GAROWE TEACHERS EDUCAITON COLLEGE (GTEC)
GAROWE, PUNTLAND, SOMALIA

LECTURE NOTE

ON

INTRODUCTION TO STATISTICS

Introduction

Our life is full of events and phenomena that enhance us to study either natural or artificial phenomena could be studied using different fields one of them is statistics. For example, study on fishing.
Since statistics is used in almost every field of endeavor, the educated individual should be knowledgeable about the vocabulary, concepts, and procedures of statistics.

What is Statistics?

> Statistics is a branch of science dealing with collecting, organizing, summarizing, analysing and making decisions from data.

Statistics is divided into two main areas, which are **descriptive** and **inferential** statistics.

Descriptive Statistics

> Descriptive statistics deals with methods for collecting, organizing, and describing data by using tables, graphs, and summary measures.

Variable

> A variable is a characteristic under study that takes different values for different elements.

For example, if we collect information about income of households, then income is a

variable. These households are expected to have different incomes; also, some of them may have the same income.

Value

> The value of a variable for an element is called an observation or measurement.

The following is an example to explain the difference in the meaning between variable and the measurement.

Types of Variables

Quantitative Variable: It gives us numbers representing counts or measurements.
Qualitative Variable: It gives us names or labels that are not numbers representing the observations.

Qualitative Data	Quantitative Data
• Deals with descriptions. • Data can be observed but not measured. • Colours, textures, smells, tastes, appearance, beauty, etc. • Qualitative —► Quality	• Deals with numbers. • Data which can be measured. • Length, height, area, volume, weight, speed, time, temperature, humidity, sound levels, cost, members, ages, etc. • Quantitative — Quantity

Raw Data

Data recorded in the sequence in which they are collected and before they are processed or ranked are called **raw data**.
Example
Suppose we collect information on the scores of 20 students from semester one at GTEC. The data values, in the order they are collected, are recorded in Table 2.1.

Table 2.1: Scores of 20 students

20	25	18	17	15	19	21	22	27	30
13	15	17	20	25	12	18	13	29	21

The table 2.1 represents quantitative raw data.

Organizing Data

Organizing data is used to organize data set.

Frequency Table
A **frequency table** is a way of organizing collected data.

To do this we draw a table with three columns:

- The first column is for the different items in the data set.
- The second column is for the tally marks.
- The last column is the frequency column where we can add up the tally marks and write in the corresponding frequencies.

We can add up all of the frequencies to find the **total frequency** of the set of data.

For example,

Let's say we wanted to collect some data on the ways of transport used by students to get to school. We would start collecting data and writing the answers in a list.

The best way to sort this data set is to use tally charts. The data collection looks like this:

walk, bus, bike, walk, bike, bus, walk, car, walk, bike, bike, bus, walk, walk, walk, car, bus, walk, bus, bus, walk, car, car, walk, walk, train, bike, bus, walk, walk

The tally chart for the same data looks like this:

Transport	Tally	Frequency
Walk	𝍸 𝍸 | | |	13
Bus	𝍸 | |	7
Car	| | | |	4
Bike	𝍸 |	5
Train	|	1

For example,

A frequency table showing the ages of 25 students on a college course.

Age	Frequency
18	15
19	6
20	4
	Total = 25

Relative Frequency

Relative frequency can be defined as *the number of times an event occurs divided by the total number of events occurring in a given scenario. Relative frequency can be calculated as below:*

$$\text{Relative frequency of a category} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}}$$

Percentage Relative Frequencey

It is the relative frequency multiply by 100%.

