Statistics

Introduction to Statistics

Ungrouped Data

Ungrouped data is data in its original or raw form. The observations are not classified into groups.

For example, the ages of everyone present in a classroom of kindergarten kids with the teacher is as follows:

3, 3, 4, 3, 5, 4, 3, 3, 4, 3, 3, 3, 3, 4, 3, 27.

This data shows that there is one adult present in this class and that is the teacher. Ungrouped data is easy to work when the data set is small.

Grouped Data

In grouped data, observations are organized in groups.

For example, a class of students got different marks in a school exam. The data is tabulated as follows:

Mark Interval	No. of students
0 – 20	13
21 – 40	9
41 – 60	36
61 – 80	32
81 – 100	10

This shows how many students got the particular mark range. Grouped data is easier to work with when large amount of data is present.

Frequency

Frequency is the number of times a particular observation occurs in a data.

Class Interval

Data can be grouped into class intervals such that all observations in that range belong to that class.

Class width = upper class limit - lower class limit

Mean

Finding mean for Grouped Data when class Intervals are not given

For grouped data without class intervals,
Mean, $\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$ where f_i is the frequency of i^{th} observation x_i.

Finding mean for Grouped Data when class Intervals are given

For grouped data with class intervals,
Mean, $\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$
Where f_i is the frequency of i^{th} class whose class mark is x_i.
Class mark = $\frac{Upper\ class\ limit+Lower\ class\ limit}{2}$

Direct method of finding mean

Step 1: Classify the **data into intervals** and find the corresponding **frequency of each class**.

Step 2: Find the **class mark** by taking the **midpoint of the upper and lower class limits.**

Step 3: Tabulate the product of class mark and its corresponding frequency for each class. Calculate their sum ($\sum x_i f_i$).

Step 4: Divide the above sum by the sum of frequencies ($\sum f_i$) to get the mean.

Assumed mean method of finding mean

Step 1: Classify the data into intervals and find the corresponding frequency of each class.

Step 2: Find the class mark by taking the midpoint of the upper and lower class limits.

Step 3: Take one of the x_i's (usually one in the middle) as assumed mean and denote it by $'a'$.

Step 4: Find the deviation of $'a'$ from each of the $x_i's$
$d_i = x_i - a$

Step 5: Find the mean of the deviations

$\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$

Step 6: Calculate the mean as
$\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$

Relation between Mean of deviations and mean

$d_i = x_i - a$
Summing over all $x_i's$,
$\sum d_i = \sum x_i - \sum a$
Dividing throughout by $\sum f_i = n$, Where $'n'$ is the total number of observations.
$\bar{d} = \bar{x} - a$
$\Rightarrow \bar{x} - \bar{d} = a$

Step-Deviation method of finding mean

Step 1: Classify the data into intervals and find the corresponding frequency of each class.

Step 2: Find the class mark by taking the midpoint of the upper and lower class limits.

Step 3: Take one of the $x_i's$ (usually one in the middle) as assumed mean and denote it by $'a'$.

Step 4: Find the deviation of a from each of the $x_i's$
$d_i = x_i - a$

Step 5: Divide all deviations $-d_i$ by the class width (h) to get $u_i's$.
$u_i = \frac{x_i - a}{h}$

Step 6: Find the mean of $u_i's$
$\bar{u} = \frac{\sum f_i u_i}{\sum f_i}$

Step 7: Calculate the mean as
$\bar{x} = a + h \times \frac{\sum f_i u_i}{f_i} = a + h\bar{u}$

Relation between mean of Step- Deviations (u) and mean

$u_i = \frac{x_i - a}{h}$

$\bar{u} = \frac{\sum f_i \frac{x_i - a}{h}}{\sum f_i}$

$\bar{u} = \frac{1}{h} \times \frac{\sum f_i x_i - a \sum f_i}{\sum f_i}$

$\bar{u} = \frac{1}{h} \times (\bar{x} - a)$

Important relations between methods of finding mean

- All three methods of finding mean yield the same result.
- Step deviation method is easier to apply if all the deviations have a common factor.
- Assumed mean method and step deviation method are simplified versions of the direct method.

Median

Finding median of Grouped Data when class Intervals are not given

Step 1: Tabulate the observations and the corresponding frequency in ascending or descending order.

Step 2: Add the cumulative frequency column to the table by finding the cumulative frequency up to each observation.

Step 3: If the number of observations is odd, the median is the observation whose cumulative frequency is just greater than or equal to $(\frac{n+1}{2})$.

If the number of observations is even, the median is the average of observations whose cumulative frequency is just greater than or equal to $(\frac{n}{2})$ and $\frac{n}{2} + 1$.

Cumulative Frequency

Cumulative frequency is obtained by adding all the frequencies up to a certain point.

Finding median for Grouped Data when class Intervals are given

Step 1: find the cumulative frequency for all class intervals.

Step 2: the median class is the class whose cumulative frequency is greater than or nearest to $\frac{n}{2}$, where n is the number of observations.

Step 3: $Median = l + \frac{\frac{n}{2} - cf}{f} \times h$

Where,
$l =$lower limit of median class,
$n =$ number of observations,
$cf =$ cumulative frequency of class preceding the median class,
$f =$ frequency of median class,
$h =$ class size (assuming class size to be equal).

