بسم الله الرحمن الرحيم

السلام عليكم ورحمة الله وبركاته

Biostatistics

L II

4th July 2022

PROF. DR. WAQAR AL-KUBAISY

This include:

Presentation of data by

1. Graph and or

2. Tables

3. Calculation or numerical summaries, such as Frequency, Average, Mean, Median, Mode Percentages

Descriptive statistics

Biostatistics consist of

1-Collection of data .

2-Presentation of data

3-.Estimation of data

Presentation of Data
table
graph, chart or
Numerical Description

Graphical Techniques

- some times table presentation will give some difficulties to the reader especially to non numerical readers

- Picture speaks lauder than thousand words .

- Graph have powerful impact on the imagination of population .

- Relationships, Trends and Contrasts are often more
- readily appreciated from diagram than table ..

An important thing is the type of the variable concerned.

Nominal and Ordinal Data

Charting

Pie Chart

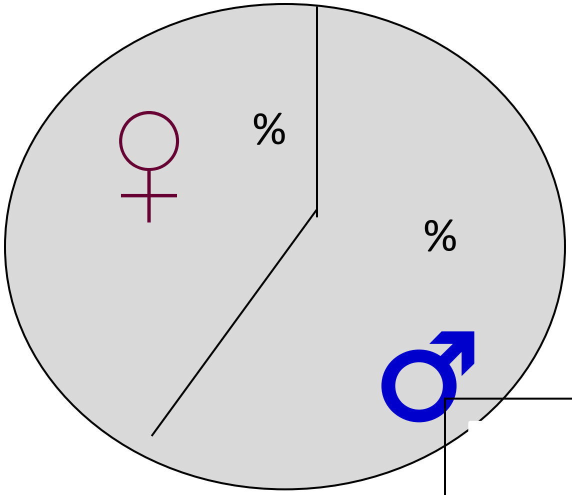
Here the circular is divided into sectors, pie shaped pieces

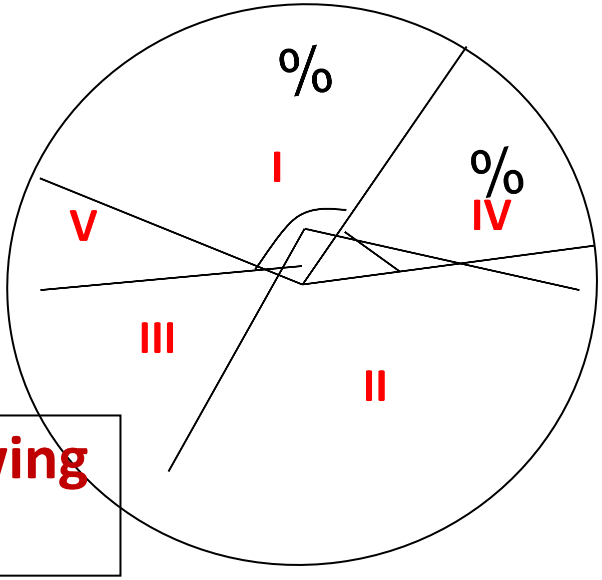
Size of pie proportional to frequency, percentage of that variable.

Disadvantage of pie chart

it can only represented one variable

(sex of children





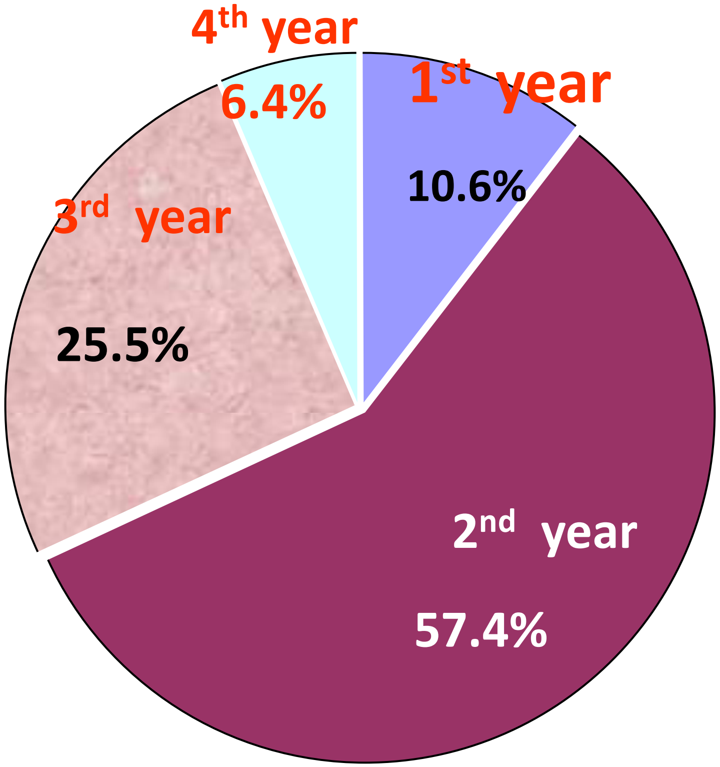
in showing comparison

Pie Charts

- Displays data in percentages.
- Statistics Class Data:
  - 5: 1st year, 10.6%
  - 27: 2nd year, 57.4%
  - 12: 3rd year, 25.5%
  - 3: 4th year, 6.4%
- Should add to 100%, adds to 99.9% due to round-off error

Excellent in showing part vs. whole comparisons

Percentage of students in each class level in a Statistics class



2- THE BAR CHART:

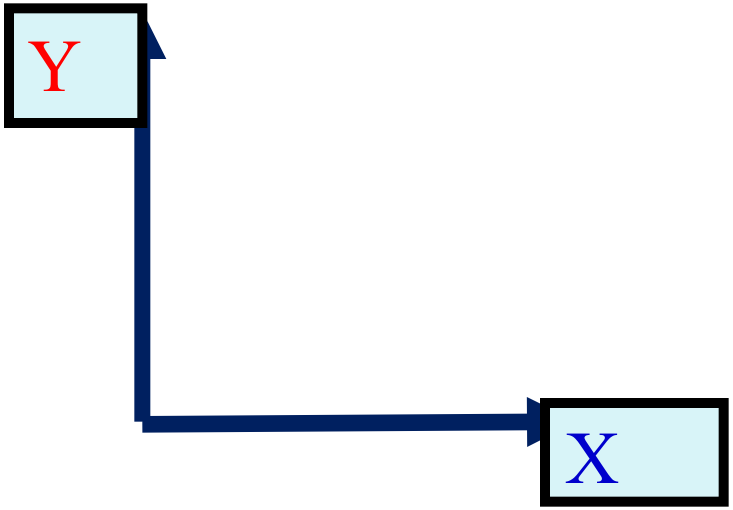
- This type of graph is suitable to represent data of the two subtypes of qualitative and quantitative discrete type.
- Each category in the table is represented by a bar or column or rectangle,
- So the height of the bar is opposite to the corresponding frequency on the Y axis.

- All bars must have the same width and a space must be left between every two consecutive bars,
- the width of that space is about same or half the width of the bar.

nominal and ordinal data

Bar Chart

Two axis



- **Horizontal,** X
- **plotting the variable .**

- **Vertical,** Y
- **plotting the**
- **frequency, Relative frequency or %**
- **Then draw** a **Rectangles (bar) .**

The length of rectangle (bar) corresponding to the frequency of the variable

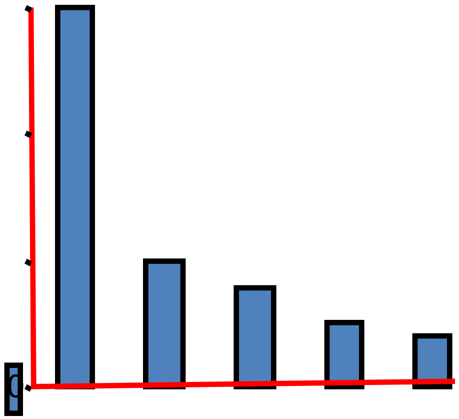
Used for

- **frequency or**
- **Relative frequency or**
- **% .**

Charting

nominal and ordinal data

Bar chart

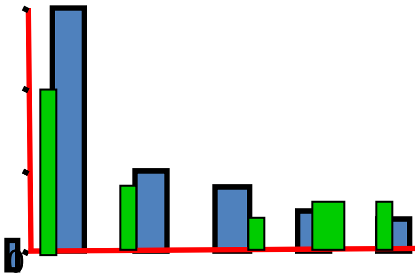


I. Simple bar chart used

-when we have one variable (sex of child)

