Introduction to Statistics

Why Statistics?

- To develop an appreciation for variability and how it effects products and processes.
- Study methods that can be used to help solve problems,
- build knowledge and continuously improve products and processes.
- Build an appreciation for the advantages and limitations of informed observation and experimentation.

Why Statistics?

- Determine how to analyze data from designed experiments in order to build knowledge and continuously improve.
- Develop an understanding of some basic ideas of statistical reliability and the analysis data.

Data and Statistics

Data consists of information coming from observations, counts, measurements, or responses.

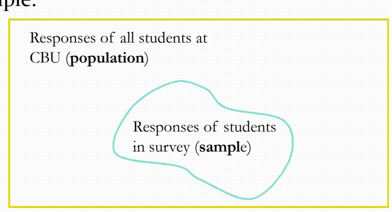
Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

A **population** is the collection of *all* outcomes, responses, measurement, or counts that are of interest.

A **sample** is a subset of a population.

Populations & Samples

Example: In a recent survey, 250 university students at CBU were asked if they smoked cigarettes regularly. 35 of the students said yes. Identify the population and the sample.



Parameters & Statistics

A **parameter** is a numerical description of a *population* characteristic.

A **statistic** is a numerical description of a *sample* characteristic.

Parameter ⟶ **Population**

Statistic ⟶ **Sample**

Parameters & Statistics

Example:
Decide whether the numerical value describes a population parameter or a sample statistic.

a.) A recent survey of a sample of 450 university students reported that the average weekly income for students is K600,000.

Because the average of K600,000 is based on a sample, this is a sample statistic.

b.) The average weekly income for all students is K500, 000

Because the average of K500, 000 is based on a population, this is a population parameter.

Types of sampling techniques

- Random Sampling
  - Sampling in which the data is collected using chance methods or random numbers.
- Systematic Sampling
  - Sampling in which data is obtained by selecting every *k*th object.

Types of sampling techniques

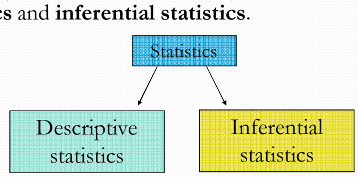
- Convenience Sampling
  - Sampling in which data which is readily available is used.
- Stratified Sampling
  - Sampling in which the population is divided into groups (called strata) according to some characteristic.
  - Each of these strata is then sampled using one of the other sampling techniques.

Types of sampling techniques

- Cluster Sampling
  - Sampling in which the population is divided into groups (usually geographically).
  - Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Branches of Statistics

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics**.



Descriptive statistics: Involves the organization, summarization, and display of data.

Inferential statistics: Involves using a sample to draw conclusions about a population.

Descriptive and Inferential Statistics

Example:
In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep. Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics.

The statement "four times more likely to answer incorrectly" is a descriptive statistic. An inference drawn from the sample is that all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.

Descriptive and Inferential Statistics

Note: The development of **Inferential Statistics** has occurred only since the early 1900's.

Examples:

1. The medical team that develops a new vaccine for a disease is interested in what would happen if the vaccine were administered to all people in the population.
2. The marketing expert may test a product in a few "representative" areas, from the resulting information, he/she will draw conclusion about what would happen if the product were made available to all potential customers.

The Essential Elements of a Statistical Problem

The objective of statistics is to make inferences (predictions, and/or decisions) about a population based upon the information contained in a sample. A statistical problem involves the following

1. A clear definition of the objectives of the experiment and the pertinent population. For example, clear specification of the questions to be answered.
2. The design of experiment or sampling procedure. This element is important because data cost money and time.

The Essential Elements of a Statistical Problem

3. The collection and analysis of data.
4. The procedure for making inferences about the population based upon the sample information.
5. The provision of a measure of goodness (reliability) of the inference. The most important step, because without the reliability the inference has no meaning and is useless.

Note, above steps to solve any statistical problem are sequential.

Data Classification

Types of Data

Data sets can consist of two types of data: **qualitative data (Attribute)** and **quantitative (Numerical) data**.

Data

Qualitative Data

Consists of attributes, labels, or nonnumerical entries.

Quantitative Data

Consists of numerical measurements or counts.

Types of Data

DATA: Consist of information coming from observations, counts, measurements or responses.

1. **Attribute (Qualitative):** Consists of qualities such as religion, sex, color, etc. **No way to rank this type of data**.
2. **Numerical Data (Quantitative):** Consists of numbers representing counts or measurements. **Can be ranked**. There are two types of numerical data.

