{ th Item}$$

$$= 136 \text{ th Item}$$

$$= 40 \text{ Years}$$

Quartiles for Discrete Series (Grouped Data)

- $Q2 = \left[\frac{N+1}{2}\right]$ th Item

  $= \left[\frac{(543+1)}{2}\right]$ th Item

  = 272 th Item

  = 50 Years

Quartiles for Discrete Series (Grouped Data)

- $Q3 = 3\left[\frac{N+1}{4}\right]$ th Item

  $= 3\left[\frac{(543+1)}{4}\right]$ th Item

  = 3 x136 th Item

  = 408 th Item

  = 60 Years

Quartiles for Discrete Series (Grouped Data)

Weight(in Kg)	40	50	60	70	80	90	100
No of Patients	15	26	12	10	8	9	5

Quartiles for Continuous Series (Grouped Data)

- $Q1 = L + \dfrac{\left(\dfrac{n}{4} - C \right)}{f} h$
- $Q2 = L + \dfrac{\left(\dfrac{n}{2} - C \right)}{f} h$
- $Q3 = L + \dfrac{\left(\dfrac{3n}{4} - C \right)}{f} h$

Quartiles for Continuous Series (Grouped Data)

- L = lower limit of the Quartile Class

  n = $\sum f$ = total no of observations in the data set

  C = Cumulative frequency in the class immediately before the Quartile class

  f = Frequency of the Quartile Class

  h = Length of the class interval of the Quartile Class

Quartiles for Continuous Series (Grouped Data)

- Calculate the Quartiles for the wages of the Labours

Wages (In Rs)	30 -32	32- 34	34- 36	36-38	38- 40	40-42	42- 44
No of Laborers	12	18	16	14	12	8	6

Quartiles for Continuous Series (Grouped Data)

- Rearranging Data in the Form of Frequency Distribution Table

x	f	cf
30-32	12	12
32-34	18	30
34-36	16	46
36-38	14	60
38-40	12	72
40-42	8	80
42-44	6	86

Quartiles for Continuous Series (Grouped Data)

- First we will calculate Q1 (Quartile 1)
- $Q1 = L + \left[\frac{\frac{n}{4} - C}{f} \right] h$

First Find Quartile Class

$Q1 = \frac{N+1}{4} = \frac{(86+1)}{4} = 21.75$

Hence 32- 34 is the Quartile Class

$= 32 + \left[\frac{\frac{86}{4} - 12}{18} \right] 2 = 33.05$

Quartiles for Continuous Series (Grouped Data)

- First we will calculate Q2 (Quartile 2)
- $$Q2 = L + \left| \frac{\frac{n}{2} - C}{f} \right| h$$

First Find Quartile Class

$$Q2 = \frac{N+1}{2} = \frac{(86+1)}{2} = 43.5$$

Hence 34- 36 is the Quartile Class

$$= 34 + \left| \frac{\frac{86}{2} - 30}{16} \right| 2 = 35.625$$

Quartiles for Continuous Series (Grouped Data)

- First we will calculate Q3 (Quartile 3)
- $Q3 = L + \left(\dfrac{\dfrac{3n}{4} - C}{f} \right) h$

First Find Quartile Class

$$Q3 = \frac{3(N+1)}{4} = \frac{3(86+1)}{4} = 3 \times 21.75 = 65.25$$

Hence 38- 40 is the Quartile Class

$$= 38 + \left(\frac{\frac{3 \times 86}{4} - 60}{12} \right) 2 = 38.75$$

RECAP

QUESTIONS???

THANK YOU