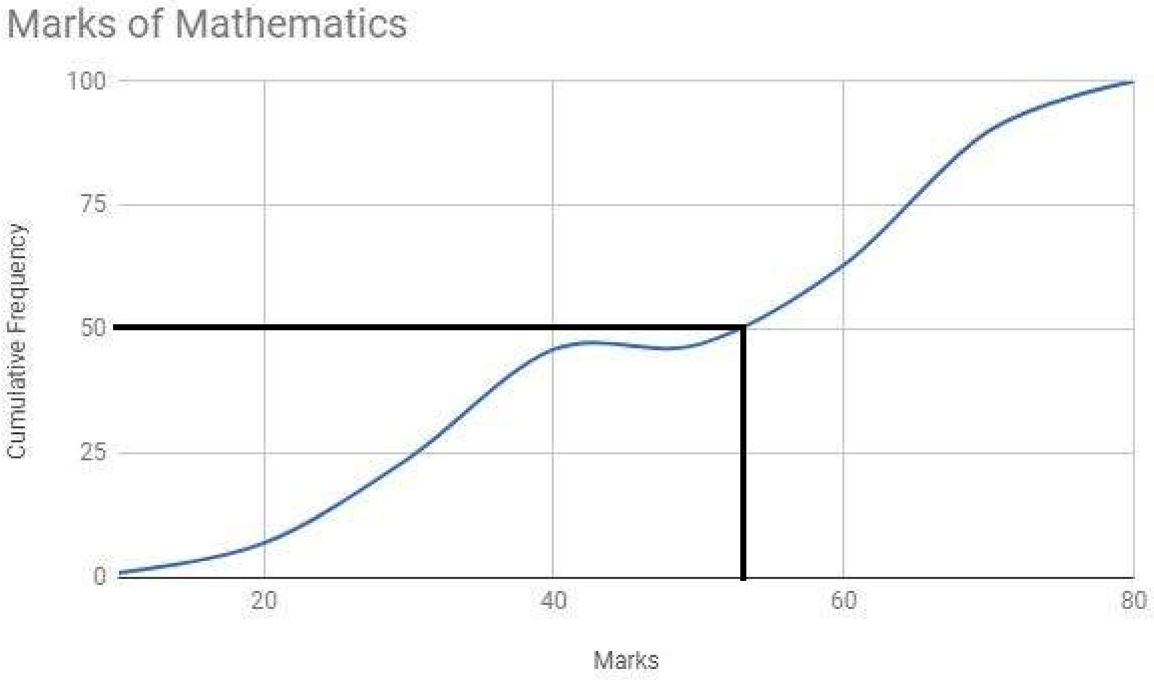
Cumulative Frequency distribution of less than type

Cumulative frequency of the less than type indicates the number of observations which are less than or equal to a particular observation.

Cumulative Frequency distribution of more than type

Cumulative frequency of more than type indicates the number of observations which are greater than or equal to a particular observation.

Visualising formula for median graphically



Median from Cumulative Frequency Curve

Step 1: Identify the median class.

Step 2: Mark cumulative frequencies on the y-axis and observations on the x-axis corresponding to the median class.

Step 3: Draw a straight line graph joining the extremes of class and cumulative frequencies.

Step 4: Identify the point on the graph corresponding to $cf = \frac{n}{2}$.

Step 5: Drop a perpendicular from this point on to the x-axis.

Ogive of less than type

The graph of a cumulative frequency distribution of the less than type is called an '**ogive of the less than type**'.

Ogive of more than type

The graph of a cumulative frequency distribution of the more than type is called an '**ogive of the more than type**'.

Relation between the less than and more than type curves

The point of intersection of the ogives of more than and less than types gives the median of the grouped frequency distribution.

Mode

Finding mode for Grouped Data wen class intervals are not given

In grouped data without class intervals, the observation having the largest frequency is the mode.

Finding mode for Ungrouped Data

For ungrouped data, the mode can be found out by counting the observations and using tally marks to construct a frequency table.
The observation having the largest frequency is the **mode**.

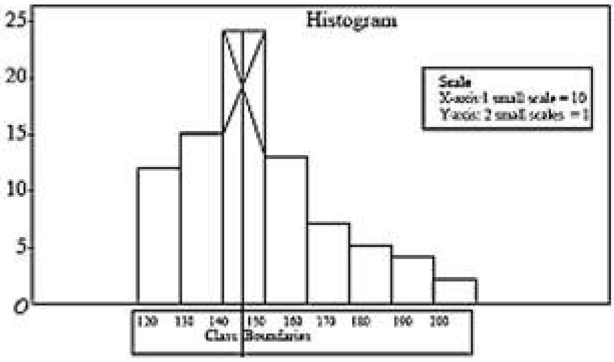
Finding mode for Grouped Data when class intervals are given

For, grouped data, the class having the highest frequency is called the modal class. Mode can be calculated using the following formula. Formula valid for equal class intervals and when the modal class is unique.

$Mode = l + (\frac{f_1 - f_0}{2f_1 - f_0 - f_2}) \times h$
$l =$ lower limit of modal class
$h =$ class width
$f_1 =$ 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

Visualising formula for mode graphically

Graphical Method for finding mode



Step 1: Express the class intervals and frequencies as a histogram.

Step 2: Join the top corners of the modal class to the diagonally opposite corners of the adjacent classes

Step 3: Drop a perpendicular from the point of intersection of the above on the horizontal x-axis.

Measures of Central Tendency for Grouped Data

i) Mean is the average of a set of observations.
ii) Median is the middle value of a set of observations.
iii) Mode is the most common observation.

Best suited measure of central tendency in different cases and the Empirical relationship between them

i) The mean takes into account all the observations and lies between the extremes. It enables us to compare distributions.

ii) In problems where individual observations are not important, and we wish to find out a 'typical' observation where half the observations are below and half the observations are above, the median is more appropriate. Median disregards the extreme values.

iii) In situations which require establishing the most frequent value or most popular item, the **mode** is the best choice.

Mean, mode and median are connected by the empirical relationship

3 Median = Mode + 2 Mean