Types of Data

a. **Discrete Data:** Can take on a finite number of values or a countable infinity (as many values as there are whole numbers such as 0, 1, 2..). Examples:

1. Number of kids in the family.
2. Number of students in the class.
3. Number of calls received by the switch board each day
4. Number of flaws in a yard of material.

Types of Data

b. Continuous Data: Can assume all possible values within a range of values without gaps, interruptions, or jumps. Examples: all kind of measurements such as, time, weight, distance, etc.

1. Yard of material.
2. Height and weight of students in a class.
3. Duration of a call to a switch board.
4. Body temperature.

Qualitative and Quantitative Data

Example:The grade point averages of five students are listed in the table. Which data are qualitative data and which are quantitative data?

Student	GPA
Sally	3.22
Bob	3.98
Cindy	2.75
Mark	2.24
Kathy	3.84

Qualitative data ←   → Quantitative data

Levels of Measurement

The level of measurement determines which statistical calculations are meaningful. The four levels of measurement are: **nominal**, **ordinal**, **interval**, and **ratio**.

Levels of Measurement → **Nominal**, **Ordinal**, **Interval**, **Ratio**

Lowest to highest

Nominal Level of Measurement

Data at the **nominal level of measurement** are qualitative only.

Levels of Measurement → **Nominal**

Calculated using names, labels, or qualities. No mathematical computations can be made at this level.

Colors in the US flag

Names of students in your class

Textbooks you are using this semester

Nominal Scale

- Numbers are used simply to label groups or classes.
- For example, gender
  - 1 = male, 2 = female.
- Color of eyes of a person
  - 1 = blue, 2 = green, 3 = brown

Ordinal Level of Measurement

Data at the **ordinal level of measurement** are qualitative or quantitative.

Levels of Measurement → **Ordinal**

Arranged in order, but differences between data entries are not meaningful.

- Class standings: freshman, sophomore, junior, senior
- Numbers on the back of each player's shirt
- Top 50 songs played on the radio

Ordinal Scale

- In addition to classification
  - members can be ordered according to relative size or quality.
  - For example, products ranked by consumers 1 = best, 2 = second best etc.

Interval Level of Measurement

Data at the **interval level of measurement** are quantitative. A zero entry simply represents a position on a scale; the entry is not an inherent zero.

Levels of Measurement → **Interval**

Arranged in order, the differences between data entries can be calculated.

- Temperatures
- Years on a timeline
- Atlanta Braves World Series victories

Ratio Level of Measurement

Data at the **ratio level of measurement** are similar to the interval level, but a zero entry is meaningful.

Levels of Measurement → **Ratio**

A ratio of two data values can be formed so one data value can be expressed as a ratio.

- Ages
- Grade point averages
- Weights

Summary of Levels of Measurement

Level of measurement	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Displaying data

FREQUENCY DISTRIBUTIONS

- Frequency (f) is used to describe the number of times a value or a range of values occurs in a data set.
- Cumulative frequencies are used to describe the number of observations less than, or greater than a specific value

Frequency Measures

- Absolute frequency is the number of times a value or range of values occurs in a data set.
- The relative frequency is found by dividing the absolute frequency by the total number of observations (n).
- The cumulative frequency is the successive sums of absolute frequencies.

Frequency distribution

- The cumulative relative frequency is the successive sum of cumulative frequencies divided by the total number of observations.
- Frequency distributions are portrayed as Frequency tables, Histograms or Polygons.

Frequency distribution table

- The first step in drawing a frequency distribution is to construct a frequency table.
- A frequency table is a way of organizing the data
- by listing every possible score as a column of numbers and
- the frequency of occurrence of each score as another.

Frequency distribution table

- Computing the frequency of a score is simply a matter of counting the number of times that score appears in the set of data.
- It is necessary to include scores with zero frequency in order to draw the frequency polygons correctly.
- For example, consider the following set of 15 scores which were obtained by asking a class of students their shoe size, shoe width, and sex (male or female).

Shoe Size	Shoe Width	Gender
10.5	B	M
6.0	B	F
9.5	D	M
8.5	A	F
7.0	B	F
10.5	C	M
7.0	C	F
8.5	D	M
6.5	B	F
9.5	C	M
7.0	B	F
7.5	B	F
9.0	D	M
6.5	A	F
7.5	B	F

Frequency distribution table

- To construct a frequency table,
- start with the smallest shoe size and list all shoe sizes as a column of numbers.
- The frequency of occurrence of that shoe size is written to the right.