Example

Number of Children y	*Frequency f*	*Relative Frequency rf = f/n*	*Percentage Frequency p = 100 * rf*
0	1	0.10	10.00
1	0	0.00	00.00
2	2	0.20	20.00
3	1	0.10	10.00
4	2	0.20	20.00
5	4	0.40	40.00
	$n = 10$	1.00	100.00%

Cumulative Frequency

Cumulative frequency is *the total of a frequency and all frequencies in a frequency distribution until a certain defined class interval.*

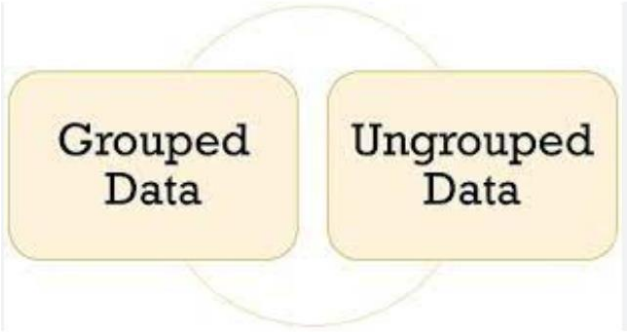
School Grade	Frequency of Students	Cumulative Frequency
1	23	23
2	20	23 + 20 = 43
3	15	43 + 15 = 58
4	12	58 + 12 = 70
5	10	70 + 10 = 80
6	8	80 + 8 = 88

Exercise

Complete the table by filling in the blanks then answer the following question.

Temp (°F)	Tally	Frequency	Cumulative Frequency
100-105	𝍸		
106-111		12	
112-117	𝍸 𝍸 𝍸 |		
		14	
124-129		2	
		1	

Concept of Group



Quantitative data can be classified into ungrouped data and grouped data.

- Ungrouped data is the type of distribution in which the data is individually given in a raw form. Example

Age	Frequency
18	15
19	6
20	4
	Total = 25

- Grouped Data: It is a way of organising a large set of data into more manageable groups. Examples

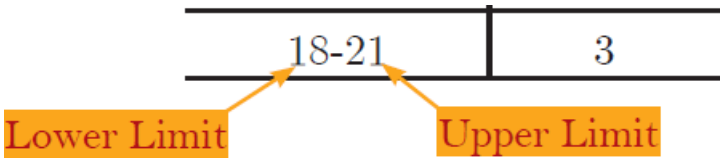
Marks	Tally	Frequency
1 to 5	||	2
6 to 10	𝍸 |||	8
11 to 15	𝍸 ||	7
16 to 20	|||	3
		Total = 20

Class

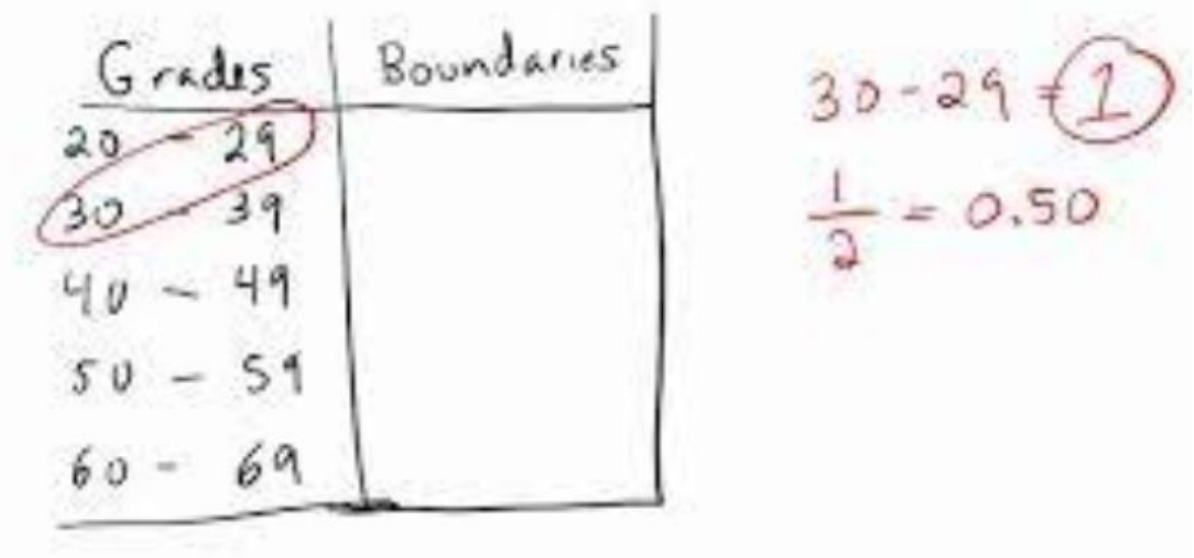
In statistics, a class is **a grouping of values by which data is binned for computation of a frequency distribution**

