Statistics

Statistics – Theory Units

PART :I (Statistics)

1. Frequency Distribution and Measures of Central Tendency
2. Measures of Dispersion
3. Skewness and Kurtosis
4. Correlation and Regression
5. Testing of Hypothesis

UNIT 1

Frequency Distribution and Measures of Central Tendency

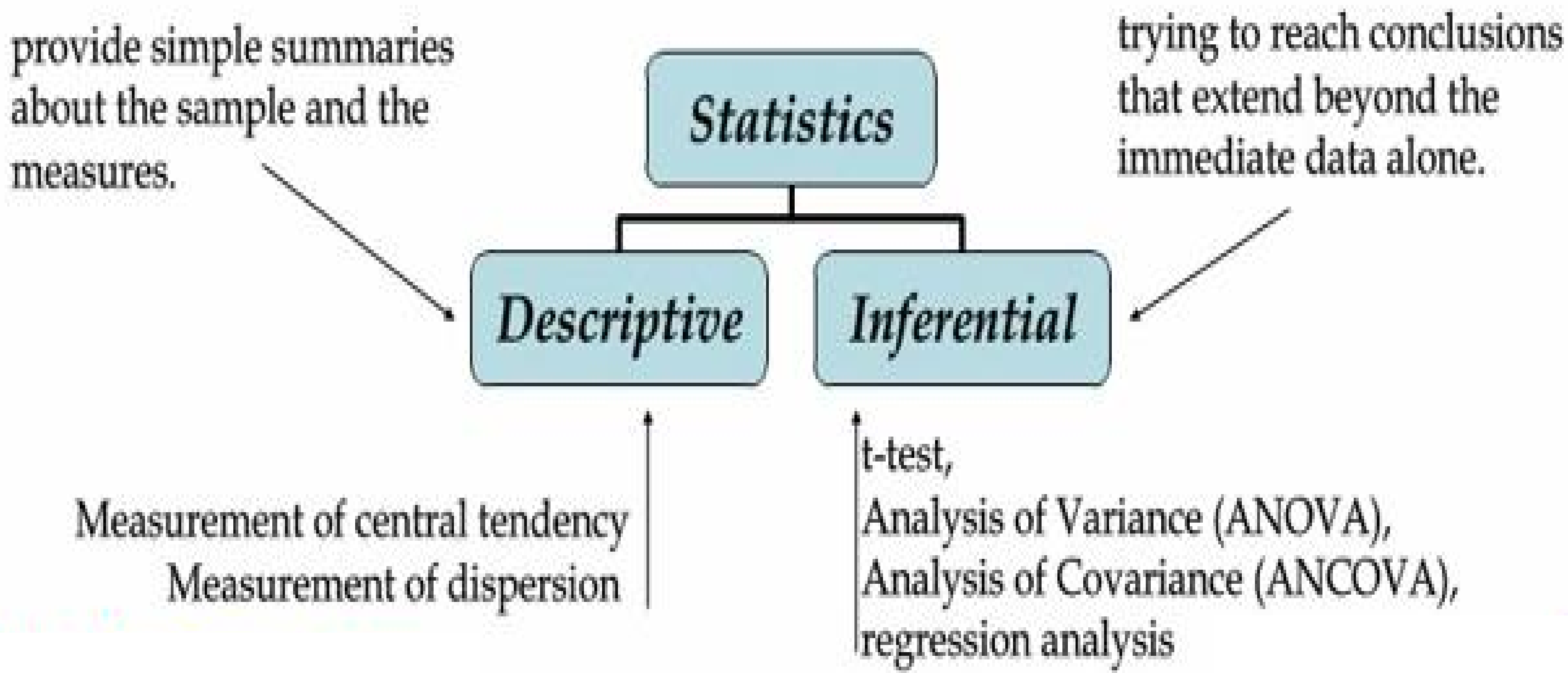
UNIT:I - Frequency Distribution and Measures of Central Tendency

- Continuous Frequency Distribution
- Histogram
- Frequency Polygon
- Mean, Median, Mode

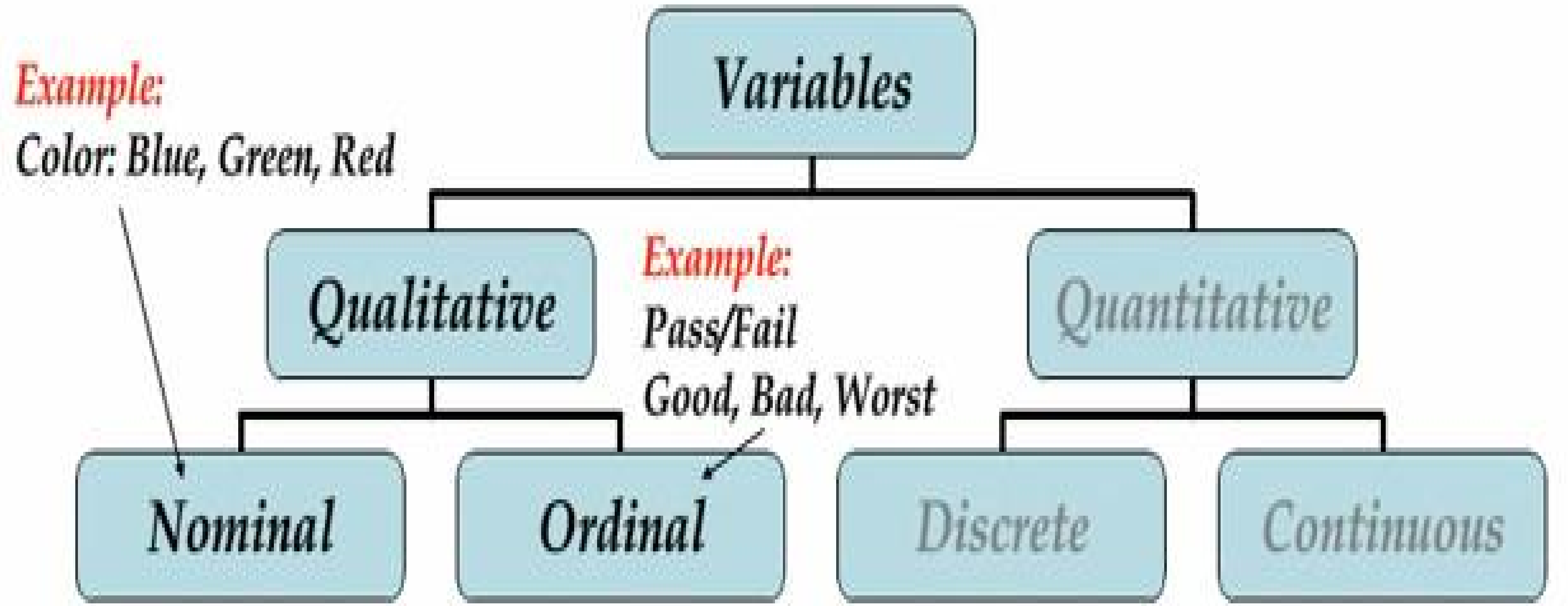
What is Statistics?

- Statistics is the science of counting
- Statistics is the science of estimates and probabilities
- The method of collecting, organizing, analyzing, and interpreting data, as well as drawing
- conclusions based on the data is called statistics.
- Methodology is divided into two main areas.
  - **Descriptive Statistics: Collecting, organizing, summarizing, and presenting data. (Use all data)**
  - **Inferential Statistics: Making generalizations about and drawing conclusions from the data collected. (Use Sample of data)**