Shoe Size	Absolute Frequency	Cumulative Frequency	Relative Freq
6.0	1	1	0.07
6.5	2	3	0.13
7.0	3	6	0.2
7.5	2	8	0.13
8.0	0	8	0
8.5	2	10	0.13
9.0	1	11	0.07
9.5	2	13	0.13
10.0	0	13	0
10.5	2	15	0.13
Total	15		

Frequency distribution table

- Note that the sum of the column of frequencies is equal to the number of scores or size of the sample (N = 15).
- This is a necessary, but not sufficient, property in order to insure that the frequency table has been correctly calculated.
- It is not sufficient because two errors could have been made, canceling each other out.

Grouped Frequency Distributions

- These distributions used for data sets that contain a large number of observations.
- The data is grouped into a number of classes.

Grouped Frequency Distributions

- Guidelines for classes
  - There should be between 5 and 20 classes.
  - The class width should be an odd number.
    - This will guarantee that the class midpoints are integers instead of decimals.

Grouped Frequency Distributions

- The classes must be mutually exclusive.
  - This means that no data value can fall into two different classes
- The classes must be all inclusive or exhaustive.
  - This means that all data values must be included.

Grouped Frequency Distributions

- The classes must be continuous.
  - There are no gaps in a frequency distribution.
  - Classes that have no values in them must be included (unless it's the first or last class which are dropped).

Grouped Frequency Distributions

- The classes must be equal in width.
  - The exception here is the first or last class.
  - It is possible to have an "below ..." or "... and above" class.
  - This is often used with ages.

Creating a Grouped Frequency Distribution

- Find the largest and smallest values
- Compute the Range = Maximum - Minimum
- Select the number of classes desired.
  - This is usually between 5 and 20.

Creating a Grouped Frequency Distribution

- Find the class width by dividing the range by the number of classes and rounding up.
  - There are two things to be careful of here. You must *round up*, not off.
  - If the range divided by the number of classes gives an integer value (no remainder), then you can either add one to the number of classes or add one to the class width.

Creating a Grouped Frequency Distribution

- Pick a suitable starting point less than or equal to the minimum value.
  - Your starting point is the lower limit of the first class.
  - Continue to add the class width to this lower limit to get the rest of the lower limits.

Creating a Grouped Frequency Distribution

- To find the upper limit of the first class, subtract one from the lower limit of the second class.
  - Then continue to add the class width to this upper limit to find the rest of the upper limits.
- Find the boundaries by subtracting 0.5 units from the lower limits and adding 0.5 units from the upper limits.
  - The boundaries are also half-way between the upper limit of one class and the lower limit of the next class.

Creating a Grouped Frequency Distribution

- Tally the data.
- Find the frequencies.
- Find the cumulative frequencies.
- If necessary, find the relative frequencies and/or relative cumulative frequencies.

Example

- Thirty AA batteries were tested to determine how long they would last. The results to the nearest minute were recorded as follows:
- 423, 365, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390

Example

- After following the steps we have the following classes:
- 360-368, 369-377, 378-386, 387-395, 396-404, 405-413, 414-422, 423-432

Life of AA batteries, in Minutes

Battery life, minutes (x)	Absolute frequency	Cummulative frequency
360-368	1	1
369-377	4	5
378-386	3	8
387-395	7	15
396-404	4	19
405-413	5	24
414-422	3	27
423-432	3	30

Visualizing Data

- The three most commonly used graphs in research are:
- The histogram.
- The frequency polygon.
- The cumulative frequency graph, or ogive (pronounced o-jive).

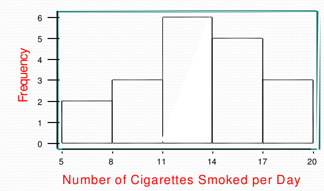
histogram

- The histogram is a graph that displays the data by using vertical bars of various heights to represent the frequencies.

histogram

- A histogram is drawn by plotting the scores (midpoints) on the X-axis and the frequencies on the Y-axis.
- A bar is drawn for each score value, the width of the bar corresponding to the real limits of the interval and the height corresponding to the frequency of the occurrence of the score value.
- An example histogram is presented below

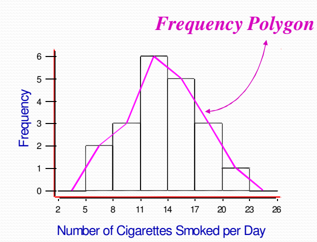
Example of a Histogram



frequency polygon

- A frequency polygon is a graph that displays the data by using lines that connect points plotted for frequencies at the midpoint of classes.
- The frequencies represent the heights of the midpoints.