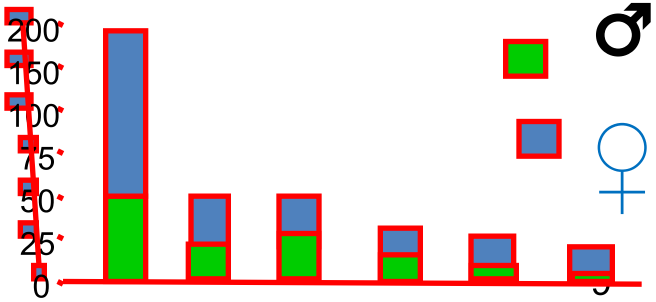
-width of bares should be equal and

-space between bars be the same



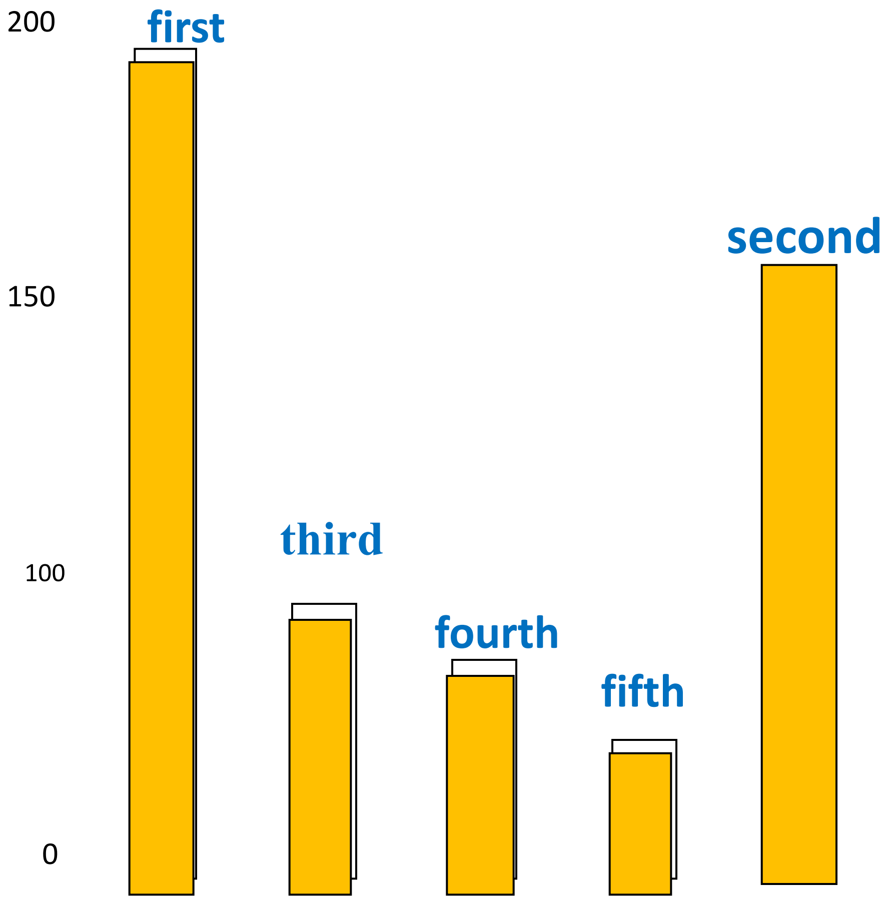
II Clustered bar chart

Used when more than one variable example sex with different class year



III Stacked bar chart

nominal and ordinal data



Excellent for showing

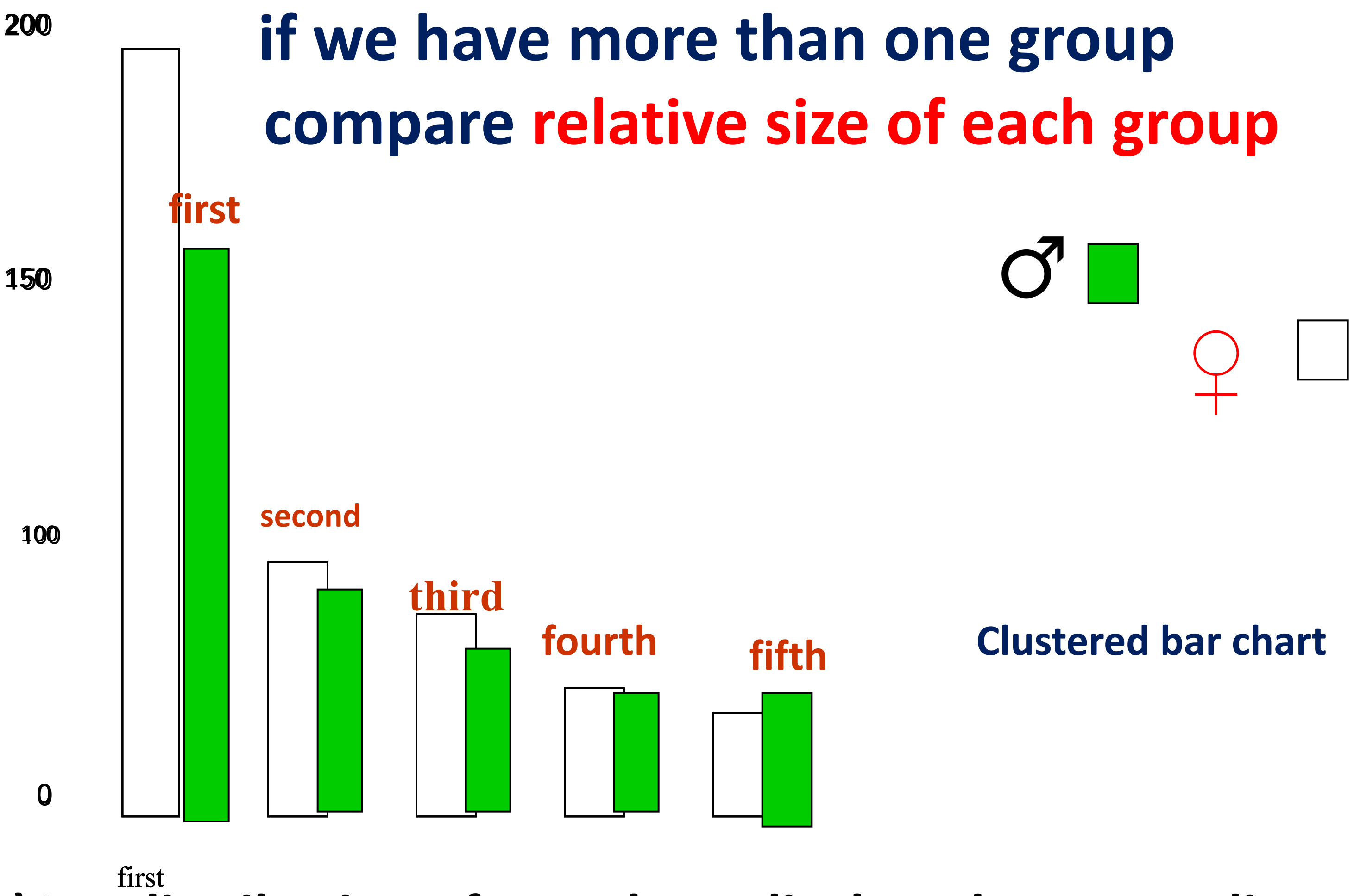
Magnitude differences

(I)Mutah medical student according to their year level 2021

nominal and ordinal data

Allows easier comparisons between data sets of different sizes.