2. Descriptive statistics and inferential statistics

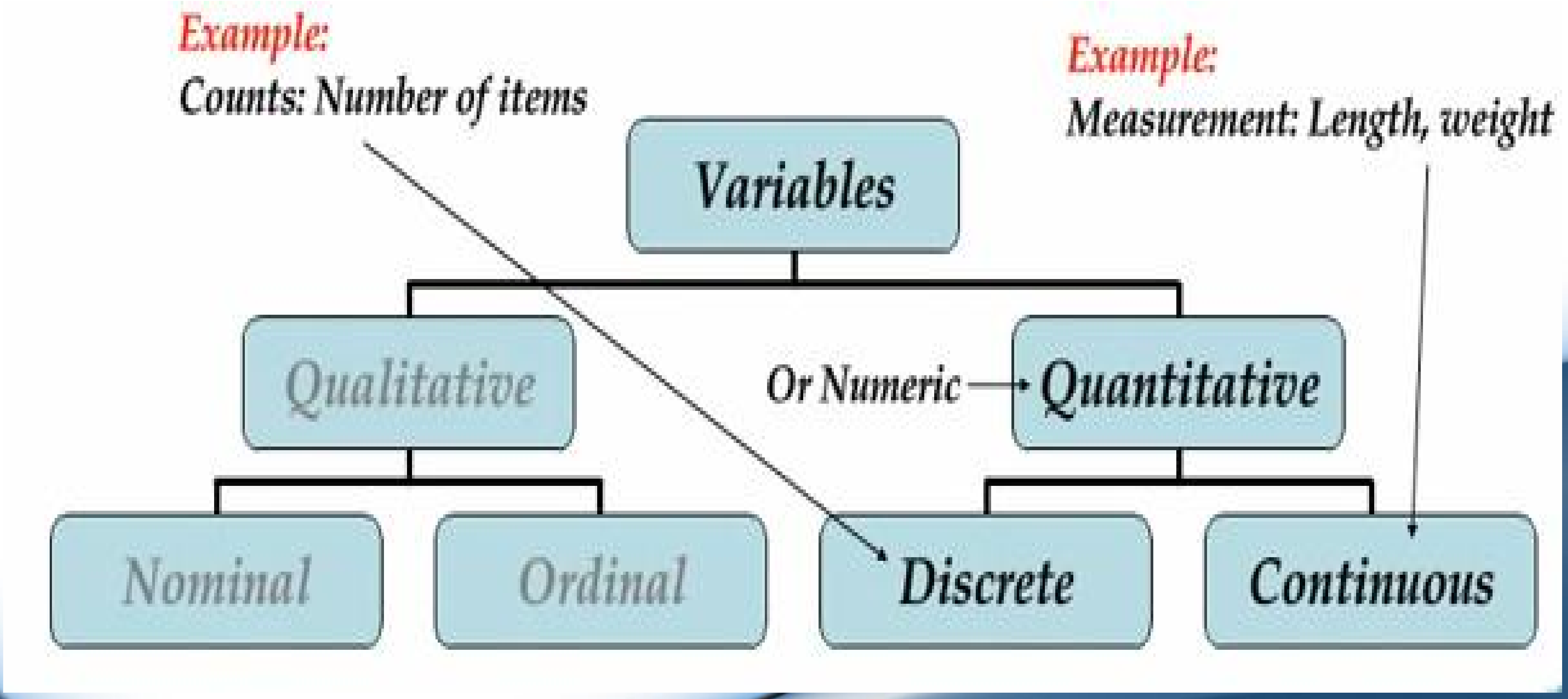


Data Types



Example:
Color: Blue, Green, Red
Variables
Qualitative
Example:
Pass/Fail
Good, Bad, Worst
Quantitative
Nominal
Ordinal
Discrete
Continuous

Data Type



Example:
Counts: Number of items
Example:
Measurement: Length, weight
Variables
Qualitative
Or Numeric
Quantitative
Nominal
Ordinal
Discrete
Continuous

Statistical terms

- **Primary/Secondary data**
- **Population**
  - complete set of individuals, objects or measurements e.g. Voting data
- **Sample**
  - a sub-set of a population (judgmental / random)e.g. Food consumption
- **Variable**
  - a characteristic which may take on different values
- **Data**
  - numbers or measurements collected
- **A parameter is a characteristic of a population**
  - e.g., the *average* height of all Indian.
- **A statistic is a characteristic of a sample**
  - e.g., the *average* height of a sample of Maharashtrians.

- A random sample is a sample obtained in such a way that every element in the population has an equal chance of being selected.

Accuracy and precision

- **Accuracy**
- Accuracy is how close a measured value is to the **actual (true) value**.

- **Precision**
- Precision is how close the measured values are **to each other**.

Example : if you are playing soccer and you always hit the left goal post instead of scoring, then you are **not** accurate, but you **are** precise!

Accuracy and precision: The target analogy

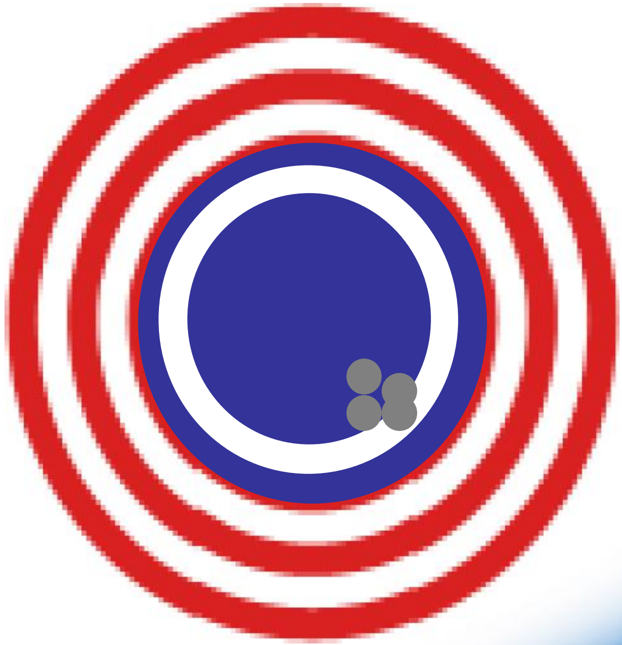
High accuracy but
low precision

High precision but
low accuracy

What does High accuracy and high precision look like?

Example: Measurement of height of a person using fixed scale.

Accuracy and precision: The target analogy



High accuracy and high precision

Objectives

- Organize data using frequency distributions.
- Represent data in frequency distributions graphically using histograms, frequency polygons and ogives.

Arithmetic mean

- Is the average
- Data →1 1 2 3 4
- Mean= (1+1+2+3+4)/5 =
- 11/5=2.2

Median

- Positional value – middle
- Step 1 : Order the list
- Step 2: Find the middle number
- Data →100 100 200 300 400
- Median= 200
- Data -> 100 100 200 300 400 500
- Median= (200+300) / 2
- =250

Mode

- It is the most common number or most frequent number
- Data →1 1 2 3 4
- Mode= 1

Example

- Find the mean ,median and mode for the following set

a) 30,50,20,60,50,90,50,20,80,60

- Mean=50.1
- Median= 50
- Mode=50

b) 51.6, 48.7,50.3, 49.5, 48.9

- Mean=49.8
- Median= 49.5
- Mode= nonexistent

Frequency Distribution

- After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.

- This is the goal of descriptive statistical techniques.

- One method for simplifying and organizing data is to construct a **frequency distribution**.

Frequency Distributions

- When data are collected in original form, they are called raw data.

- When the raw data is organized into a frequency distribution, the frequency will be the number of values in a specific class of the distribution.

Frequency Distributions

- A frequency distribution is the organizing of raw data in table form, using classes and frequencies.
- Three Types of Frequency Distributions
  - Categorical frequency distributions
  - Ungrouped frequency distributions
  - Grouped frequency distributions

Three Types of Frequency Distributions

- Categorical frequency distributions - can be used for data that can be placed in specific categories, such as nominal- or ordinal-level data.
- Examples - political affiliation, religious affiliation, blood type etc.

Blood Type Frequency Distribution - Example

Class	Frequency	Percent
A	5	20
B	7	28
O	9	36
AB	4	16

Ungrouped Frequency Distributions

- Ungrouped frequency distributions - can be used for data that can be enumerated and when the range of values in the data set is not large.
- Examples - number of miles your instructors have to travel from home to campus, number of girls in a 4-child family etc.

Number of Miles Traveled Example

Class	Frequency
5	24
10	16
15	10

Grouped Frequency Distributions

- Grouped frequency distributions - can be used when the range of values in the data set is very large. The data must be grouped into classes that are more than one unit in width.
- Examples - the life of boat batteries in hours.

Lifetimes of Boat Batteries - Example

Frequency Table

Class limits	Class Boundaries	Frequency	Cumulative frequency
24 - 37	23.5 - 37.5	4	4
38 - 51	37.5 - 51.5	14	18
52 - 65	51.5 - 65.5	7	25

Terms In Grouped Frequency Distribution

- Class limits represent the smallest and largest data values that can be included in a class.
- In the lifetimes of boat batteries example, the values 24 and 37 of the first class are the class limits.
- The lower class limit is 24 and the upper class limit is 37.

Terms In Grouped Frequency Distribution

- The class boundaries are used to separate the classes so that there are no gaps in the frequency distribution.
- The class mark is the midpoint of the class interval. It is obtained by adding the lower and upper class limits and dividing by 2.