Class Interval	Frequency
67 to 79	3
79 to 91	5
91 to 103	8
103 to 115	9
115 to 127	5
Total	30

Class Limits: A class consist of two numbers, the first number is called **lower limit** and the second number is called the **upper limit.**



Class Boundaries: Class Boundaries are similar to class limit, where each class



boundaries divided into lower boundary and upper boundary. However, the upper boundary limit for any class always the same of the lower boundary limit of previous class. It is calculated as below:

For Example: Find the class boundary of the class 40 – 44

Solution:

Lower boundary limit= 40 – 0.5 = 39.5

Upper boundary limit= 44 + 0.5 = 44.5

Midpoint: It is calculated by adding class limits or class boundries and dividing by 2.

$$x_m = \frac{\text{Lower boundary} + \text{Upper boundary}}{2}$$
$$= \frac{\text{Lower limit} + \text{Upper limit}}{2}$$

Find the class midpoint of the class $40 - 44$.

Solution: The lower limit is 40 and the upper limit is 44, then the midpoint of this class is:

$$x_m = \frac{40 + 44}{2} = 42 \text{ or } x_m = \frac{39.5 + 44.5}{2} = 42.$$

Example on class limits, class boundaries and midpoint

Class	Frequency	Mid point	Class boundaries	
			Lower	Upper
135-139	6	137	135	139
140-144	4	142	140	144
145-149	11	147	145	149
150-154	15	152	150	154
155-159	8	157	155	159

Class Interval

Class width = Upper boundary limit – Lower boundary limit

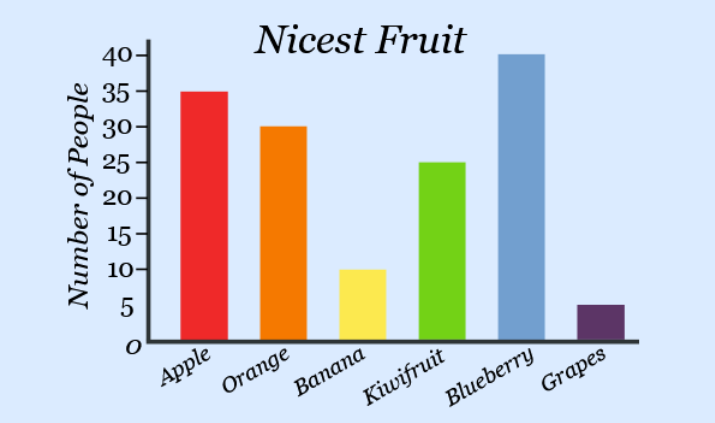
Inclusive Class: When the lower and the upper-class limits are included, then it is an inclusive class. For example: 220-234, 235 -49…. etc
Exclusive Class: When the lower limit is included, but the upper limit is excluded is an exclusive class. For Example: 150 – 153, 153 – 156, … etc.

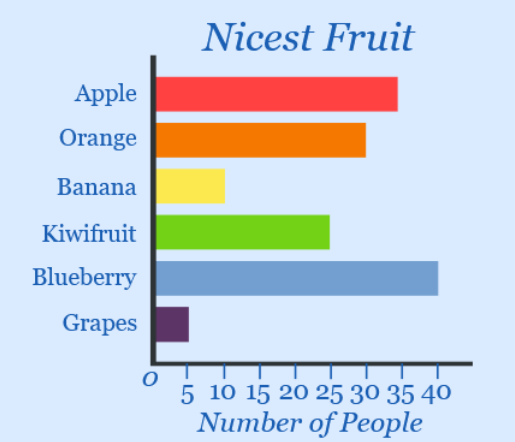
Graphical Presentation of Qualitative Data

There are many types of graphs that are used to display qualitative data; in this part we will study two graphs which they are commonly used to display the qualitative data, these graphs are the **Bar chart** and the **Pie chart**.

Bar Chart

A graph made of bars whose heights represent the frequencies of respective categories is called a bar graph.