if we have more than one group compare relative size of each group

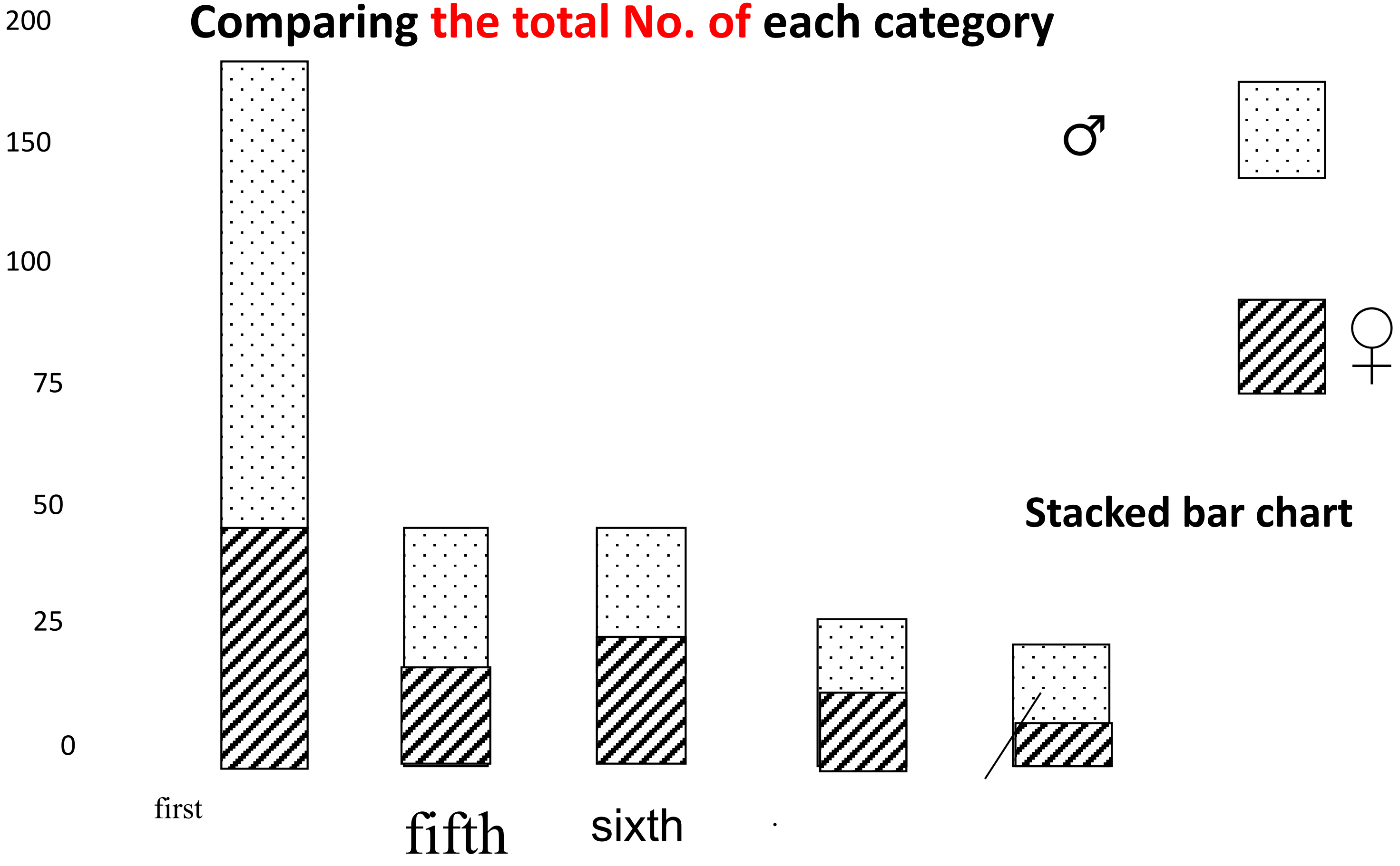


Clustered bar chart

(II)Sex distribution of Mutah medical student according to their year level 2021

nominal and ordinal data

Comparing the total No. of each category



Stacked bar chart

Sex distribution of Mutah medical student according to their year level 2021

Charting

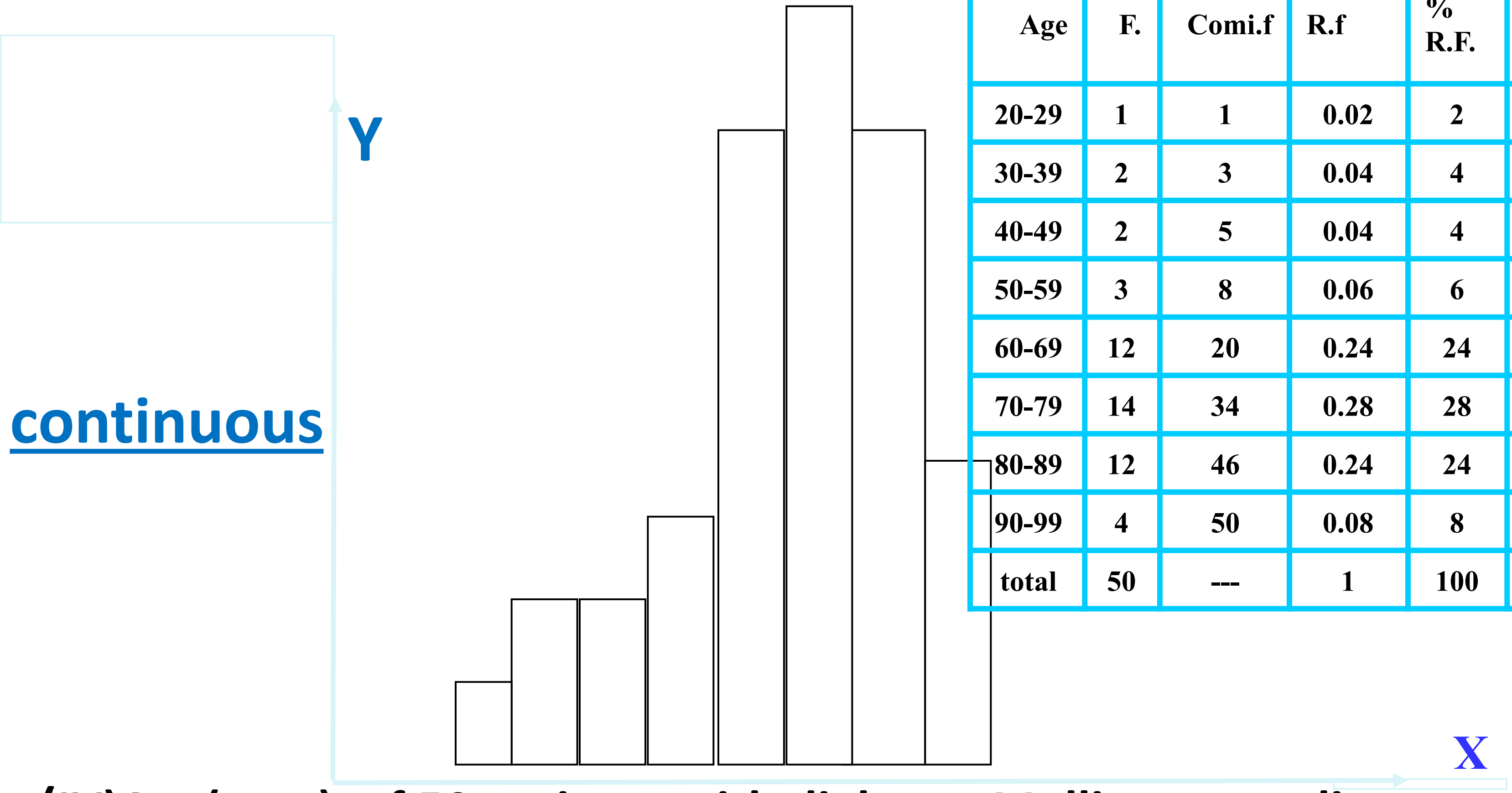
Continuous Metric Variable by

Histogram

Age (year)	F.	Commut frequenc	Relative frequenc	% R.F.	Cumulat R.F.	%cum Freq.
20-29	1	1	0.02	2	0.02	2
30-39	2	3	0.04	4	0.06	6
40-49	2	5	0.04	4	0.1	10
50-59	3	8	0.06	6	0.16	16
60-69	12	20	0.24	24	0.4	40
70-79	14	34	0.28	28	0.68	68
80-89	12	46	0.24	24	0.92	92
90-99	4	50	0.08	8	1.00	100
total	50	---	1	100	---	---

Histogram

The group frequency distribution table usually represented graphically or diagrammatically by histogram .