Terms Associated with a Grouped Frequency Distribution

- The class width for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class minus the lower (or upper) class limit of the previous class.
- e.g 37.5 - 23.5 =14

Example

Rating	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Lower Class Limits

Rating	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Upper Class Limits

Rating	Frequency
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Class
Midpoints

class mark

Rating	Frequency
0 - 1 2	20
3 - 4 5	14
6 - 7 8	15
9 - 10 11	2
12 - 13 14	1

Grouped Frequency Distribution - Example

- In a survey of 20 patients who smoked, the following data were obtained. Each value represents the number of cigarettes the patient smoked per day. Construct a frequency distribution using six classes. (The data is given on the next slide.)

Grouped Frequency Distribution - Example

10	8	6	14
22	13	17	19
11	9	18	14
13	12	15	15
5	11	16	11

Grouped Frequency Distribution - Example

- Step 1: Find the highest and lowest

  Values: $H = 22$ and $L = 5$.
- Step 2: Find the range:

  $R = H - L = 22 - 5 = 17.$
- Step 3: Select the number of classes

  desired. In this case it is equal to 6.

Grouped Frequency Distribution - Example

- Step 4: Find the class width by dividing the range by the number of classes. Width = 17/6 = 2.83. This value is rounded up to 3.

Grouped Frequency Distribution - Example

- Step 5: Select a starting point for the lowest class limit. For convenience, this value is chosen to be 5, the smallest data value. The lower class limits will be 5, 8, 11, 14, 17 and 20.

Grouped Frequency Distribution - Example

- Step 6: The upper class limits will be 7, 10, 13, 16, 19 and 22. For example, the upper limit for the first class is computed as 8 - 1, etc.

Grouped Frequency Distribution - Example

- Step 7: Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to the upper class limit.

Grouped Frequency Distribution - Example

- Step 8: Tally the data, write the numerical values for the tallies in the frequency column and find the cumulative frequencies.
- The grouped frequency distribution is shown on the next slide.

Note: The dash “-” represents “to”.

Frequency Table

Class Limits	*Class Boundaries*	*Frequency*	*Cumulative Frequency*
05 to 07	4.5 - 7.5	2	2
08 to 10	7.5 - 10.5	3	5
11 to 13	10.5 - 13.5	6	11
14 to 16	13.5 - 16.5	5	16
17 to 19	16.5 - 19.5	3	19
20 to 22	19.5 - 22.5	1	20

Example: Leaves

Alex measured the lengths of leaves on the oak tree (to the nearest cm):
9,16,13,7,8,4,18,10,17,18,9,12,5,9,9,16,1,8,17,1, 10, 5,9,11,15,6,14,9,1,12,5,16,4,16,8,15,14,17

Solution:

In order the lengths are:
1,1,1,4,4,5,5,5,6,7,8,8,8,9,9,9,9,9,9,10,10,11,12,12,13,14,14,15,15,16,16,16,16,17,17,17,18,18

The smallest value (the "minimum") is **1 cm**

The largest value (the "maximum") is **18 cm**

The range is 18−1 = **17 cm**

Grouped Frequency Distribution - Example

- Let us say we want about 5 groups.
- Divide the range by 5:
- 17/5 = 3.4
- Then round that up to **4**.

Start Value

– Pick a starting value that is less than or equal to the smallest value

- Starting at 0 and with a group size of 4 we get: **0, 4, 8, 12, 16**
- Write down the groups, include the end value of each group (must be less than the next group):
- The last group goes to 19, which is greater than the largest value, so that is good.

Length (cm)	Frequency
0-3	
4-7	
8-11	
12-15	
16-19	

- **Upper and Lower Values For Each Group**
  - Even though Alex only measured in whole numbers, the data is continuous, so "4 cm" means the actual value could have been anywhere from 3.5 cm to 4.5 cm. Alex just rounded the numbers to whole centimeters.

Length	Lower/Upper	Frequency
0-3 cm	0-3.5	
4-7 cm	3.5-7.5	
8-11 cm	7.5-11.5	
12-15 cm	11.5-15.5	
16-19 cm	15.5-19.5	

- **Example: Leaves (continued)**
- Now tally the results to find the frequencies. And do a total as well.

1,1,1,4,4,5,5,5,6,7,8,8,8,9,9,9,9,9,9,10,10,11,12,12,13,14,14,15,15,16,16,16,16,17,17,17,18,18:

Length	Lower/Upper	Mid Point	Frequency	Cumulative frequency
0-3 cm	0-3.5	1.5	3	3
4-7 cm	3.5-7.5	5.5	7	10
8-11 cm	7.5-11.5	9.5	12	22
12-15 cm	11.5-15.5	13.5	7	29
16-19 cm	15.5-19.5	17.5	9	38
	Total:		**38**	

Do it Yourself

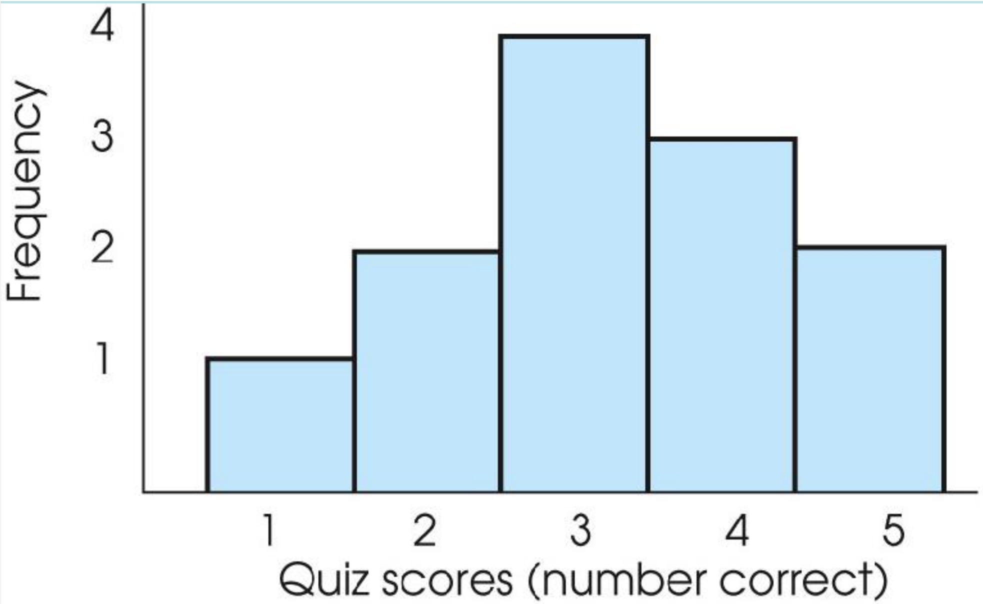
Scores on a 60 question exam for 20 students

50, 46, 58, 49, 50, 57, 49, 48, 53, 45, 50, 55, 43, 49, 46, 48, 44, 56, 57, 44 using 5 classes

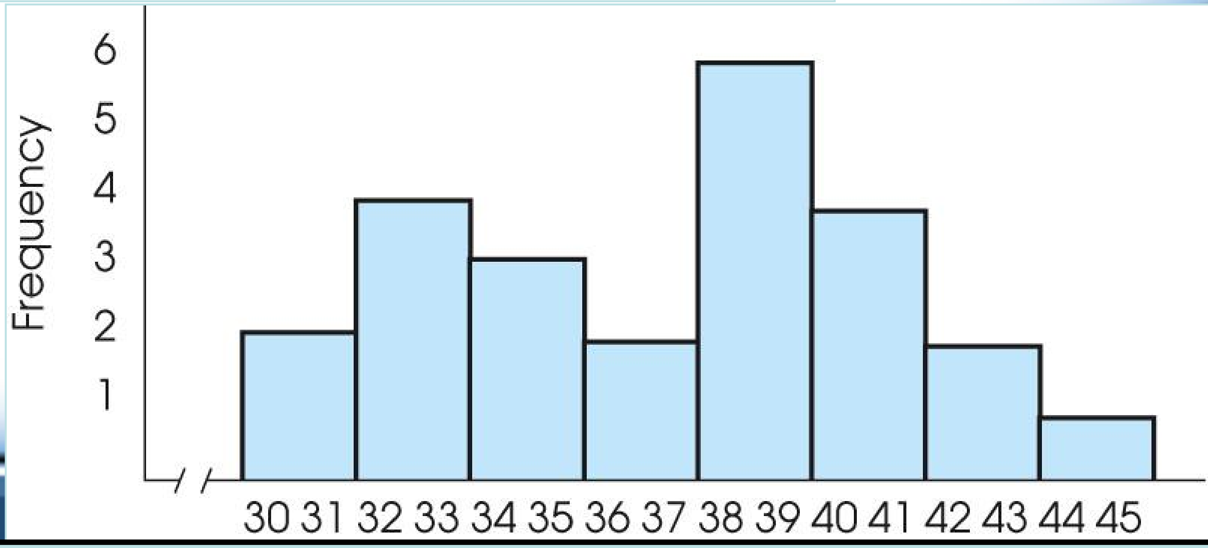
Graphical representation of Frequency Distribution

Histograms

- **Histogram** for a frequency distribution based on quantitative data is very similar to the bar chart showing the distribution of qualitative data.
- The classes are marked on the horizontal axis and the class frequencies on the vertical axis.
- The class frequencies are represented by the heights of the bars.



