Semester: V Subject: Statistics for AIDS Academic Year: 2023-2024

Estimates of Location :- Variable with measured data might have thousands of distinct values. A basic step in exploring your data is getting a "typical value" for each variable (i.e) an estimate of where the most of the data is located (i.e) its central tendancy).

* Mean
* Median
* Robust
* Outlier

The central tendancy can be found by the following:-

Mean:

The most basic estimate of location is the mean, or average value. The mean is the sum of all values divided by the number of values. The formula to compute the mean for a set of n values $x_1, x_2 \cdots x_n$ is

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

Consider the following set of numbers: {3 5 12}.
The mean is $(3+5+12) = 11/4 = 2.75$

~~Example~~:-

Median:-

* The median is the middle number on a sorted list of the data.
* The middle or the central number becomes the median.

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023-2024

* To calculate the median in an even number of data set, select the middle two numbers and find the average of it.
* If it is an odd number of data set, then the middle number is considered as median.

Example:-

Consider that we have a data set [1,2,3,4,5]

The mean for this dataset = $\frac{1+2+3+4+5}{5} = \frac{15}{5} = \boxed{3}$

If suppose a additional data point 60 is added to the existing dataset. The dataset becomes like this: [1,2,3,4,5,50]. Now the mean for the dataset is $\frac{1+2+3+4+5+50}{5} = \frac{65}{5} = \boxed{13}$

Just by adding one value in the dataset, results in a complete different format of distribution and this is harmful to do any statistics. The predictions wont be accurate. In this case we calculate the median. The median value for the above example is: $\frac{3+4}{2} = \boxed{3.5}$. Now the value 3.5 is closer to the mean 3.

Compared to the mean, which uses all the observations, the median depends only on the values in the centre of

the sorted data. This might seem to be a disadvantage, but there are many instances in which the median is a better metric for location.

In the above example the problem caused was due to the value 50. This value is known as outlier value.

An outlier is any value that is very distant from the other values in a dataset.

* Being an outlier in itself does not make a data value invalid. Still outlier are the result of data errors.
* When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will still be valid.
* In any case, outlier should be identified and are usually worthy of further investigation.

Example:-

Let consider a example to demonstrate the concept of Estimates of location.

Consider there is a batch of students who have scored the following marks in one of the subject. {20, 10, 40, 30}. The school has introduced a system that if the average score of the batch is less than 35, then extra class has to be arranged for

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023-2024

the students, so that the results can be improved. After taking the mean, we can come to the conclusion whether special class is needed for this class or not.

Mean of {20, 10, 40, 30}

$$\text{Mean} = \frac{20+10+40+30}{4} = \frac{100}{4}$$

Mean = 25

The conclusion draw from the mean value is that they very much need the special class to be conducted immediately.

Just imagine a new student is added to the batch and his score is 90, Now the new dataset is {20, 10, 40, 30, 90}.

$$\text{Mean} = \frac{20+10+40+30+90}{5} = \frac{190}{5}$$

Mean = 38

Because of just one student the Mean >35, wherein the conclusion results that the batch does not require special class. Because of this the students will suffer. The conclusion is not accurate.

Semester: V Subject: Statistics for ATDS Academic Year: 2023-2024

The solution for this problem is median. The outlier in this case is 90.

Median:

Sort the dataset = {10, 20, 30, 40, 90}.

The middle value is 30

Median = 30

30 is near to the mean value. The conclusion is that batch needs special class. Because of median the conclusion is correct.

The disadvantage faced by the mean is somehow solved by the median.

If suppose another student is added to the batch with the score of 100. Let's check the median.

Median = {10, 20, 30, 40, 90, 100}

$= \frac{30+40}{2} = 35$

Median = 35.

Mode:-

Mode is the most frequent number - that is, the number that occurs the highest number of times.

Example:

Consider the dataset {10, 20, 10, 30, 40, 10, 90}

10 → 3, 40 → 1

20 → 1, 90 → 1

30 → 1

Mode = 3

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023 - 2024

Example – Problem:

Calculate the center of tendency for the below given data:

Formulas to calculate Mean, Median and Mode for Grouped data

$$\text{Mean} = \bar{x} = \frac{\Sigma f \cdot x}{\Sigma f}$$

$$\text{Median} = L + \left[\frac{N/2 - cf}{f}\right] h$$

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) h$$

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023-2024

Example:- Mean for Grouped data

Calculate the mean for the given following data.