Age	F.	Comi.f	R.f	% R.F.	Cumi. RF.	% cum F.
20-29	1	1	0.02	2	0.02	2
30-39	2	3	0.04	4	0.06	6
40-49	2	5	0.04	4	0.1	10
50-59	3	8	0.06	6	0.16	16
60-69	12	20	0.24	24	0.4	40
70-79	14	34	0.28	28	0.68	68
80-89	12	46	0.24	24	0.92	92
90-99	4	50	0.08	8	1.00	100
total	50	---	1	100	---	---

(IV)Age(year) of 50 patients with diabetes Mellitus attending Al Karak Hospital during march 2022

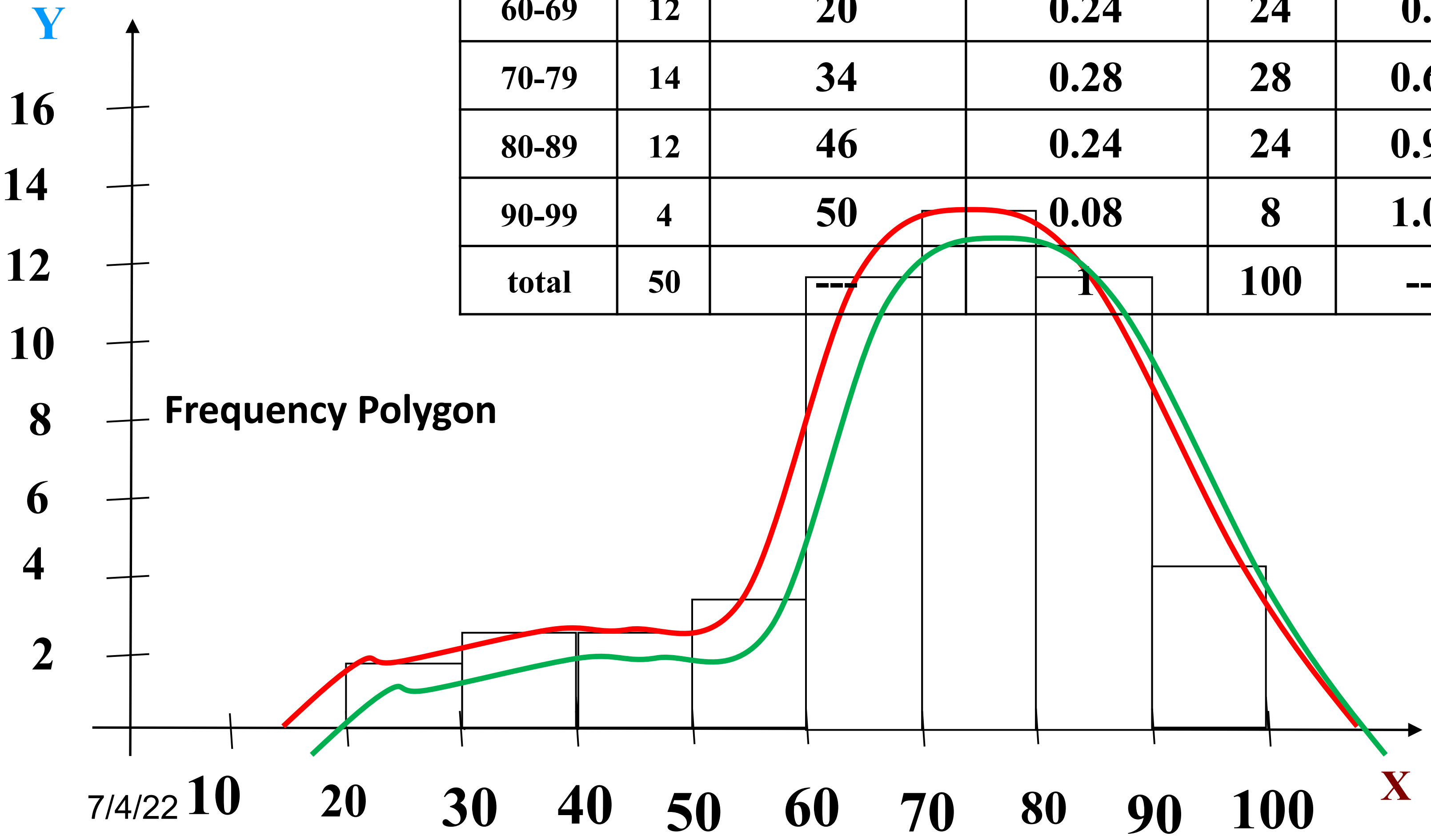
THE FREQUENCY POLYGON:

This type is used when the variable is of **continuous quantitative type and the table is of simple or complex type.**

Each category on the table represented by single point opposite its frequency on Y axis and **the mid-point of the interval on X axis.**

Then every two consecutive points are joined together by a straight line.

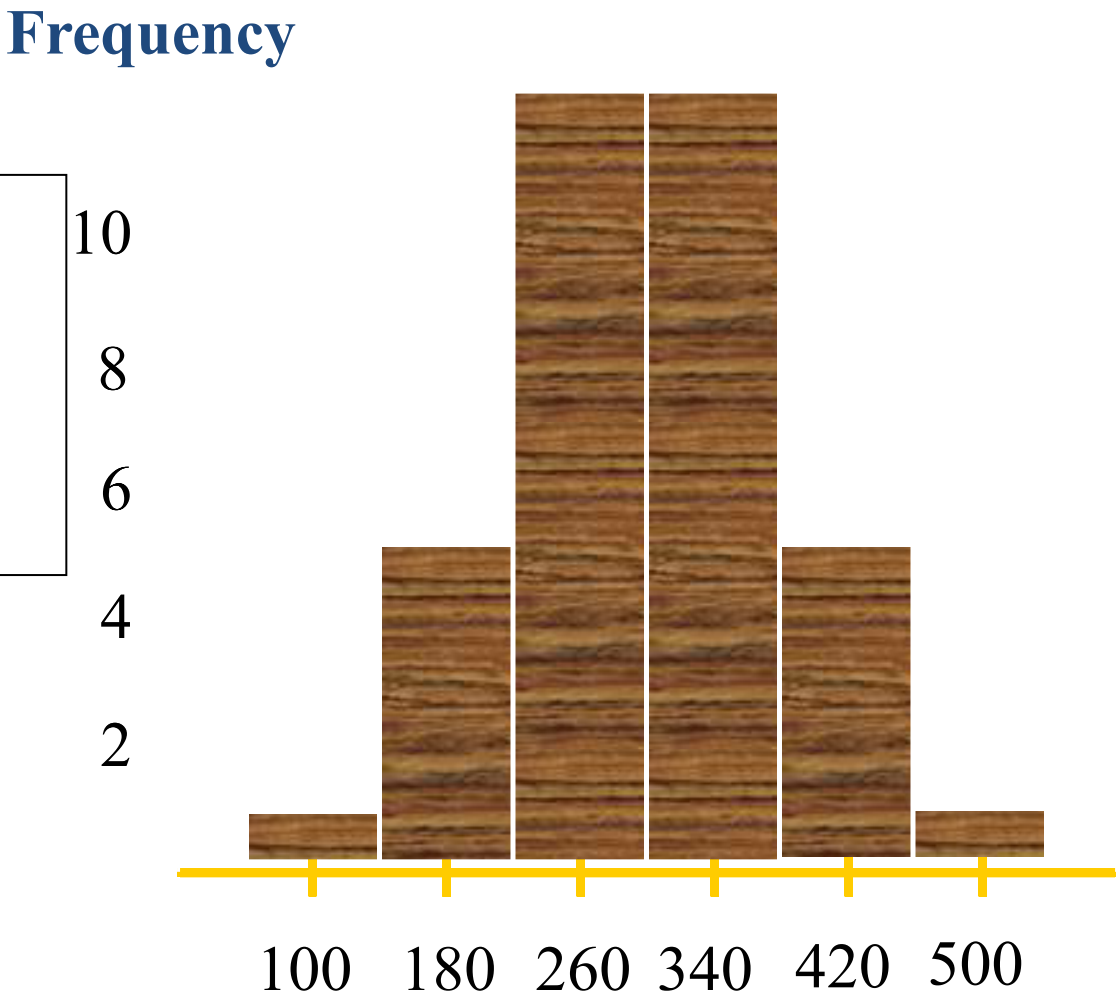
Age(year)	Freq	Commu.frequ.	Relat.Freque	% R.F.	Cumul. R.F.	%cum.Freq.
20-29	1	1	0.02	2	0.02	2
30-39	2	3	0.04	4	0.06	6
40-49	2	5	0.04	4	0.1	10
50-59	3	8	0.06	6	0.16	16
60-69	12	20	0.24	24	0.4	40
70-79	14	34	0.28	28	0.68	68
80-89	12	46	0.24	24	0.92	92
90-99	4	50	0.08	8	1.00	100
total	50	---	1	100	---	---