X	f
5	2
4	3
3	4
2	2
1	1



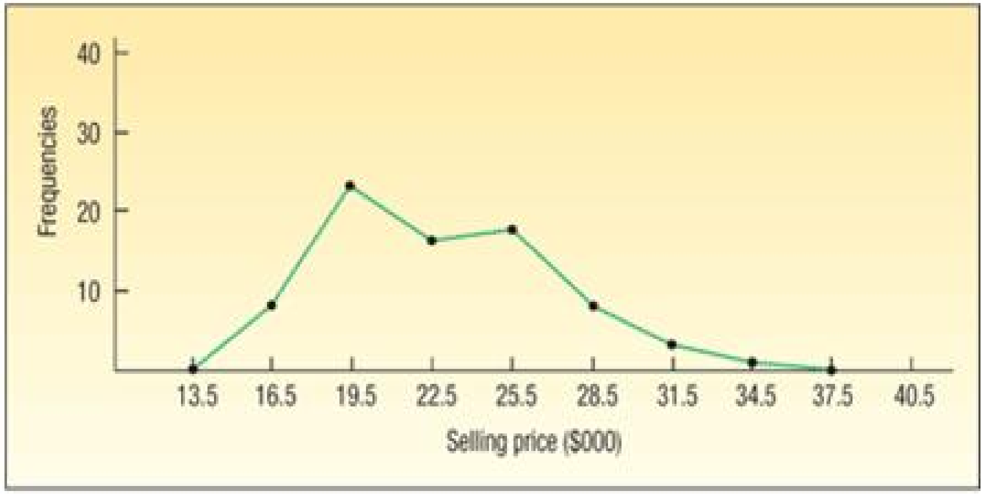
X	f
44–45	1
42–43	2
40–41	4
38–39	6
36–37	2
34–35	3
32–33	4
30–31	2

Frequency Polygon

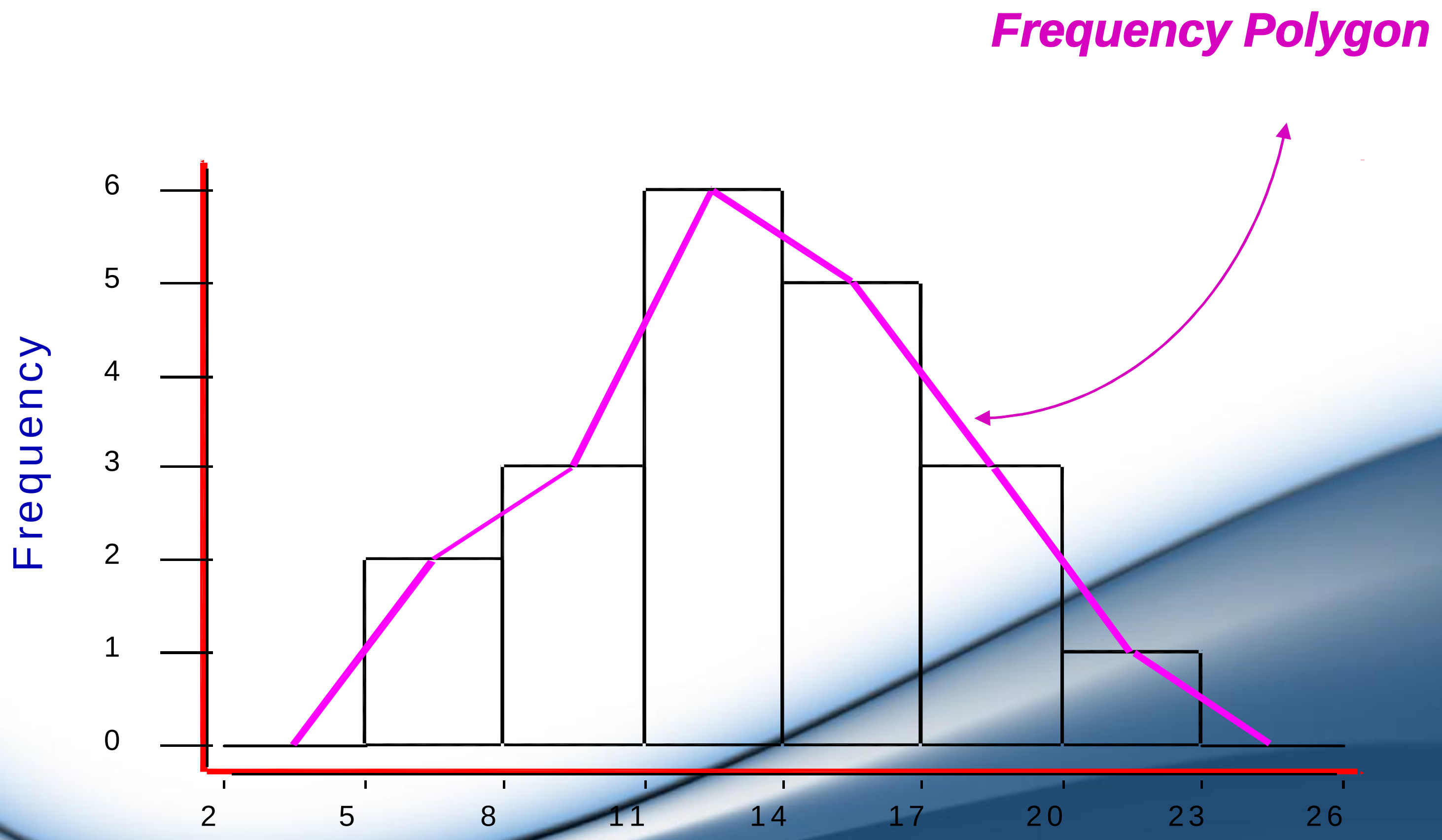
- A **frequency polygon** also shows the shape of a distribution and is similar to a histogram.

- It consists of line segments connecting the points formed by the intersections of the class midpoints and the class frequencies.

Selling Price ($ thousands)	Midpoint	Frequency
15 up to 18	16.5	8
18 up to 21	19.5	23
21 up to 24	22.5	17
24 up to 27	25.5	18
27 up to 30	28.5	8
30 up to 33	31.5	4
33 up to 36	34.5	2
Total		80

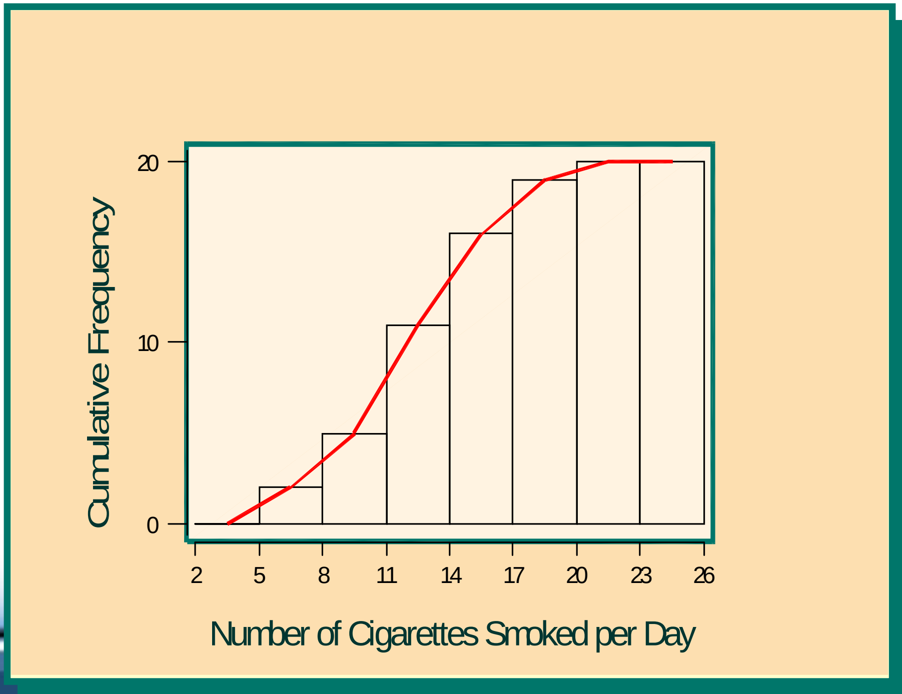


Frequency Polygon



Cumulative Frequency Graph or Ogive

- A cumulative frequency graph or Ogive is a graph that represents the cumulative frequencies for the classes in a frequency distribution.