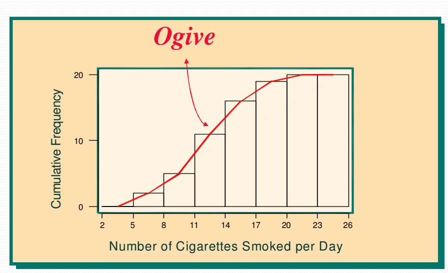
Example of a Frequency Polygon



cumulative frequency graph

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

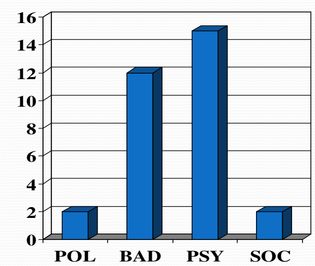
Example of an Ogive



Other Types of Graphs

- A bar chart or bar graph is a chart with rectangular bars with lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.
- Bar graph use frequency distributions of discrete variables, often nominal or ordinal data.
- Bars represent separate groups, so they should be separated

Bar graphs



- Number of students in statistics class from each of four majors, fall, 2005

Other Types of Graphs

- Pie graph - A pie graph is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

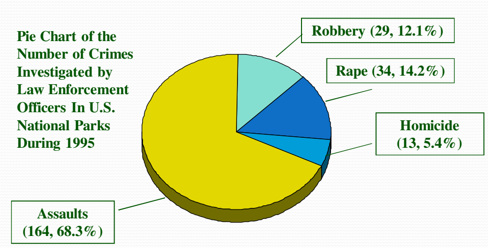
A *pie chart*

- A *pie chart* (or a *circle graph*) is a circular chart divided into sectors, illustrating proportion.
- In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents.

A *pie chart*

- When angles are measured with 1 turn as unit then a number of percent is identified with the same number of centiturns.
- Together, the sectors create a full disk.
- It is named for its resemblance to a pie which has been sliced.

Other Types of Graphs - Pie Graph



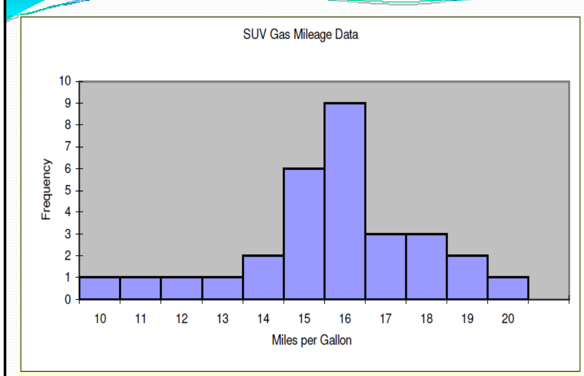
Measures of Central Tendency

- There are three main measures of central tendency: the mean, median, and mode.
- The purpose of measures of central tendency is to identify the location of the center of various distributions.

- For example, let's consider the data below.
- This data represents the number of miles per gallon that 30 selected four-wheel drive sports utility vehicles obtained in city driving.

12	17	16	14	16	18
16	18	17	16	17	15
15	16	16	15	16	19
10	14	15	11	15	15
19	13	16	18	16	20

- In its current form it is difficult to determine where the center for the above data set lies.
- Thus, one way to help us get a better idea as to where the center of a distribution is located is to graph the data.
- Because the data is numerical, the most appropriate method of graphing the data would be to create a histogram.
- The histogram for the gas mileage data is given below



- If we rely on sight alone, it seems that the middle of the distributions lies at around 15 to 16 miles per gallon;
- however, because our senses can sometimes deceive us,
- we want to be a little more scientific in our methodology.

Mode

- The mode is the observation that occurs most frequently.
- Thus, to find the mode for the above data set we simply locate the observation that occurs most frequently.
- In this case, the number 16 occurs 9 times, which is more than any other observation.
- Therefore, the mode of the data is 16.