Frequency Polygon

Shapes of Histograms I

Symmetrical,
normal,
or bell-shaped

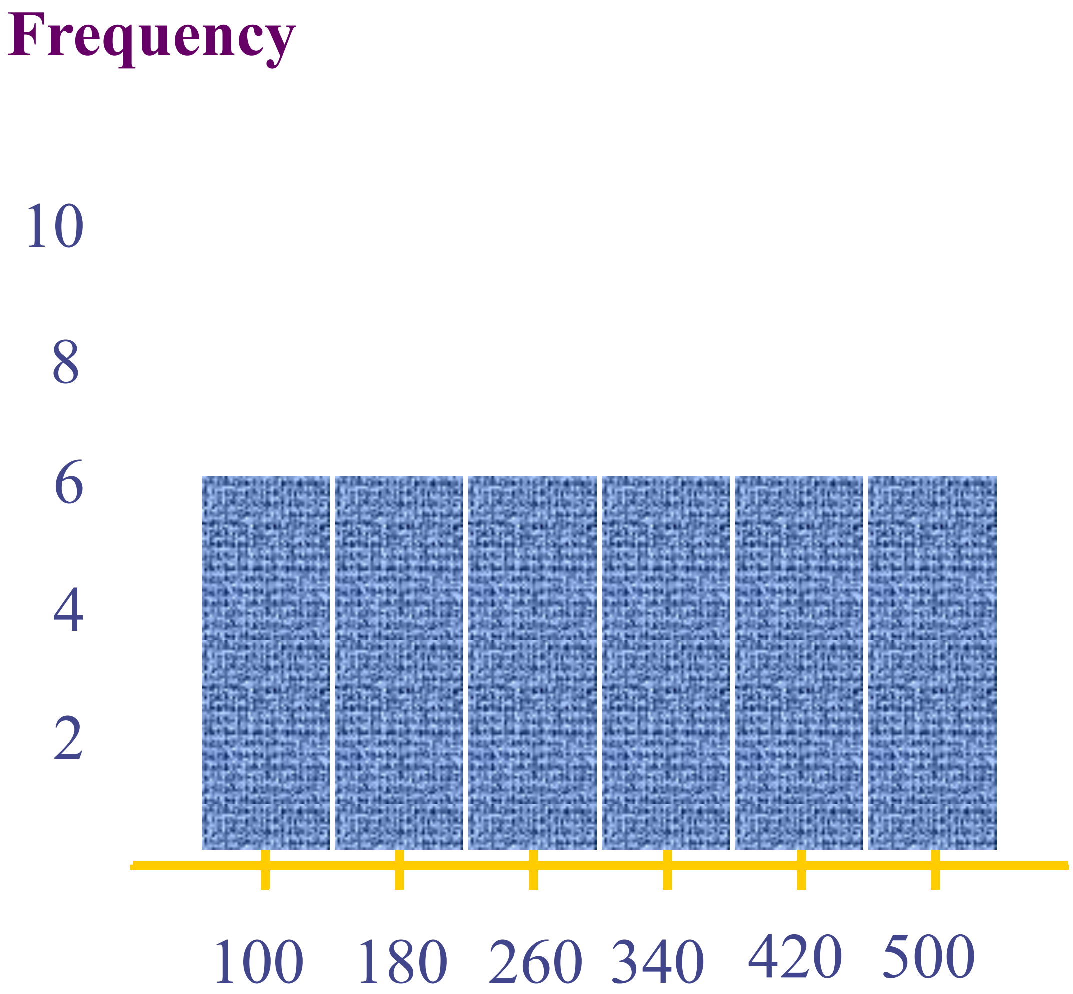


Shapes of Histograms II

Uniform

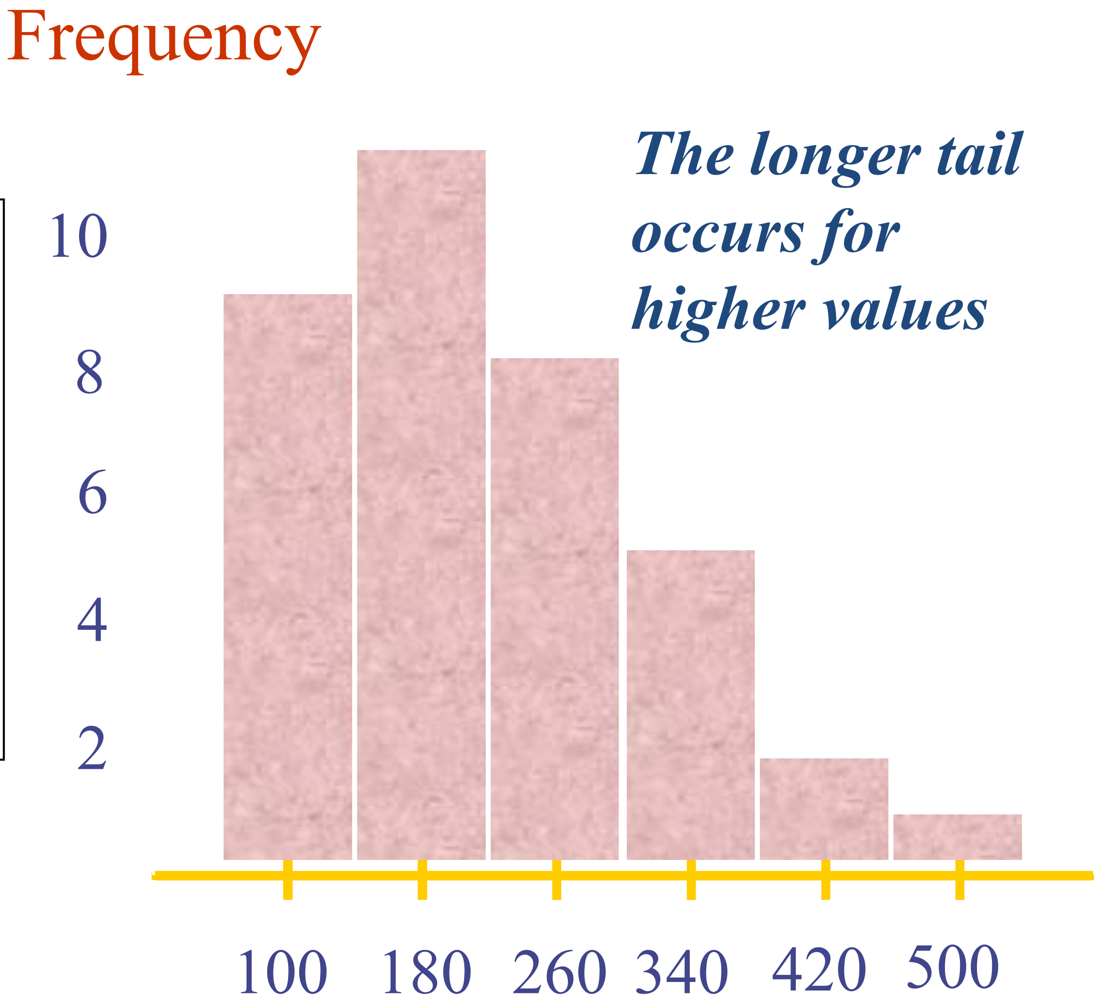
or

rectangular



Shapes of Histograms III

Skewed right
or
Positively
skewed



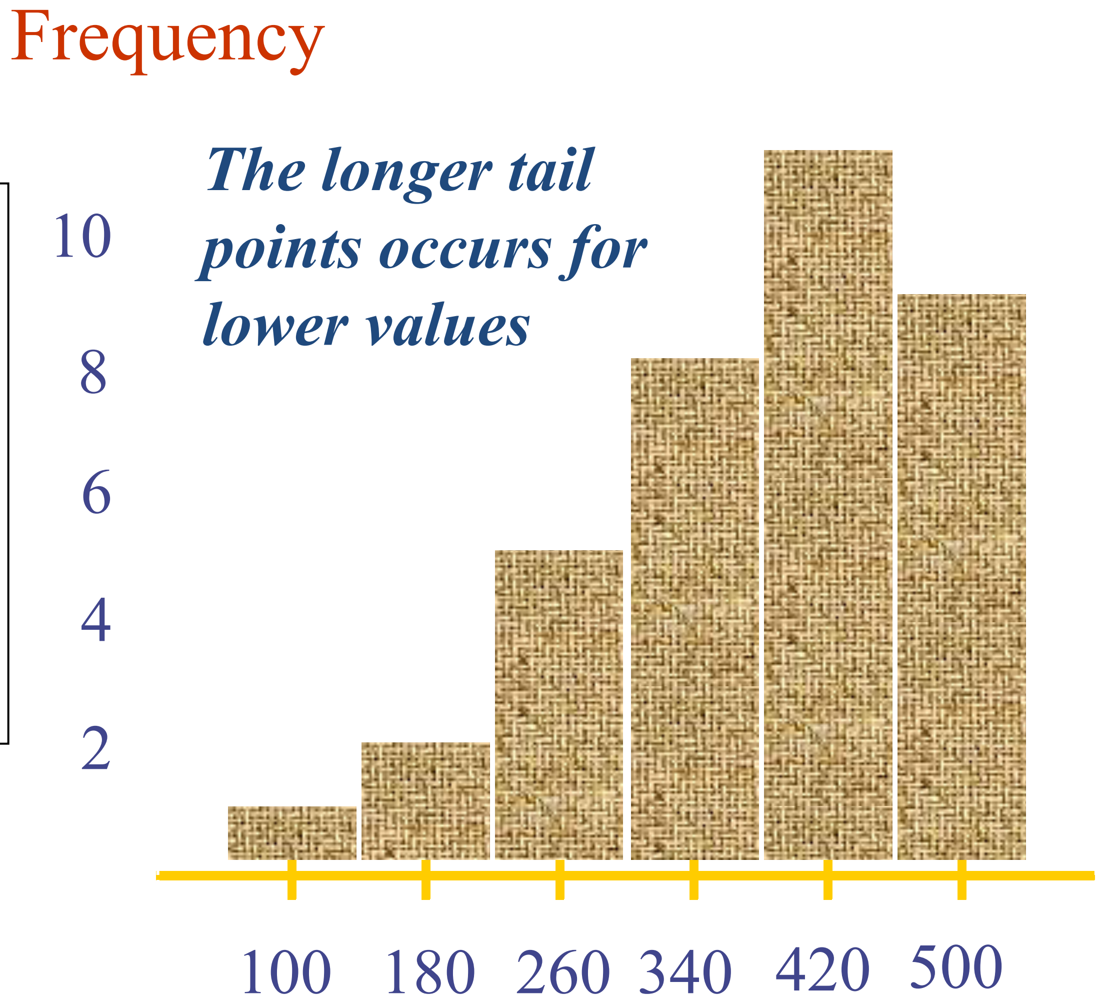
The longer tail occurs for higher values