Example

- Draw the histogram, frequency polygon and ogive for the following data

Weekly wages	20-40	40-60	60-80	80-100	100-120	120-140	140-160	160-180	180-200
No. of employees	8	12	20	30	40	35	18	7	5

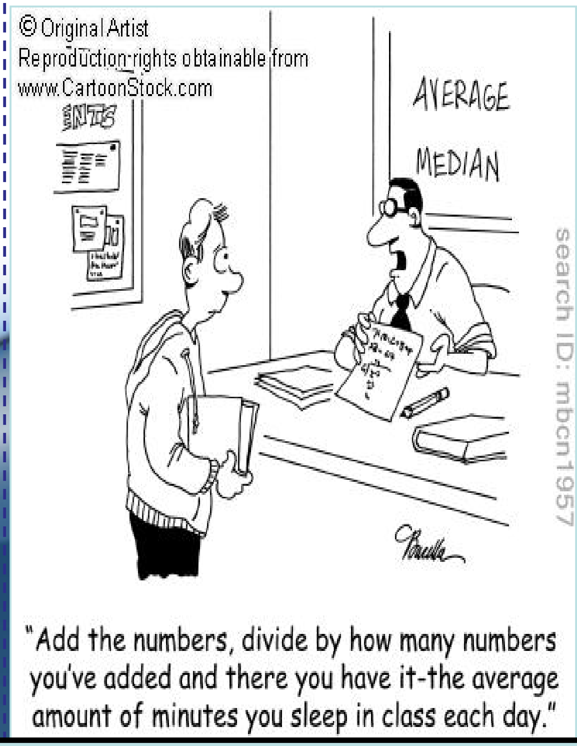
Measures of Central Tendency

Measures of Central Tendency

Measures of central tendency are numerical descriptive measures which indicate or locate the center of a distribution or data set.

In layman's term, a measure of central tendency is an AVERAGE. It is a single number of value which can be considered typical in a set of data as a whole.

For example, in a class of 40 students, the average height would be the typical height of the members of this class as a whole.



MEAN (Average)

The **MEAN** of a set of values or measurements is the sum of all the measurements divided by the number of measurements in the set.

Among the three measures of central tendency, the mean is the most popular and widely used. It is sometimes called the arithmetic mean.

If we compute the mean of the population, we call it the parametric or population mean, denoted by μ (read "mu").

If we get the mean of the sample, we call it the sample mean and it is denoted by $\bar{x}$ (read "x bar").

Properties of Mean

1. Mean can be calculated for any set of numerical data, so it always exists.
2. A set of numerical data has one and only one mean.
3. Mean is the most reliable measure of central tendency since it takes into account every item in the set of data.
4. It is greatly affected by extreme or deviant values (*outliers*)

Arithmetic mean

- Arithmetic mean of a set of observations is their sum divided by the number of observations.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

- In case of frequency distribution $x_i \mid w_i$,i=1,2,3…..n where w_i is the frequency of the variable x_i

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

which means:

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}.$$

Example

- Sargodha's Temperature Last week was (Celc):

  38, 42, 35, 39, 42, 44, 36

  find the mean temperature using A.M.

$$\text{A.M.} = \frac{\sum x}{n}$$

$$= \frac{38+42+35+39+40+43+36}{7}$$

$$= 39$$

Hence mean temperature of the week is **39**

Direct Method:

- It is the simplest method to find the A.M. value.

$$A.M. = \bar{x} = \frac{\sum fx}{\sum f}$$

where: x = given values f = frequency of groups

- **Example**

Ages (Years)	13	14	15	16	17
No. of Students	2	5	13	7	3

Ages (years) x	Number of students f	fx
13	2	26
14	5	70
15	13	195
16	7	112
17	3	51
Total	$\sum f = 30$	$\sum fx = 454$

Solution:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{454}{30} = 15.13\text{ y}$$

Examples

- Find the arithmetic mean of the numbers 8,3,5,12,10
- Find the arithmetic mean of the following frequency distribution:

x:	1	2	3	4	5	6	7
f:	5	9	12	17	14	10	6

- Calculate the arithmetic mean from the following table:

Marks	No. of Students
0-10	12
10-20	18
20-30	27
30-40	20
40-50	17
50-60	06

Arithmetic mean

- Find the arithmetic mean of the numbers 8,3,5,12,10
- Is the average
- Data → 8, 3, 5, 12, 10
- Mean= (8+3+5+12+10)/5 = 38/5 =7.6
- Inference is
  - – All numbers are closest to 7.6

- Find the arithmetic mean of the following frequency distribution:

x:	1	2	3	4	5	6	7
f:	5	9	12	17	14	10	6

		Direct
Marks	*f*	*fx*
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	73	299

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

$\bar{x} =$ 299/73 = 4.09

Calculate the arithmetic mean from the following table:

Marks	No. of Students
0-10	12
10-20	18
20-30	27
30-40	20
40-50	17
50-60	06

			Direct
Marks	*f*	*Mid point x*	*fx*
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	06	55	330
Total (Σ)	**100**		2800

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

$\bar{x} =$ 2800/100 = 28

Step deviation method to calculate Arithmetic mean from **ungrouped** data

- If the values of *x* or *f* are large, the calculation of mean is quite time consuming and tedious.
- The mean is given by-

$$\text{Mean } \overline{X} = A + [(\sum f_i d_i)/N]$$

Where A= Assumed mean

N= Number of items (Total Frequency)

f_i= frequency

$d_i = (x_i - A)$

deviation of the given value from the assumed mean

Step deviation method to calculate Arithmetic mean from **grouped** data

- If the values of *x* or *f* are large, the calculation of mean is quite time consuming and tedious.
- The mean is given by-

$$\text{Mean } \overline{X} = A + h[(\sum f_i d_i)/N]$$

Where A= Assumed mean

h= Class size

N= Number of items (Total Frequency)

f_i= frequency

d_i= $(x_i - A)/h$

deviation of the given value from the assumed mean

Steps to calculate AM

Mean X = $A + h[(\sum f_i d_i)/N]$

Step1 : compute $d=(x_i - A)/h$

Step2: multiply d by f_i .

Step3: find sum of df_i

Step4: divide result of step 3 by N.

Step 5: Add A to value of step 4.

The result gives the arithmetic mean of the distribution.

A.M. of Grouped Data

- Grouped data is data that has been organized into groups known as classes.
- Formula of A.M.

Method	**Formula**
Direct Method	
Short-Cut Method	
Step Deviation Method	

Short-Cut Method

$$\bar{x} = A + \frac{\sum fD}{\sum f}$$

Where

A = Assumed Mean

f = Frequency of different groups

D = Deviation form of **A**

$$D = (X - A)$$

Step-Deviation Method

$$\bar{x} = A + \frac{\sum fu}{\sum f} \times h$$

Where

A = Assumed Mean

f = Frequency of different groups

u = Step Deviation

$$u = \frac{x - A}{h}$$

h = Size of Class interval

Example of Short-Cut and Step Deviation

The following frequency distribution showing the marks obtained by 50 students in statistics at a certain college. Find the arithmetic mean

Marks	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80-89**
Frequency	1	5	12	15	9	6	2

			Direct	Short-Cut Method		Step-Deviation	
Marks	*f*	*x*	*fx*	*D=x-A*	*fD*	$u=\frac{x-A}{h}$	*fu*
20 – 29	1	24.5	24.5	-30	-30	-3	-3
30 – 39	5	34.5	172.5	-20	-100	-2	-10
40 – 49	12	44.5	534.5	-10	-120	-1	-12
50 – 59	15	54.5	817.5	0	0	0	0
60 – 69	9	64.5	580.5	10	90	1	9
70 – 79	6	74.5	447.5	20	120	2	12
80 – 89	2	84.5	169.5	30	60	3	6
Total (Σ)	**50**		$\sum fx = 2745$		$\sum fD = 20$		$\sum fu = 2$