Pie Chart

A circle divided into portions that represents the relative frequencies or percentages of a population or a sample of different categories is called a pie chart. The pie chart

Construct a Pie Graph (Chart)

1. Draw a circle.
2. Find the central angle for each category by the following equation:

   $$\text{Measure of the central angle} = (\text{Relative frequency}) \times 360^0$$

3. Draw sectors corresponding to the angles that obtained in step 2.

is one of the most commonly used charts when we need to display percentages, or display frequencies and relative frequencies.

Example

Step1: Draw a circle.

Step2: Find the central angle for each category by the equation:

Measure of the central angle= (Relative frequency) ×360 %

Applying step 2 for each category, we get

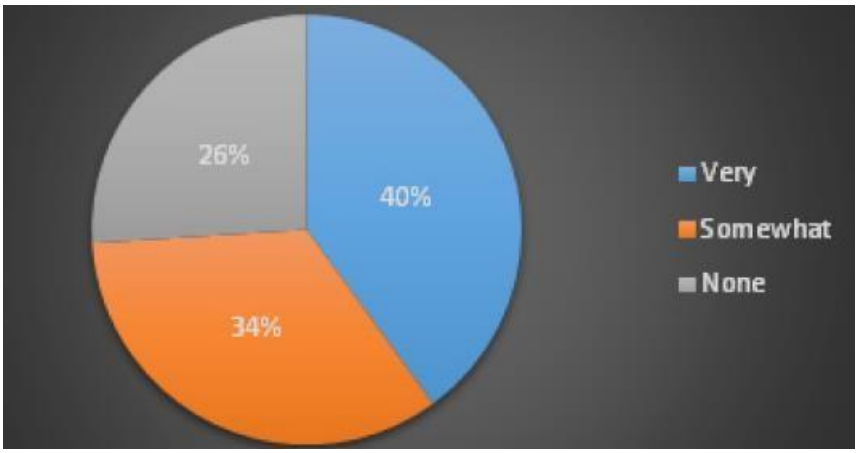
Type of Satisfaction Variable	Number of Students Frequency
Very high Satisfaction (v)	20
Somewhat satisfaction (s)	12
No satisfaction (n)	18
	Sum=50

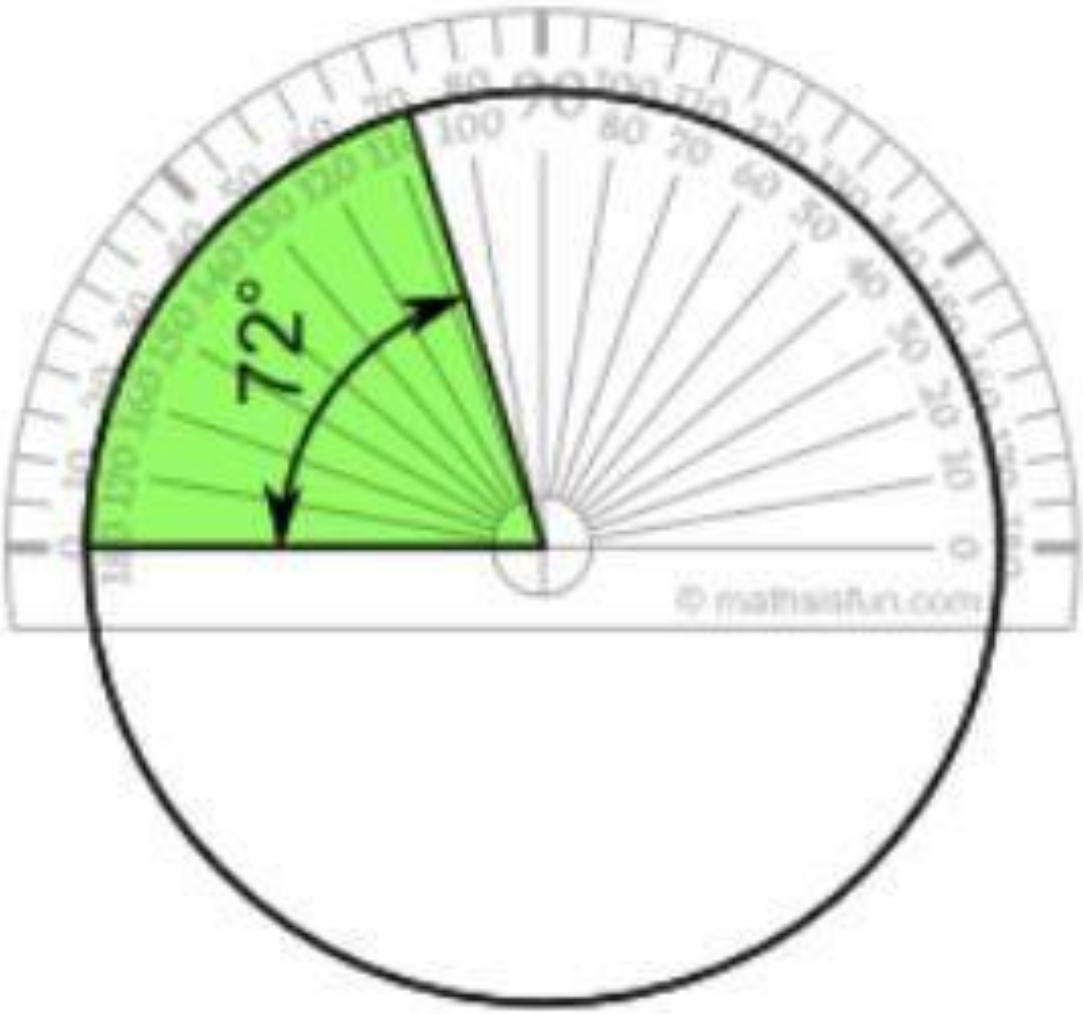
For category (v) the measure angle is (0.40)(360 %)=144 %