The Median

- The median is the middle observation in the data.
- This means that 50% of the data is below the median and 50% of the data is above the median.
- To find the median, we must first organize the data in order from the smallest to the largest observation.
- For example, the above gas mileage data would take on the following form:

- 10 11 12 13 14 14 15 15 15 15 15 15 16 16 16 16 16 16 16 16 16 17 17 17 18 18 18 19 19 20

- To find the middle, or halfway point, is to divide n+1 by 2.
- In this case we have 30 observations so our halfway point is 15.5
- Ie. 30+1/2=15.5
- Next, to find the center, we count in 15.5 spaces or observations from the starting or ending points of the data.

- This will put us directly between the two highlighted 16's.
- 10 11 12 13 14 14 15 15 15 15 15 15 16 16 **16 16** 16 16 16 16 16 17 17 17 18 18 18 19 19 20
- We want the number between these two 16's

- To overcome this problem we add up the two middle points and divide by 2 (essentially taking the average of the two middle observations).
- Thus, the median for the data set is 16 + 16/2=16

The Mean

- The mean is the arithmetic average of all the observations in the data.
- It is also the fulcrum or, “balancing point”, of the data.
- For instance, if you were to place the histogram of the gas mileage data onto a seesaw,
- the mean would be the point that would allow the histogram to be perfectly balanced.

- The mean is found by adding up all of the observations and dividing by the total number of observations,
- either N or n depending upon whether you are dealing with the population or sample.

- The formula for the population and sample mean are:

$$u = \sum \frac{x_i}{N}$$

Sample mean

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

Population mean

- For the mean of a grouped distribution,
- we assume that all of the data points in a given interval are located at the midpoint of that interval.
- If x represents the midpoints, f denotes the frequencies, and $n = \sum f$

Is the total number of data points, then

$$\bar{x} = \frac{\sum(xf)}{n}.$$

Measures of Central Tendency

Use the formula

$$\mu = \Sigma_i(f_i * x_i) / \Sigma_i f_i$$

Where x_i = the midpoint of the i^{th} class

and f_i = the number of items in the i^{th} class

Measures of Central Tendency

From the table we obtain

Class	Class Midpoint (x)	Total (f)	Frequency	f*x
64.5 - 69.5	67	6	0.100	402
69.5 – 74.5	72	11	0. 183	792
74.5 – 79.5	77	20	0.333	1540
79.5 – 84.5	82	13	0.217	1066
84.5 – 89.5	87	9	0.150	783
89.5 – 94.5	92	1	0.0167	92
		60		4675

$$\mu = \Sigma_i(f_i * x_i) / \Sigma_i f_i = 4675/60 = 77.917$$

Mode of Grouped Data

- The mode of given data is the observation which is repeated maximum number time.
- This can be found just by observing the data carefully when the data is ungrouped.

Mode of Grouped Data

- For finding the mode of grouped data, first of all we have to determine the modal class.
- The class interval whose frequency is maximum is known by this name.
- The mode lies in between this class.
- Then the mode is calculated by the following formula.

$$l + \left(\frac{f_1 - f_o}{2f_1 - f_o - f_2}\right) x h$$

Mode of Grouped Data

- Here,
- l = lower limit of modal class
- f_1 = frequency of modal class
- f_0 = frequency of class preceding the modal class.
- f_2 = frequency of class succeeding the modal class
- h = size of class interval.

Example:

- Find the mode of following data

Class interval (C. I)	5 - 10	10 - 15	15 - 20	20 -25	25 - 30
Frequency (f_i)	3	5	7	2	4

Solution:

- Here frequency of class interval 15 - 20 is maximum.
- So, it is the modal class
- l = the lower limit of modal class = 15
- f_1 = frequency of modal class = 7
- f_0 = frequency of class preceding the modal class = 5
- f_2 = frequency of class succeeding the modal class = 2
- h = size of class intervals = 5

So,

$$\text{Mode} = l + \left(\frac{f_1 - f_o}{2f_1 - f_o - f_2}\right) x h$$

Mode = 15 + [(7 - 5) / (2 x 7 - 5 - 2)] x 5

Mode = 15 + [2 / (14 - 7)] x 5

Mode = 15 + (2 / 7) x 5

Mode = 15 + (10 / 7)

Mode = 15 + 1.42
Mode = 16.42

Median of Grouped Data

- On arranging the data in ascending or descending order median is the middle – most observation.
- If the number of observations are odd then the median is (n+1 / 2)th observation
- where 'n' is the number of observations.
- If number of observations are even then median is the average of (n / 2)th and (n / 2 + 1)th observation.