Shapes of Histograms IV

Skewed left

or

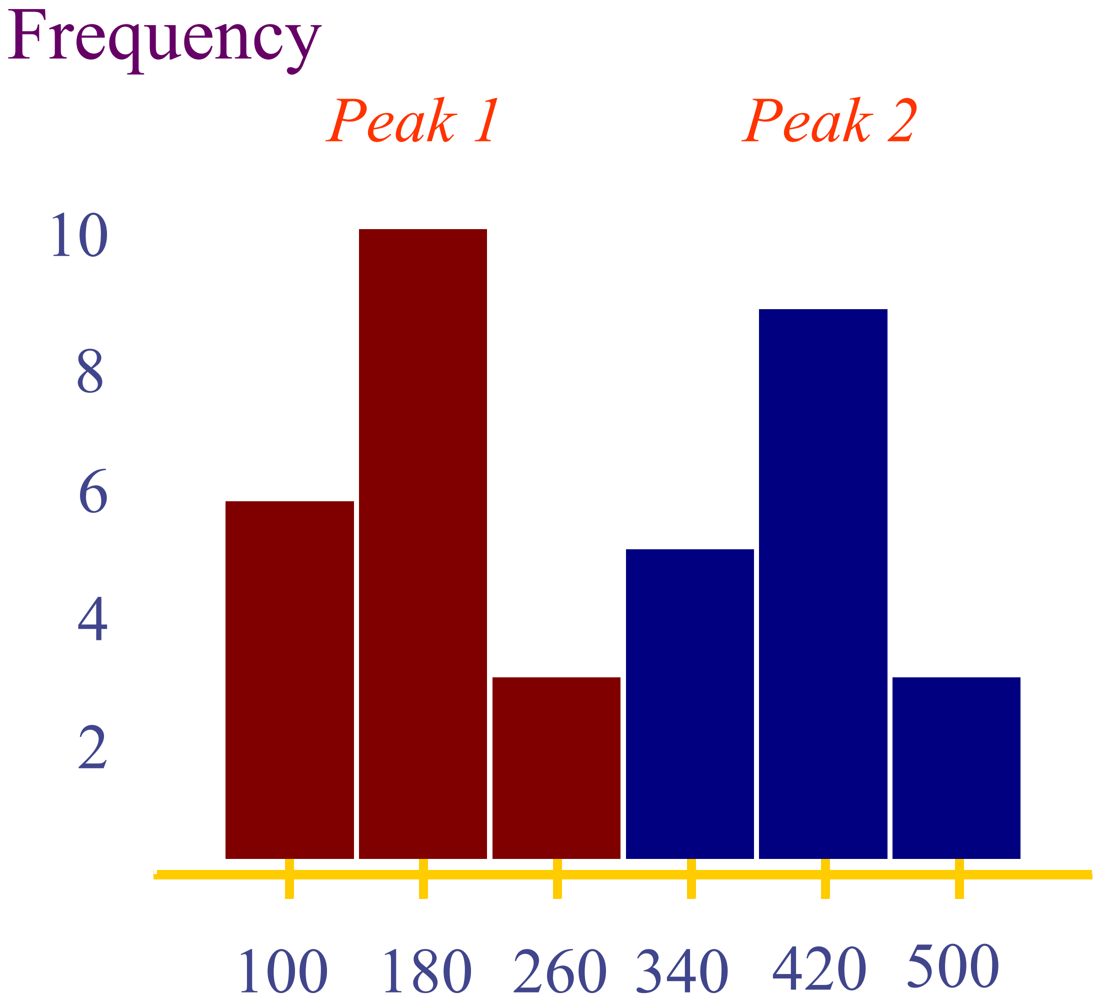
Negatively skewed



The longer tail points occurs for lower values

Shapes of Histograms V

Bimodal



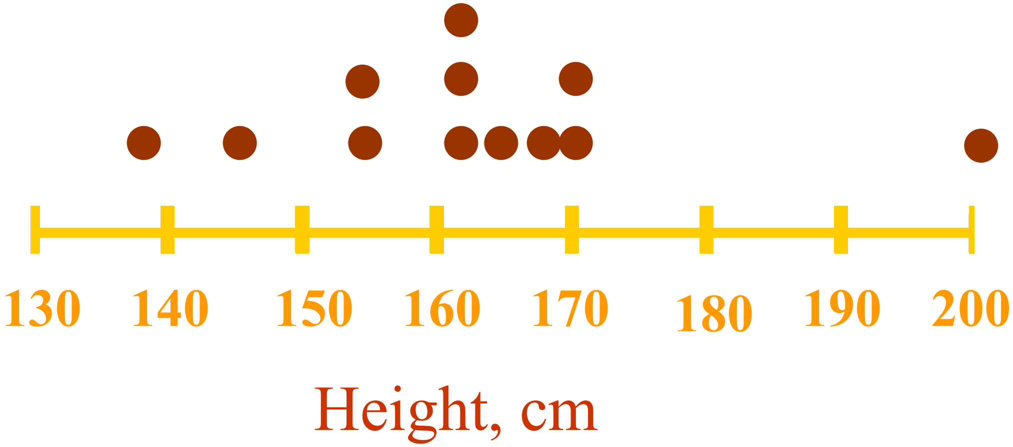
Dotplot

- Number line with dots representing data points
- Can visualize the "spread" of the data
- Data: Height of of 12 female students measured in (cm)