For category (s) the measure angle is (0.34)(360 %)=122.4 %

For category (n) the measure angle is (0.26)(360 %)=93.6 %

Step2: Draw the sectors corresponding to above angles, then the pie chart for such data is constructed in figure 2.2:




72°
© mathsisfun.com

Graphical Presentation of Quantitative Data

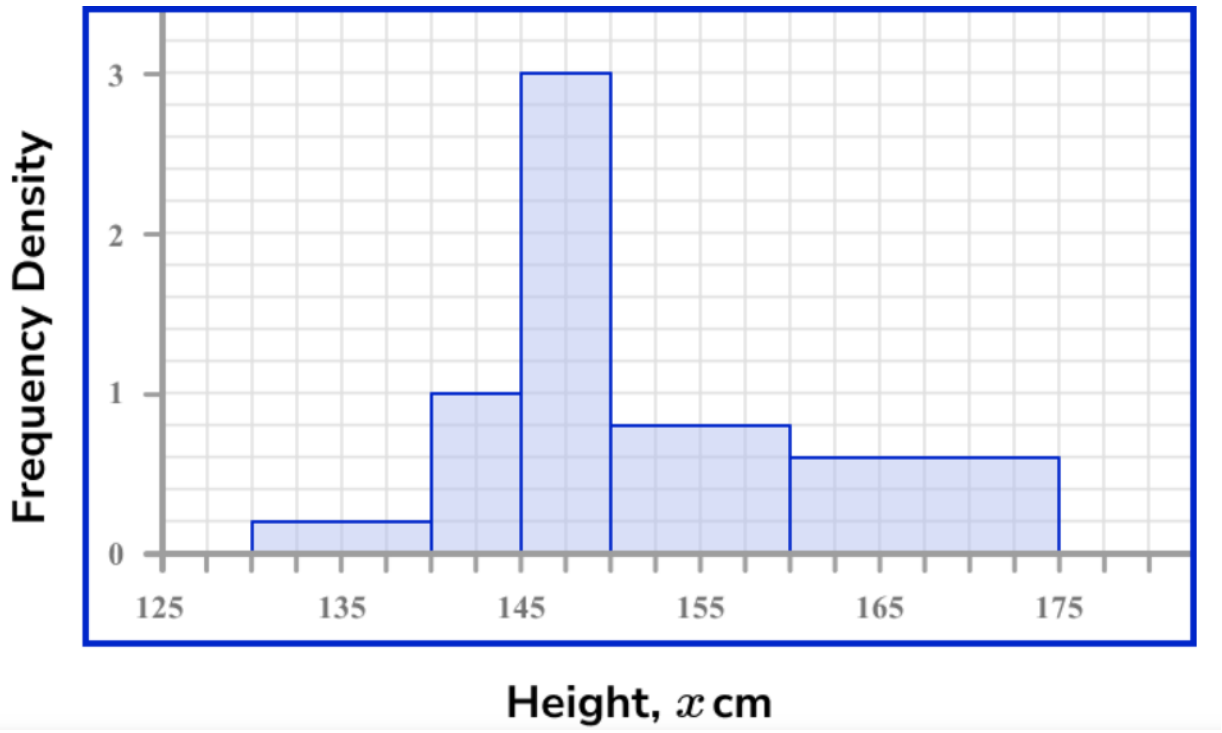
Histogram

Histogram is a graphical technique that used for representation of a frequency, a relative frequency, or a percentage of frequency distribution table of grouped data.

A **histogram** is similar to a bar chart but is used to display quantitative **continuous data (numeric data)**, whereas a bar chart (or bar graph) is used to display qualitative or quantitative **discrete data**. For example,