Example:

- Find the median of the given data.

Wages of workers	3800	4100	4400	4900	5200	5500	6000
Number of workers	12	13	25	17	15	12	6

Solution

Wages of workers	Number of workers
3800	12
4100	13
4400	25
4900	17
5200	15
5500	12
6000	6
	Total = 100

Solution

- Here, the number of observations (n) = 100
- This is an even number, so the median is average of (n / 2)th and (n / 2 + 1)th observations
- i.e. average of (100 / 2)th and [(100 / 2) + 1]th observation.
- i.e. average of 50th and 51th observations.
- To find these observations let us arrange the data in the following manner.

Wages of workers	Number of workers
3800	12
upto 4100	12 +13 = 25
upto 4400	25 + 25 = 50
upto 4900	50 + 17 = 67
upto 5200	67 + 15 = 82
upto 5500	82 + 12 = 94
upto 6000	94 + 6 = 100

Median of unGrouped Data

- The frequencies arranged in above manner are known as cumulative frequencies.
- So, the 50th observation is 4400 and 51th observation is 4900
- Median = 4400 + 4900 / 2
- Median = 9300 / 2
- Median = 4650
- This means 50% workers got wages less than Rs. 4650 and another 50% got more than Rs. 4650.

Less and More than cumulative frequencies

- We will know the method for calculating median of grouped data. But before that let
- us know about the cumulative frequency of
- (a) Less than type
- (b) More than type
- for the given grouped data

The grouped data is

Marks	Number of Student
0 -10	2
10 -20	12
20 - 30	22
30 - 40	8
40 - 50	6

Less than cumulative frequencies

- Let us construct a cumulative frequency table of less than type for the above data.
- Here 2 students got the marks between 0 and 10 which means 2 students have marks less than 10.
- Now 12 students got marks between 10 - 20.
- So the students who got marks less than 20 are (2 + 12) i.e. 14 students.
- Proceeding in the similar way, we get the following cumulative frequency table.

the table is known as cumulative frequency table of less than type

Marks	Number of Student
Less than10	2
Less than 20	2 +12 = 14
Less than 30	14 + 22 = 36
Less than 40	36 + 8 = 44
Less than 50	44 + 6 = 50

More than cumulative frequencies

- Let us know the method of calculating cumulative frequency table of more than type for the above data.
- Here, all students got marks more than or equal to 0.
- Which means 50 students got more than or equal to 0 marks.
- Since 2 students got less than 10 marks.

More than cumulative frequencies

- Thus (50 -2) i.e. 48 students got more than equal to 10 marks.
- Preceding in the same way 48 -2 = 36 students got more than 20 marks.
- Hence, we get the cumulative frequency distribution table of mare than type.

More than cumulative frequencies

Marks	Number of Student
More than 0	50
More than 10	50 -2 = 48
Less than 20	48 - 12 = 36
Less than 30	36 - 22 = 14
Less than 40	14 - 6 = 6

calculating median of grouped data.

- Now, let us know the method of calculating median of grouped data.
- For this we require to calculate cumulative frequencies of less than type.
- After than we calculate n / 2,

•with its help we determine the class whose cumulative frequency is nearly equal to n /2.
This class is known as median class.
Then, the median is calculated by the following formula.

$$\textbf{Median} = l + \left(\frac{\frac{n}{2} - cf}{f}\right) \times h$$

calculating median of grouped data.

- l = lower limit of median class
- cf = cumulative frequency of class prior to median class.
- f = frequency of median class.
- h = class size.

calculating median of grouped data.

- Calculate the median of following grouped data.

Marks	Number of Student
0 -10	2
10 -20	12
20 - 30	22
30 - 40	8
40 - 50	6

Solution

Marks (C. I.)	f	cf
0 -10	2	2
10 -20	12	14
20 - 30	22	36 ?
30 - 40	8	44
40 - 50	6	50

calculating median of grouped data.

- Here n / 2 = 50 / 2 = 25
- So, 20 -30 is the median class.
- Now, l =20
- h = 10
- cf = 14
- f = 22

calculating median of grouped data.

Median = $l + \left(\frac{\frac{n}{2} - cf}{f} \right) \times h$

Median = 20 + [(25 - 14) / 22] x 10

Median = 20 + (11 / 22) x 10

Median = 20 + 5

Median = 25

calculating median of grouped data.

- This means 50% of the students got less than 25 marks and other 50% got more than 25 marks.