Find average number of defective bulb in each boxes.

No. of defective Bulbs.	No. of boxes (f)	Class mark (x) (midvalue)	f.x (Total no of defective bulb in all boxes = 3 defective bulb)
0-2	3	(1)	3
2-4	4	3	12
4-6	5	5	25
6-8	3	7	21
8-10	1	9	9
Total	$\Sigma(f) = 16$		$\Sigma f.x = 70$

Class Interval.

$$\text{Mean} = \bar{X} = \frac{\Sigma f \cdot x}{\Sigma f}$$

$$\bar{X} = \frac{70}{16} \quad \frac{\text{(Total No of defective bulbs)}}{\text{Total No. of boxes.}}$$

$$= 4.38 \approx 4.$$

Average Number of defective bulb (Mean) } = 4.

Semester: V Subject: Statistics for IIDS Academic Year: 2023-2024

Example:- Median for Grouped data

Find the median for the below given data

Class Interval	Frequency (f)	Cumulative Frequency
(0-10)	5	5
10-20	10	15
20-30	12	27 → cf
(30-40)	15	42 (Median class)
(40-50)	18	60
Total	$\sum f_i = 60$	

$\frac{1}{2}$ · Total No of students / 2

$\frac{60}{2} = 30$

Here $N = 60$.

Find $N/2 = \frac{\text{Total No. of students}}{2} = \frac{60}{2} = 30$

Check when the cumulative frequency crosses 30. The cumulative frequency crosses 30 in class interval 30-40.

$$\text{Median} = L + \left[\frac{\frac{N}{2} - cf}{f}\right] h$$

L → lower limit → 30
N/2 → 30
cf → cf of previous class → 27.
f → frequency of median class → 15
h → class width. 10 units

$L = 30,\ N/2 = 30,\ cf = 27,\ f = 15,\ h = 10.$

$$\text{Median} = 30 + \left[\frac{30 - 27}{15}\right] \times 10.$$

$$= 30 + \left[\frac{3}{15}\right] \times 10 = 32$$

Median - 32

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023-2024

Example:- Mode for Grouped data

Consider the below given data and calculate the Mode.

Marks	Frequency	
0-10	2	
10-20	5	→ f_0
20-30	6	→ Modal class (f_1)
30-40	5	→ f_2
40-50	2	

$$Mode = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) h$$

f_1 → value of modal class → 6

f_0 → frequency of previous class − 5

f_2 → frequency of next class − 5

h → class width → 10.

L → Lower limit → 20

$$Mode = 20 + \left[\frac{6-5}{2\times6-5-5}\right] \times 10$$

$$= 20 + \left[\frac{1}{12-5-5}\right] \times 10 = 20 + \frac{1}{2} \times 10^{5}$$

Mode = 25

Semester: V Subject: Statistics for NIDS Academic Year: 2023-2024

Example: Calculate the median

Class interval	Frequency (f_i)	cf
0-20	5	5
20-40	8	13
40-60	15	28
60-80	16	44
80-100	6	50
Total	$\sum f_i = 50$	

(40-60: Median class)

$N = 50 = N/2 = 25$.

$$\text{Median} = l + \left[\frac{N/2 - cf}{f}\right] h.$$

$$= 40 + \left[\frac{25 - 13}{15}\right] \times 20$$

$$= 40 + \left[\frac{12^{4} \times 20^{4}}{15_{3}}\right]$$

$$= 40 + 16 = 56.$$

Median = 56

Semester: 4 Subject: Statistics for AIDS Academic Year: 2023-2024

Example2: (Mode)

Calculate the Mode for the following data:

Class Interval	Frequency f_i
3-4	1
4-5	7
5-6	28
6-7	78
7-8	84 → Modal class (f_1)
8-9	45
9-10	28
10-11	7
11-12	2

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)h$$

$f_1 \rightarrow 84$, $f_0 = 78$, $f_2 = 45$, $h = 1$, $L = 7$.

$$\text{Mode} = 7 + \left(\frac{84 - 78}{(2 \times 84) - 78 - 45}\right)$$

$$= 7 + \left(\frac{6}{168 - 78 - 45}\right) = 7 + \frac{6}{45}\frac{2}{15}$$

$$= \frac{105 + 2}{15} = \frac{107}{15}$$

$$\boxed{\text{Mode} = 7.133}$$