  139, 161, 170, 201,

  161, 168, 170, 155,

  165, 145, 155, 161

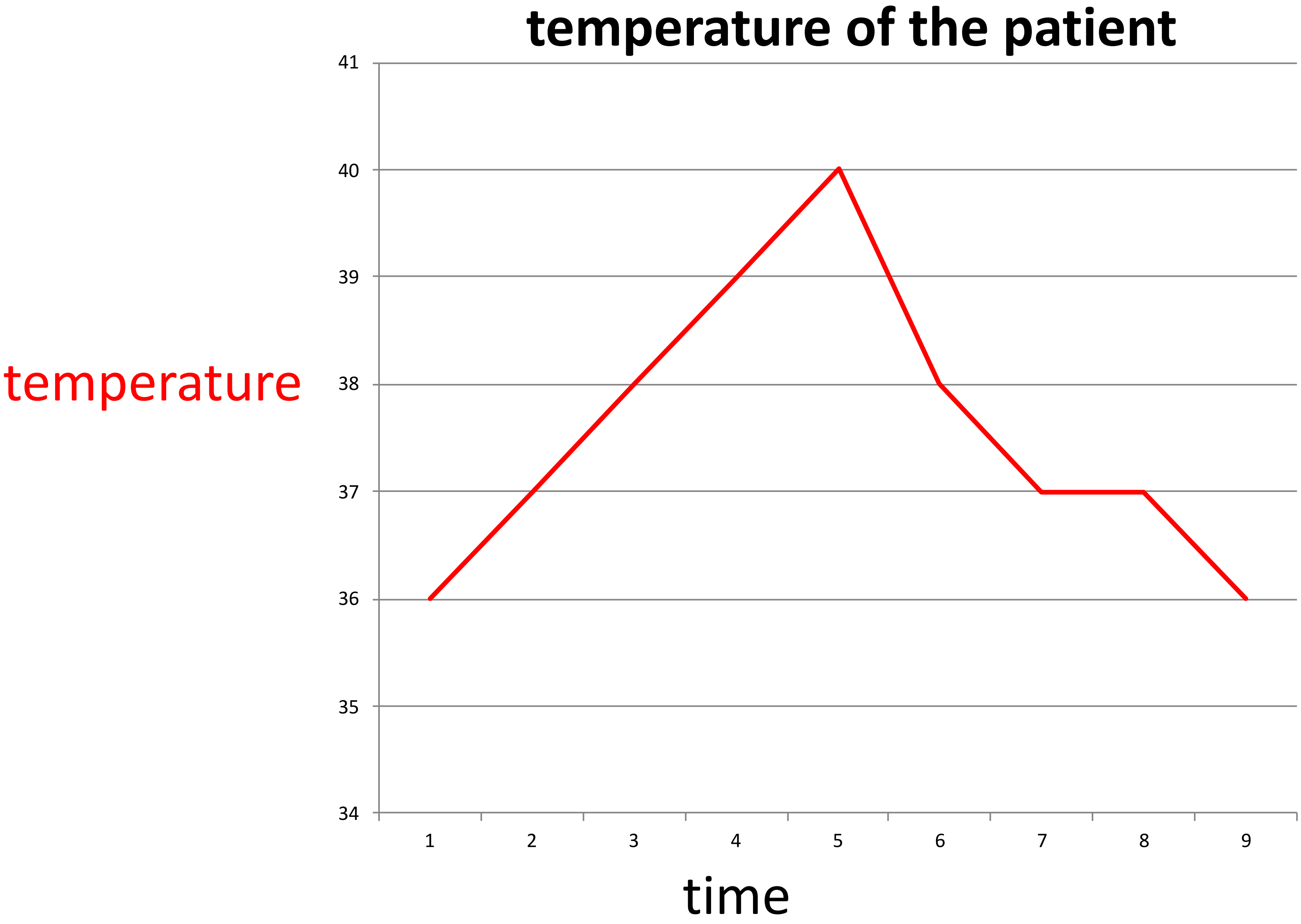


THE LINE GRAPH

- This type is specifically used when we are dealing with a certain observation that varies according to ***time***.
- That is when we are dealing with a time variable.
- (The time variable is a special type of continuous quantitative variable)
- Usually the time variable is put on the **horizontal axis** (**X-axis**) and the **other variable** is put on the vertical axis (**Y-axis**),
- then each observation is shown on the graph by means of a point opposite to the **exact time value** on the horizontal axis and opposite the corresponding value on the vertical axis,
- then every two consecutive points are joined by a straight line.

Example of this is a temperature chart of the patient. It is also used in study of trends of birth and death rate

Time	temperature
1	36
2	37
3	38
4	39
5	40
6	38
7	37
8	37
9	36



temperature of the patient
41
40
39
38
37
36
35
34
temperature
1
2
3
4
5
6
7
8
9
time

Evaluation of table or graph

Can this table or graph stand alone ?

It should be self explanatory, Through,

Labeling it properly .

Begin with title and carried on through out table or graph

No.
I II III …Graph
1 2 3 … Table

Title should contain :

what kind of data is this .

who were involved .

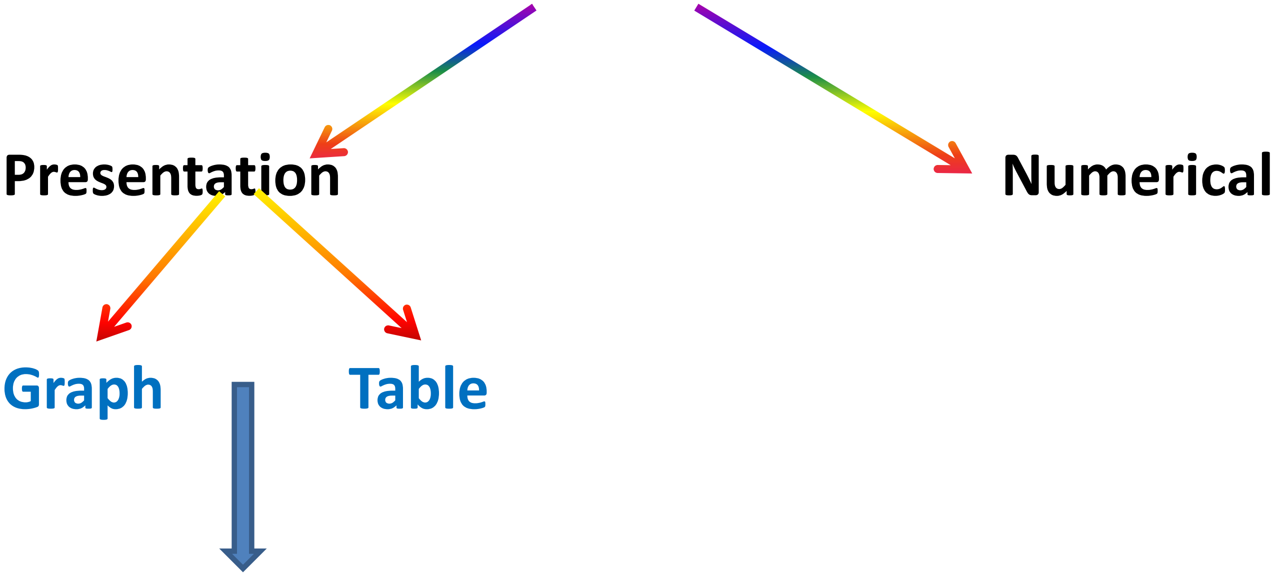
where it was collected .

when it was done .

Title
above table
Below graph .

Foot note may needed .