Where:

A = 54.5

h = 10

$\sum f = 50 \quad \sum fx = 2745 \quad \sum fD = 20 \quad \sum fu = 2$

Direct Method:

$$\overline{x} = \frac{\sum fx}{\sum f} = \frac{2750}{50} = \mathbf{54.9 \cong 55}$$

Short-Cut Method:

$$\overline{x} = A + \frac{\sum fD}{\sum f} \qquad \textit{where } A = 54.5$$

$$= 54.5 + \frac{20}{50} = 54.5 + 0.4 = \mathbf{54.9}$$

Step-Deviation Method:

$$\overline{x} = A + \frac{\sum fu}{\sum f} \times h \qquad \textit{where } A = 54.5 \;\&\; h = 10$$

$$= 54.5 + \frac{2}{50} \times 10 = 54.5 + 0.4 \times 10$$

$$= \mathbf{54.9}$$

Do it Yourself

- Calculate the arithmetic mean from the following table: (Assume Arbitratary point)

Marks	No. of Students
0-8	8
8-16	7
16-24	16
24-32	24
32-40	15
40-48	7

Properties of Arithmetic Mean

- Algebraic sum of the deviations of a set of values from their arithmetic mean is zero.

- The sum of the squares of the deviations of a set of values is minimum when taken about mean.
- Mean of the Combined Series
  - If n_1 and n_2 are the sizes and $\overline{X}_1$ and $\overline{X}_2$ are the respective means of two series then the mean of the combined series of size n_1+n_2 is given by:

$$\text{Mean} = \frac{(n_1\overline{X}_1 + n_2\overline{X}_2)}{n_1 + n_2}$$

Do it Yourself

- Find the A.M. of following deviations:

  25, 30, 20, 63, 52, 29, 18, 8, 41   Ans: 31.77

- The following data shows distance covered by persons to perform their routine jobs.   Ans :24

Distance (Km)	1-10	11-20	21-30	31-40
No. of Persons	10	20	40	30

- Marks of Stats subject obtained by student in mid exams:
- Ans:27.7

Marks	1-10	11-20	21-30	31-40	41-50
Students	**5**	**4**	**35**	**17**	**6**

- Scores of Statistics Final Exam of

83	72	72	100	73
85	98	70	100	80
77	77	74	100	89
90	98	83	80	78
91	86	89	77	90
88	75	91	74	86
100	91	86	85	89
77	96	78	98	93
71	88	75	84	91
96	71	85	90	93

Examples

- The mean salary paid to 1000 employees of an establishment was found to be Rs. 180.40. Later on, after disbursement of salary, it was discovered that the salary of two employees was wrongly entered as Rs. 297 and 165. Their correct salaries were Rs. 197 and 185. Find the correct Arithmetic mean.
- The mean of marks in statistics of 100 students in a class was 72. The mean of marks of boys was 75, while their number was 70. Find the mean of girls in the class.

Median

The median of a set of numbers arranged in order of magnitude is either the middle value or the arithmetic mean of the two middle value

MEDIAN

The **MEDIAN,** denoted Md, is the middle value of the sample when the data are ranked in order according to size.

16 17 18 19 20 21 22

↑

Median

16 17 18 19 20 21 22 23

↑

Median

$$\frac{19 + 20}{2} = 19.5$$

Arrange values from
least to greatest.

63 73 84 86 88 95 97 97 100

Find the number that is in the middle.

The median is **88.**

Half the numbers are less than the median.

Half the numbers are greater than the median.

Calculation of Median –Discrete series :

i.Arrange the data in ascending or descending order.

ii.Calculate the cumulative frequencies.

iii.Apply the formula.

$$Median(M) = Size\ of \left(\frac{N+1}{2}\right) th\ item$$

6.72 3.46 3.60 6.44

3.46 3.60 6.44 6.72

(even number of values)

no exact middle -- shared by two numbers

$$\frac{3.60 + 6.44}{2}$$

MEDIAN is 5.02

6.72 3.46 3.60 6.44 26.70

3.46 3.60 6.44 6.72 26.70

(in order - odd number of values)

exact middle

MEDIAN is 6.44

Median

- Example:-
  - Find the median for the set of numbers 4,5,4,3,8,6,8,10,8
  - Find the median for the set of numbers 12,5,9,11,5,18,15,7
- In case of ungrouped frequency distribution median is obtained by following steps:-
  - Find N/2,where $N=\sum f$
  - See the cumulative frequency just greater than N/2
  - The corresponding value of x is median

Obtain the median for the following:-

x:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
f:	**8**	**10**	**11**	**16**	**20**	**25**	**15**	**9**	**6**

Obtain the median for the following frequency distribution:

x	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

X	f	Cf
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	Total 120	

Median = N/2 = 120/2

= 60

The cumulative frequency just greter than N/2 that is 65 so the x correspoinding to 65 is 5

Calculation of median – Continuous series

For calculation of median in a **continuous frequency distribution** the following formula will be employed. Algebraically,

$$Median(M) = L1 + \frac{\frac{N}{2} - cf}{f} \times i$$

Median

- For grouped data, the median is given by-

$$Median(M) = L1 + \frac{\frac{N}{2} - cf}{f} \times i$$

where: L1 = lower class boundary of the median class

(the class corresponding cumulative frequency just greater than N/2)

N = number of items in the data

cf = sum of the frequencies of all class lower than the median class

f = frequency of the median class

I =size of the median class interval

Example: Median of a set Grouped Data in a Distribution of Respondents by age

Age Group	Frequency of Median class(f)	Cumulative frequencies(cf)
0-20	15	15
20-40	32	cf->47
L1→40-60	f->54	101
60-80	30	131
80-100	19	150
Total	N->150	

Class Width I = 20

$$\text{Median (M)} = 40 + \frac{\frac{150}{2} - 47}{54} \times 20$$

$$= 40 + \frac{75-47}{54} \times 20$$

$$= 40 + \frac{28}{54} X 20$$

= 40+0.52X20

= 40+10.37

= 50.37

Do it Yourself

- Find the median wage of the following distribution:

Wages (in Rs.)	2,000-3,000	3,000-4,000	4,000-5,000	5,000-6,000	6,000-7,000
No. of workers	3	5	20	10	5

- An incomplete frequency distribution is given as follows:

Class-interval	0-10	10-20	20-30	30-40	40-50	50-60	Total
f	10	?	25	30	?	10	100

Given that the median value is 30, N=100. Determine the missing frequencies, using the median formula.

Do it Yourself

- Find the mean of the following distribution

% of ash content	3 – 3.9	4 – 4.9	5 – 5.9	6 – 6.9	7 – 7.9	8 – 8.9	9 – 9.9	10 – 10.9	11-11.9
Frequency	1	7	28	78	84	45	28	7	2

- Find the median wage of the following distribution:

Wages (in Rs.)	2,000-3,000	3,000-4,000	4,000-5,000	5,000-6,000	6,000-7,000
No. of workers	3	5	20	10	5

Wages	frequency	cf
2000-3000	3	3
3000-4000	5	8
4000-5000	20	28
5000-6000	10	38
6000-7000	5	43
	43	

$$Median(M) = L1 + \frac{\frac{N}{2} - cf}{f} \times i$$

$$=4000+ \frac{(43/2 - 8)}{20} *1000$$

$= 4000+675$

$= 4675$

- An incomplete frequency distribution is given as follows:

  Given that the median value is 30, N=100. Determine the missing frequencies, using the median formula.