Below is a grouped frequency table and the associated histogram.

Height, cm	Frequency	Frequency Density
$130 \leq x < 140$	2	0.2
$140 \leq x < 145$	5	1
$145 \leq x < 150$	15	3
$150 \leq x < 160$	8	0.8
$160 \leq x < 175$	9	0.6



Polygon

A frequency polygon is a graph that displays the data by using line segments that connect points plotted for the frequencies at the midpoints of the classes.

Construct a Polygon

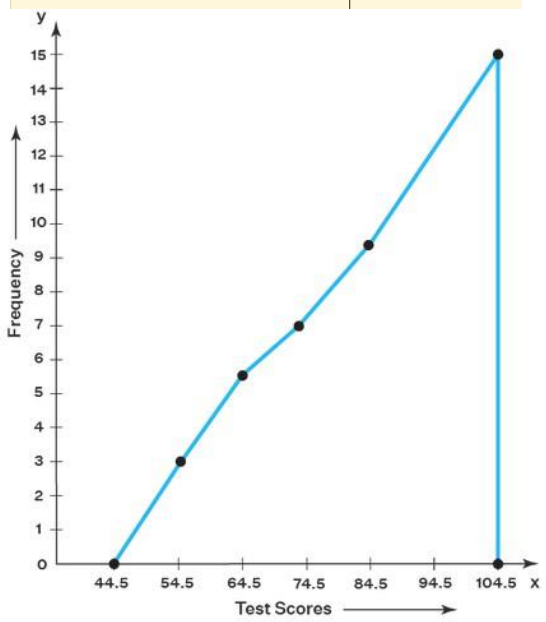
1. Find the class midpoints.
2. Mark the class midpoints on the horizontal axis.
3. Mark the frequencies on the vertical axis.
4. Plot the points (Class midpoint, Frequency).
5. Connect these points by straight lines segments.

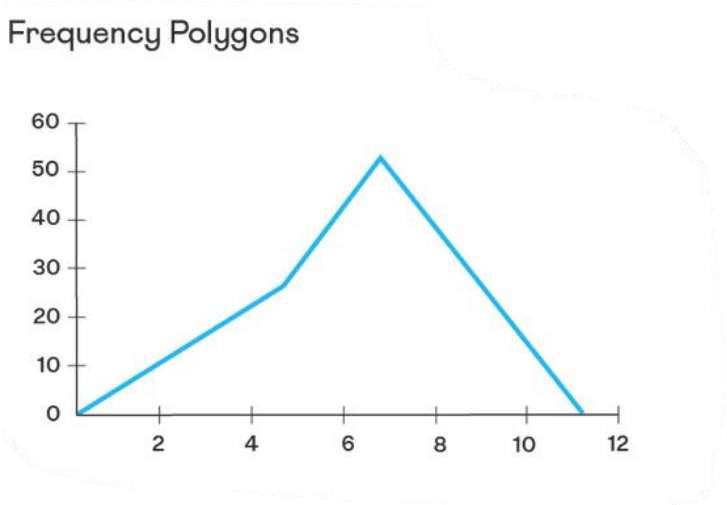
Example

Construct a frequency polygon without a histogram using the data given below:

Test Scores	Frequency
49.5 - 59.5	10
59.5 - 69.5	3
69.5 - 79.5	7
79.5 - 89.5	15
89.5 - 99.5	5

Test Scores	Frequency	Classmark
49.5 - 59.5	3	54.5
59.5 - 69.5	5	64.5
69.5 - 79.5	7	74.5
79.5 - 89.5	10	84.5
89.5 - 99.5	15	94.5





Cumulative Frequency Curve(Ogive)

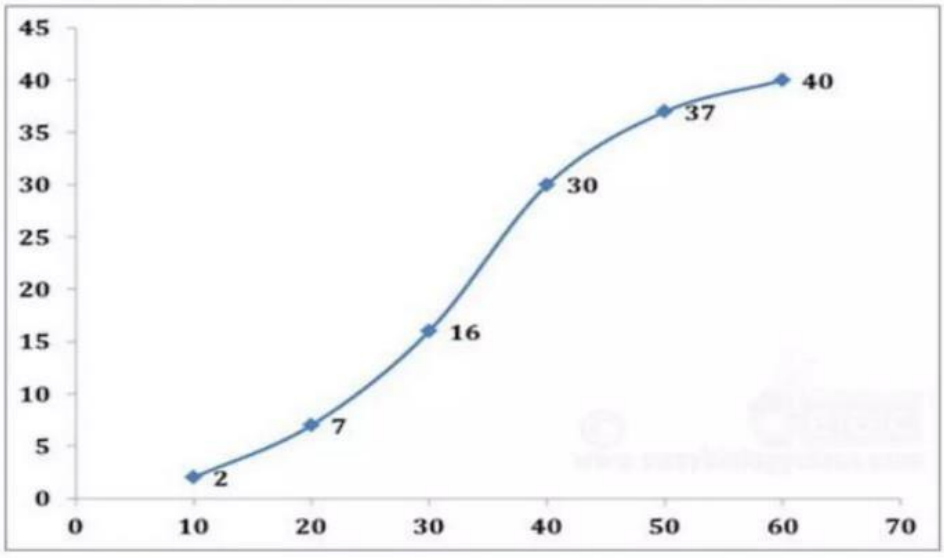
An ogive is a curve drawn for the ascending cumulative frequency of grouped data in a distribution table by first joining plotting dots marked above the upper boundaries of classes at heights equal to the ascending cumulative frequencies of respective classes, then joining these points by smooth curve

Construct Ascending Cumulative Frequency

1. Mark the upper boundaries on the horizontal axis.
2. Mark the ascending cumulative frequencies in the vertical axis.
3. Plot the points of the coordinates (upper boundary, ascending cumulative frequency).
4. Connect each two adjacent points with a curve (Smoothly).
5. Close the curve from the left to the lower limit of first class boundary.

Solution: Find the lower class limit and cumulative frequency. Then plot the lower class limit against the cumulative frequency.

Class	Frequency	Lower Limit of Class	Cumulative Frequency
10 – 20	2	10	2
20 – 30	5	20	7
30 – 40	9	30	16
40 – 50	14	40	30
50 – 60	7	50	37



Measures of Central Tendency

A measure of central tendency is very important tool that refer to the centre of a histogram or a frequency distribution curve. In This section we will discuss three measures of central tendency such as the **mean**, the **median**, and the **mode** for the two cases (grouped and ungrouped data sets).

The Mean

The most commonly used measure of central tendency is called mean (or the average).