Description statistics summarization



Presentation

Numerical

Graph

Table

-*this approach might not be enough,*

-*comparisons between one set of data & another*

-*summarize data by one more step further .*

-*presenting a set of data by a*

- *single Numerical value*

Numerical Presentation

Numerical Description

Measures of Central Tendency

Measures of Dispersion

The central value as representative value in a set of data,

1-Measures of central tendencies (Location) .

A value around which the data has a tendency to congregate (come together)or cluster

2-Measures of Dispersion, scatter around average

A value which measures

the degree to which the data are or are not, spread out

29

-single Numerical value. ??

Are we using largest value ?

Are we using lowest value ?

As a single Number representation

The central value as representative value in a set of data,

Measures of Central Tendency

A value around which the data has a tendency to congregate or cluster

1- Mean

2- Median

3- Mode

4- weighted mean

the choice of the most appropriate measure depends crucially on the type of data involved

31

Mode (Mo)

- Most frequently occurring value in a set of observation

5 1, 3, 2, 6 , 7, 10 5 ?????

Or

- the value of observation which has the highest frequency in a set of observation .

1 5 1, 3, 1, 2, 6 , 7, 10 5 ?????

- Mode is the only measure of central tendency that can be used for qualitative data ???
- is not practically useful with the metric continuous data where no two value may be the same,

➢ If the observation all having different value

5 1, 3, 2, 6 , 7, 10 ?????

So

the observation all having different value

there is no Mode 5 1 3 2 6 .

We might have one Mode 5 , 1 2, 3, 1, 6 uni modal

We might have more than one Mode

5, 1, 3, 5 7, 3, 6 , 2 Two Mode Bimodal

5, 1, 3, 5, 7, 3, 6, 2, 1 Three Mode Tri modal

5, 1, 3, 5, 7, 3, 6, 2, 1 , 3 ???

Characteristics of Mode

Advantages and Disadvantages

1-Requires no calculation just counting

2- It may not exist (No Mode)

3-It is not necessarily be unique

there may be one mode unimodal

more than one mode in a set of data

Bimodal, Tri modal ….

- It is the only measure of central tendency that can be used for qualitative data

4 -Mode is not practically useful with the metric continuous data

Median (Md)

Measures of C T
1- Mean
2- Median
3- Mode

It is the middle value in ordered data

(from the lowest to the highest values).

-Divided the observations into two halves .

So

- 1/2 of observation their values less than the value of median
- 1/2 of observation their values More than the value of median

- Median is located the center of data by count and disregards the size .
- Median is thus a measure of centrals

Steps in calculating the median

1- Arrange the value.

From the lowest to the highest value .

Exam. marks

50 10 90 20 40 → 10 20 40 50 90

2- Find the Median position by this formula

$$\boxed{\frac{n \quad 1}{2} \qquad \frac{5 \quad 1}{2} \qquad 3^{rd}}$$

Calculate the value of the third observation = 40 marks .

Odd No. we have just one median position .

Even No. we have two median position or

two median values

Median value =Average of the two values

Even No 50 10 90 20 40 95

10 20 40 50 90 95

$$\frac{n \qquad 1}{2} \qquad \frac{6 \qquad 1}{2} \qquad \frac{7}{2} \qquad 3.5$$

Median located (position)

between the 3rd and 4th .

Median value =Average of the two (3rd and 4th) values

$$Md \qquad \frac{40 \qquad 50}{2} \qquad 45$$

Characteristics

10 20

20 40 50 90 95

10 20 40 50 90 95 99 100........

10 20 40 50 70 85 90 99 100

1 20 40 50 70 85 90 99 100

10 20 40 50 70 85 90 99 1000.

two extremes

15 20 30 35 95 99 100

skewness

1 5 10 **35** 40 99 1000

Characteristics of the Median

It is always existed .

- It is always unique, there is one and only one Md .
- It is not affected by two extremes, not sensitive by two extremities .
- Not affected by skewness in the distribution or
- Not affected by presence of outliers
- It is discard a lot of information

  because it ignores most of the values apart from those in the center of distribution

-

Measures of C T
1- Mean
2- Median
3- Mode

Mean $\overline{X}$

Arithmetic Mean

- more commonly known as average
- -it is an arithmetic average of a set of observation obtained by

- Adding the values of all observation together .
- Dividing the sum by No. of observation in sample .
- *It represent the center of data according to the size of the values .*

Example :

following are the scores of five students

40 50 90 10 20

$$\overline{X} = \frac{\sum X}{N}$$

$$\overline{X} = \frac{\sum X}{N}$$

Σ = sigma = summation .

X = value of observation

N = No. of observation

$\overline{X}$ = is the sum of value of all observation divided by the total No. of observation

41

Characteristics of the Mean

Advantages and disadvantages

- Relatively easy to handle
- It is always exist
- It is always unique,

  there is one and only one Mean
- It takes into account every item in a set of data
- It uses all of the information in the data set.

- affected by skewness in the in the data set

- affected by presence of outliers

- it can not be used with the ordinal data ???

- It is affected by the two extremes by

  a very small or

  a very large value .

- It is sensitive to the extremes

1 2 3 4 5 mean = 3

1 2 3 4 50 mean = 12

1 2 3 4 500 mean = 102

- this may produce a mean that is not very representative of the general mass of data

  another disadvantage ,

- it can not be used with the ordinal data ???

  (ordinal data are not real numbers,

  so they cannot be added or divided)

Weighted mean

It is the average measure of a No. of means, when we take into consideration the frequencies of each mean .

It is used when some values of observation more important in some sense than others .

$$W.mean \quad \frac{W_1\bar{X}_1 \quad W_2\bar{X}_2 \quad W_3\bar{X}_3 \quad \ldots\ldots\ldots\ldots \quad W_k\bar{X}_k}{W_1 \quad W_2 \quad W_3 \quad \ldots\ldots\ldots \quad W_k}$$

Group	$\overline{X}$ Hb	No. of person
I	13	5
II	14	10
III	13.5	15

$$W.mean \quad \frac{5 \quad 13 \quad 10 \quad 14 \quad 15 \quad 13.5}{5 \quad 10 \quad 15} \quad \frac{407.5}{30} \quad 13.5\, gm/100\, ml$$

$$\frac{65+140+202.5}{5+10+15} = \frac{407.5}{30} = 13.58$$

Central Tendency In Grouped Data

Age (year)	F	M.P.	(M.P.)F	Cum. F	%
20-29	2	24.5	24.5 2 = 49	2	4
30-39	8	34.5	34.5 8 = 276	10	16
40-49	5	44.5	44.5 5 = 222.5	15	10
50-59	14	54.5	54.5 14 = 763	29	28
60-69	15	64.5	64.5 15 = 967.5	44	30
70-79	6	74.5	74.5 6 = 447	50	12
total	50	---		---	100

$$(M.P.)F \quad 2725$$

2725/50 =54.5 years

Choosing the most appropriate measure (Mean, Median or mode)

How do you chose the most appropriate measure of location in a given set of data ??

The main thing is to remember is that

mean can not be use with the ordinal data(**because they are not real numbers**

the median can be use for both ordinal & metric data.

and

the Median can be use for both ordinal & metric data.

when the later (metric data)

is skewed

Or

when there is outlier

the median is

more representative of data than the mean

????????

	Mode	Median	Mean
Nominal	**Yes**	**No**	**No**
Ordinal	**Yes**	**Yes**	**No**
Metric discrete	**Yes**	**Yes if distribution is markedly skewed**	**yes**
Metric continuous	**No**	**Yes** if distribution is markedly skewed	**yes**

Thank you ...

Thank
you

?
Any questions?
© Presentation-Process.com

The central value as

1-Measures of central tendencies (Location) .
A value around which the data has a tendency to congregate (come together)or cluster
2-Measures of Dispersion, scatter around average
A value which measures
the degree to which the data are or are not , spread out

1-Measures of central tendencies (Location)

75, 75, 75, 75, 75, 75, Mean = ????

75, 70, 75. 80, 85. **Mean = ????**

60, 65, 55, 70, 75, 75, ,70, 80, Mean= ????

$$\overline{X} = \frac{\sum X}{N}$$

2-Measures of Dispersion,

The central value as
1-Measures of central tendencies
2-Measures of Dispersion,

Measures of Dispersion (Measures of Variation) (Measures of Scattering) measures of spread