Class-interval	0-10	10-20	20-30	30-40	40-50	50-60	Total
f	10	?	25	30	?	10	100

Class Interval	frequency
0-10	10
10-20	f1
20-30	25
30-40	30
40-50	f2
50-60	10
Total	100

$$Median(M) = L1 + \frac{\frac{N}{2} - cf}{f} \times i$$

$$30 = 30 + \frac{(100/2 - cf) * 10}{30}$$

cf = 10+f1+25

50 = 10+f1+25

f1=15

f1+f2+75=100

f2=10

Mode

- The mode of a set of numbers is that value which occurs with the greatest frequency
- It is the most common value
- The mode may not exist, and even if it does it may not be unique.
- Mode is the value which has the greatest frequency density in its immediate neighborhood.

a. 5 5 5 3 1 5 1 4 3 5 ⇦ Mode is 5

b. 1 2 2 2 3 4 5 6 6 6 7 9 ⇦ Bimodal - 2 and 6

c. 1 2 3 6 7 8 9 10 ⇦ No Mode

Example

- Find the mode of the following distribution:

Size (x)	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f)	3	8	15	23	35	40	32	28	20	45	14	6

Mode

- From a frequency distribution or histogram the mode can be obtained from the formula

$$\text{Mode} = L_1 + \left(\frac{f_1 - f_0}{(f_1 - f_0) + (f_1 - f_2)} \, h \right)$$

where L_1 = lower class boundary of the modal class (class that has maximum of all the frequencies)

f_1 = Frequency of the modal Class

f_0 = Frequency of the class preceding the modal class

f_2 = Frequency of the class succeeding the modal class

h = size of the modal class interval

Example: Calculate Mode for the distribution of monthly rent Paid by Libraries in Karnataka

Monthly rent (Rs)	Number of Libraries (f)
500-1000	5
1000-1500	10
1500-2000	8
2000-2500	16
2500-3000	14
3000 & Above	12
Total	65

$$Z = L_1 + \frac{f1 - f0}{2f1 - f0 - f2} \times i$$

$$Z = 2000 + \frac{16-8}{2(16)-8-14} \times 500$$

$$Z = 2000 + \frac{8}{32-8-14} X500$$

$$Z = 2000 + \frac{8}{10} \times 500$$

$$Z = 2000 + 0.8 \times 500 = 400$$

Z=2400

Do it Yourself

Modal class

- Find the mode for the following distribution:

Class-interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	5	8	7	12	28	20	10	10

- The median and mode of the following wage distribution are known to be Rs. 3350 and Rs. 3400 respectively. Find the values of f3, f4 and f5

Wages	0-1000	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000	Total
No. of Employees	4	16	f3	f4	f5	6	4	230

Relation between Mean, Median, Mode

$$Mode = 3\ Median - 2\ Mean$$

Do it Yourself

- In a frequency distribution of 100 families given below, the number of families corresponding to expenditure group 20 – 40 and 60 – 80 are missing from the table. However the median is known to be 50. Find out the missing frequencies

Expenditure	**0 - 20**	**20 - 40**	**40- - 60**	**60 - 80**	**80 - 100**
No. of families	14	?	27	?	15

Do it Yourself

- Find mean, median and mode of the following data

Marks	0 - 20	20 - 40	40- - 60	60 - 80	80 - 100
No. of students	15	8	15	16	6

Do it Yourself

- Find mean, median and mode of the following data

Marks	0 - 20	20 - 40	40- - 60	60 - 80	80 - 100
No. of students	15	8	15	16	6

Statistics
Measures of Dispersion

- Player1 – 48,49,50,50,50,50,51,52
  - Mean=50
  - Median=50
  - Mode=50
- Player2 – 0,25,26,50,50,99,100
  - Mean=50
  - Median=50
  - Mode=50

Player 1 is consistent player but Player2 is not.

Thus mean, median and mode do not give a clear picture

Purpose of Measuring Dispersion

- A measure of dispersion appears to serve two purposes.
- First, it is one of the most important quantities used to characterize a frequency distribution.
- Second, it affords a basis of comparison between two or more frequency distributions.
- The study of dispersion bears its importance from the fact that various distributions may have exactly the same averages, but substantial differences in their variability.

Dispersion

- Measures of central tendency are inadequate to give us a complete idea of the distribution.
- They give us an idea about the concentration about the central part of the distribution.
- They must be supported by some other measures
- One such measure is Dispersion ("scatteredness")
- The degree to which numerical data tend to spread about an average value .

Measures of Dispersion

- Range
  - Percentile range
  - Quartile deviation
  - Mean deviation
  - Variance and standard deviation

- Relative measure of dispersion
  - Coefficient of variation
  - Coefficient of mean deviation
  - Coefficient of range
  - Coefficient of quartile deviation

Range

- The simplest and crudest measure of dispersion is the range. This is defined as the difference between the largest and the smallest values in the distribution. If $x_1, x_2, \ldots\ldots\ldots, x_n$ are the values of observations in a sample, then range (R) of the variable X is given by:

$$R\left(x_1, x_2, \ldots\ldots\ldots, x_n\right) = \max\left\{x_1, x_2, \ldots\ldots\ldots\ldots, x_n\right\} - \min\left(x_1, x_2, \ldots\ldots\ldots\ldots, x_n\right\}$$

Range

- The range of a set of numbers is the difference between the largest and smallest numbers in the set.
- Player1 – 48,49,50,50,50,50,51,52
  - Range= 52-48=4
- Player2 – 0,25,26,50,50,99,100
  - Range=100-0=100
- Not a reliable measure

Quartile Deviation

- Quartiles :Three points that divide the series into four equal parts.
- Deciles : nine points that divide the series into ten equal parts
- Percentiles : ninety nine points that divide the series into hundred equal parts.

Quartile

- First Quartile (Q1) : $\frac{N}{4}$
- Second Quartile (Q2) : $\frac{N}{2}$
- Third Quartile (Q3) : $\frac{3N}{4}$

Quartile Deviation

- Quartile deviation is given by

$$Q=(Q_3-Q_1)/2$$

where Q_1 and Q_3 are the first and third quartiles of the distribution frequency

- (Q_3-Q_1) is known as Inter- Quartile Range

- Quartiles tell us how compressed the data is about the median
- Gives us measure of spread.
- Find Quartiles for:
  - Sample A : 4,6,10,14,15,16,17,17,18,20,20
  - Sample B : 2,3,5,5,7,9,12,13,15,18,18,21

Quartile Deviation following data:

x	0	1	2	3	4	5	6	7	8	
f	1	9	26	59	72	52	29	7	1	
c.f	1	10	36	95	167	219	248	255	256	

– N= 256/2=128. cumulative frequency just greater than 128 is 167. thus Median=Q2=4

– N/4=64. c.f. just greater than 64 is 95. thus Q1=3

– N*(3/4)=192. c.f. just greater than 192 is 219. thus Q3=5

- Find the Inter-quartile range & Quartile deviation of the following:

Class-Interval	0-15	15-30	30-45	45-60	60-75	75-90	90-105
f	8	26	30	45	20	17	4

x	F	CF
0-15	8	8
15-30	26	34
30-45	30	64
45-60	45	109
60-75	20	129
75-90	17	146
90-105	4	150
	150	

Q1 -> N/4 = 150/4 = 37.5

Q1=Median= L1+ (N/4 – cf) *I/f
= 30 + (37.5 -34)* 15/30
= 31.75

Q2 -> N/2 = 150/2 = 75
Q2=Median= L1+ (N/2 – cf) *I/f
= 45+(75-64) *15/45
=48.66

Q3->3N/4 = 3*150/4 =112.5
Q3=Median= L1+ (3N/4 – cf) *I /f
=60+(112.5-109)*15/20
=62.62

Quartile Deviation = 1/2(Q3-Q1) =1/2 (62.62-31.75) = 15.43

Mean Deviation

- Mean deviation= $\dfrac{\sum |x_i - \bar{x}|}{N}$
- Mean deviation in case of frequency distribution or grouped frequency distribution:

$$M.D = \frac{\sum f_i |x_i - \bar{x}|}{N}$$

Calculate the mean deviation for the following raw data 15, 20, 17, 19, 21, 13, 12, 10, 17, 9, 12

Mean = ∑x/n = (**15 + 20 + 17 + 19 + 21 + 13 + 12 + 10 + 17 + 9+ 12)/ 11 = 15**

M.D = ∑(xi-x)/n = 38/11=3.455

X	X-M	|X-M|
15	0	0
20	5	5
17	2	2
19	4	4
21	6	6
13	-2	2
12	-3	3
10	-5	5
17	2	2
9	-6	6
12	-3	3
		38

Example

- Calculate Quartile deviation and Mean deviation from mean and median, for the following data: (May 2014)

Marks	No. of students
0-10	6
10-20	5
20-30	8
30-40	15
40-50	7
50-60	6
60-70	3

		Short-Cut Method					
Marks	*f*	*X*	*D=(x-A)/10*	*fD*	$\|X-\bar{X}\|$	$f\|X-\bar{X}\|$	*Cf*
0 – 10	6	5	-3	-18	28.4	170.4	6
10 – 20	5	15	-2	-10	18.4	92.0	11
20 – 30	8	25	-1	-8	8.4	67.2	19
30 – 40	15	35	0	0	1.6	24.0	34
40 – 50	7	45	1	7	11.6	81.2	41
50 – 60	6	55	2	12	21.6	129.6	47
60 – 70	3	65	3	9	31.6	94.8	50
Total (Σ)	**50**			-8		659.2	

$\bar{x}$= A+fd/N * I
=35+(-8)*10/50
= 33.4

Q1 -> N/4 = 50/4 = 12.25 - > 20+10/8(12.25-11) -> 22.875
Q3 -> 3N/4 = 150/4 = 37.5 - > 40+10/7(37.5-34) -> 45
Q.D = ½ (Q3-Q1) = ½(45-22.875) = 11.56

M.D = 1/N (∑f($X-\bar{X}$) = 1/50 (659.2) = 13.184

Standard Deviation:

- Standard deviation measures the absolute dispersion or variability of a distribution.