The Mean for Ungrouped Data

Mean for sample data: $\overline{x} = \frac{\sum x}{n}$

Where $\sum x$ is the sum of all values, N is the population size, and n is the sample size, μ is the population mean, and $\overline{x}$ is the sample mean

Example

Find the mean score of 10 students in a midterm exam in a class if their scores are:

25	27	30	23	16	27	29	14	20	28

To sum all scores

$$\begin{aligned}\sum x &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \\ &= 25 + 27 + 30 + 23 + 16 + 27 + 29 + 14 + 20 + 28 \\ &= 239\end{aligned}$$

Since the given data set includes all scores of the students, it represents the population. Hence, $N = 10$. We have

$$\mu = \frac{\sum x}{N} = \frac{239}{10} = 23.9$$

The Median for Ungrouped Data

The median is the value of the middle term in a data set that has been ranked in increasing or decreasing order.

Steps

- Rank the given data sets (in increasing or decreasing order)
- Find the middle term for the ranked data set that obtained in step 1.
- The value of this term represents the median.

The median of the ranked data $x_1, x_2, \ldots, x_n$ is given by

$$\text{Median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example: Find the median of the following data

$$312, 257, 421, 289, 526, 374, 497$$

Solution: First, the data set after we have ranked in increasing order is:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
257	289	312	374	421	497	526

Median=374

The Mode
The mode is the value that occurs most often in a data set.

1. What is the mode for given data?

   77 69 74 81 71 68 74 73

2. What is the mode for given data?

   77 69 68 74 81 71 68 74 73

Solution:

1. Mode = 74 (this number occurs twice): ***Unimodal***

2. Mode = 68 and 74: ***Bimodal***

The Mean for Grouped Data

Example
The following table gives the frequency distribution of the number of orders received each day during the past 50 days at the office of a mail-order company. Calculate the mean

Formula

$$\bar{x} = \frac{\sum fx}{n}$$

Number of order	f
10 – 12	4
13 – 15	12
16 – 18	20
19 – 21	14
	$n = 50$

Number of order	f	x	fx
10 – 12	**4**	**11**	**44**
13 – 15	**12**	14	168
16 – 18	**20**	17	340
19 – 21	**14**	20	280
	$n = 50$		**= 832**

X is the midpoint of the class. It is adding the class limits and divide by 2.

$$\bar{x} = \frac{\sum fx}{n} = \frac{832}{50} = 16.64$$

The Median for Grouped Data

Step 1: Construct the cumulative frequency distribution.
Step 2: Decide the class that contain the median.
Class Median is the first class with the value of cumulative frequency equal at least n/2.
Step 3: Find the median by using the following formula:

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

Where:
n = the **total frequency**
F = the **cumulative frequency *before*** class median
f_m = the **frequency** of the class median
i = the class width
L_m = the **lower boundary** of the class median

Example: Based on the grouped data below, find the median:

Time to travel to work	Frequency
1 – 10	**8**
11 – 20	**14**
21 – 30	**12**
31 – 40	**9**
41 – 50	**7**

Solution:

1st Step: Construct the cumulative frequency distribution

Time to travel to work	Frequency	Cumulative Frequency
1 – 10	8	8
11 – 20	14	22
21 – 30	12	34
31 – 40	9	43
41 – 50	7	50

$\frac{n}{2} = \frac{50}{2} = 25 \longrightarrow$ class median is the 3rd class

So, $F = 22$, $f_m = 12$, $L_m = 20.5$ and $i = 10$

Therefore,

$$\text{Median} = L_m + \left(\frac{\frac{n}{2} - F}{f_m} \right) i$$

$$= 21.5 + \left(\frac{25 - 22}{12} \right) 10$$

$$= 24$$

Thus, 25 persons take less than 24 minutes to travel to work and another 25 persons take more than 24 minutes to travel to work.

The Mode for Grouped Data

Formula

$$\text{Mode} = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

Where:

i is the class width

Δ_1 is the difference between the frequency of class mode and the frequency of the class **after** the class mode

Δ_2 is the difference between the frequency of class mode and the frequency of the class **before** the class mode

L_{mo} is the **lower boundary** of class mode

Example: Based on the grouped data below, find the mode

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

Based on the table,

$L_{mo} = 10.5$, $\Delta_1 = (14 - 8) = 6$, $\Delta_2 = (14 - 12) = 2$ and $i = 10$

$$\text{Mode} = 10.5 + \left(\frac{6}{6 + 2} \right) 10 = 17.5$$