- A small standard deviation means high degree of uniformity in the observation as well as homogeneity of the series

- Standard deviation is the positive square root of the average of squared deviation taken from arithmetic mean.
- The standard deviation is given by the :-

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n}(X_i - \bar{X})^2}$$

$$\pm = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

- Calculate the standard deviation of the following observations on a certain variable:
- 10,20,30,40

$$\pm = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

$\sum x^2 = 10^2+20^2+30^2+40^2$
$= 100+400+900+1600$
$= 3000$

$(\sum x)^2 = (10+20+30+40)^2$
$= (100)^2$
$= 10000$

$$\sigma = \sqrt{\frac{3000}{4} - \frac{10000}{16}}$$

$= 11.18$

Standard Deviation

$$\sigma = \sqrt{\frac{\sum f(x-\overline{x})^2}{\sum f}}$$

$$\pm = i^2 \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Examples

- Calculate the standard deviation and mean of the following table

Age in years	No. of members
20 – 30	3
30 – 40	61
40 - 50	132
50 – 60	153
60 – 70	140
70 – 80	51
80 – 90	2

Age Group	f	*X*	D=(x-A)/10	fD	fD^2
20 – 30	3	25	-3	-9	27
30 – 40	61	35	-2	-122	244
40 – 50	132	45	-1	-132	132
50 – 60	153	55	0	0	0
60 – 70	140	65	1	140	140
70 - 80	51	75	2	102	204
80 – 90	2	85	3	6	18
Total (Σ)	**542**			**-15**	765

$$\pm = i^2 \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$\sigma = 100 * \sqrt{\frac{765}{542} - \frac{-15^2}{542^2}}$$

=11.88 years

Mean=A+ ∑fd *h/N
= 55+(-15*10)/542
= 54.72

Do it Yourself

Calculate the Mean deviation and Standard deviation for the following distribution.(May 2008, May 2009)

No of cold Experience in 12 months	0	1	2	3	4	5	6	7	8	9
No of persons	**15**	**46**	**91**	**162**	**110**	**95**	**82**	**26**	**13**	**2**

Standard Deviation of Combined mean

- n_i= size
- X_i=mean
- σ_i = S.D.

- For k series , S.D. of combined series is :

Square of combined S.D

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3 \ldots\ldots + n_k\bar{X}_k}{n_1 + n_2 + n_3 + .. + n_k}$$

$$\sigma^2 = \frac{n_1(\sigma_1^{\,2} + d_1^{\,2}) + n_2(\sigma_2^{\,2} + d_2^{\,2}) + \ldots\ldots + n_k(\sigma_k^{\,2} + d_k^{\,2})}{n_1 + n_2 + n_3 + .. + n_k}$$

where "σ" is S.D. for each series

where $d_i = \bar{X}_i - \bar{X}$

n1

n2

The means of two samples of sizes 50 and 100 respectively are 54.1 and 50.3 and the S.D. are 8 and 7. Obtain the S.D of the sample of size 150 obtained by combing the two samples.

n_1+n_2

σ 1

σ 2

$\overline{X1}$

$\overline{X2}$

Combined mean=51.57
Std. Deviation= 7.5625

respectively are 54.1 and 50.3 and the S.D. are 8 and 7. Obtain the S.D of the sample of size 150 obtained by combing the two samples.

Combine Mean = $n1\overline{x1} + n1\overline{x2}$ / (n1+n2)

= 50 *54.1 + 50.3*100 / (100+50)

= 51.57

d1= x1-x = 54.1-51.57= 2.53

d2 =x2-x =50.3-51.57 = -1.27

Combine std σ = $\sqrt{n1(\sigma1^2 + d1^2) + n2(\sigma2^2 + d2^2) / (n1+n2)}$

= $\sqrt{50 (64+6.4)+100(49+1.61)/150}$

= 7.56

- For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores 43 and 35 were misread as 34 and 53 respectively. Find the corrected mean and standard deviation to the corrected figures.

n=200

$\overline{X}=40=(\sum x_i)/n$ → incorrect Mean $\sum x_i = 8000$

SD=15

Correct mean = (8000-34-53+43+35)/200 = 7991/200 =**39.995**

$SD^2 = (\sum x^2)/n - (\sum x/n)^2$

$SD^2 = (\sum x^2)/n - \overline{X}^2$

$\sum x^2 = (SD^2 + \overline{X}^2) * n$

$\sum x^2 = 200(15^2 + 40^2) = 365000$

Corrected $\sum x^2 = 365000-(34)^2-(53)^2+(43)^2+(35)^2 = 364109$

$sd^2 = 364109/200 - 39.995^2 = 224.14$ sd= **14.97**

The first of the two samples has 100 items with mean 15 and S.D 3. If the whole group has 250 items with mean 15.6 and S.D $(13.44)^{½}$, find the S.D of the second group

n1+n2= 250

Combined mean = 15.6

Combined SD = $13.44^{½}$

Combined mean ,15.6 = [(100*15) + (150 * X_2)]/250

$X_2 = 16$

d1= x1-x= 15-15.6 =-0.6

d2= x2-x=16-15.6

Combine std σ = n1($σ1^2$ + $d1^2$) + n2 ($σ2^2$ + $d2^2$) / (n1+n2)

$(13.44)^{½}$ = 100(9+0.36) + 150($σ2^2$ + 0.16**) /250**

150 $σ_2^2$ = 250*13.44-100*9.36-150*0.16 = 2400

$σ_2^2 = 16$

$σ_2 = 4$

Variance

- The square of the standard deviation is called the variance

$$\pm^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N} \right)^2$$

Coefficient of Dispersion:

- Used to compare the variability of two series which differ widely in their averages .
- Used to compare if they are measured in different units.
- Coefficients are independent of the units of measurement.

- Coefficient of Variance is Relative measurement of Dispersion.

- Coefficient of variation is given as: Ratio of sample standard deviation to sample mean multiplied by 100.

$$C.V = 100 * \frac{\sigma}{\bar{x}}$$

- The series having lesser C.V. is said to be more consistent (homogeneous) than the other.

Examples

- The following is the record goals scored by team A in a football season.

No of goal scored by Team A	0	1	2	3	4
No of matches played in month	1	9	7	5	3

For Team B the average number of goals scored per match was 2.5 with a S.D of 1.25 goals. Find which team may be considered as more consistent.

No of goals(x)	***Match played(f)***	***Fx***	***fx²***
0	1	0	0
1	9	9	9
2	7	14	28
3	5	15	45
4	3	12	48
	25	**50**	**130**

$$\pm = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

$$\pm = \sqrt{\frac{130}{25} - \left(\frac{50}{25}\right)^2}$$

$$=1.09$$

Coefficient of Team A = 100 * S.D/Mean = 1.09 /2 = 54.5%

Coefficient of Team B = 100 * S.D/Mean = 1.25/2.5 = 50%

Team	**Mean**	**SD**	**C.V**
A	2	1.09	54.5
B	2.5	1.25	50

Team B is better in average scores as well as consistency of performance because coefficient of variance of team B is lower.

- Coefficient of Range = $\frac{x_{\max} - x_{min}}{x_{max} + x_{min}}$
- Coefficient based on Quartile = $\frac{Q3 - Q1}{Q3 + Q1}$
- Coefficient of Mean deviation = $\frac{MD_x}{\bar{x}}$
- Coefficient of Std. Deviation= $\frac{\sigma}{\bar{x}}$
- Coefficient of variation = $100 * \frac{\sigma}{\bar{x}}$

Do it Yourself

1. Calculate the coefficient of variation for the following data [may 2009]

Daily wages	**30-35**	**35-40**	**40-45**	**45-50**	**50-55**	**55-60**	**60-65**	**65-70**	**70-75**	**75-80**
No of worker	**17**	**27**	**42**	**61**	**72**	**65**	**47**	**34**	**22**	**13**

THANK YOU