Basic Statistics

Measures of Central Tendency

Variable

A variable is a quantity which can vary from one value to another. The common examples of variables are barometer readings, temperature, rainfall records, wages, heights etc. Variables are of two kinds.

(i) Continuous variables: The quantities which can assume any numerical value within a certain range are called continuous variables. For example, if we consider the height H of a child at various ages, we observe that as the child grows from 110 cm to 160 cm (say), his height takes all possible values within this range. Thus, the height H of a child at various ages is an example of a continuous variable.

(ii) Discrete (or discontinuous) variables: The quantities which cannot assume all possible values are called discrete variables. For example, the number of children x in a family can take any of the values 0, 1, 2, 3 , ..., but cannot be 2.1, 2.46, 3.783. Thus, the number of children x is an example of a discrete variable.

Constant: A quantity is a constant if it can assume only one value.

Frequency Distribution

A frequency distribution is defined when the following two information are specified:

(i) The value which the variable takes, and
(ii) The number of times (i.e. frequencies) a value is taken by a variable.

Consider the marks obtained by 50 students in a statistics paper which are arranged according to their roll numbers, the maximum marks allotted to the paper being 100: 17, 71, 70, 14, 22, 00, 55, 59, 23, 87, 93, 21, 22, 50, 87, 54, 70, 52, 87, 74, 63, 87, 28, 04, 17, 49, 85, 81, 76, 52, 21, 86, 50, 87, 26, 87, 50, 60, 32, 40, 30, 90, 27, 89, 81, 21, 14, 30, 37, 22.

The data given above is in raw form, i.e. we cannot conclude anything. Such type of data is called Ungrouped data. When the data is arranged in ascending or descending order of magnitude, the data is said to be arranged in an array. Suppose these data is arranged in the intervals 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100.

This is done by the method called 'tally method.' In this method, we prepare a frequency table as follows:

Marks	Class	Number of Students (frequency)	Cumulative Frequency
0 — 10	II	2	2
10 — 20	II II	4	6
20 — 30	IIII IIII	10	16
30 — 40	II II	4	20
40 — 50	II	2	22
50 — 60	IIII III	8	30
60 — 70	II	2	32
70 — 80	IIII	5	37
80 — 90	IIII IIII I	11	48
90 — 100	II	2	50
Total		50	

In the above example, 'marks' is the variable x and the 'number of students' against the marks is called frequency f. Thus, the frequency of the class 50-60 is 8. The value of the variable which is mid-way between the upper and lower limits is called mid-

point (or mid-value) of the class. In the above example, the mid-values of the classes are respectively 5, 15, 25, ..., 95.

The cumulative frequency corresponding to a class is the total of all the frequencies upto and including that class. In the above example, the cumulative frequency of the class 0-10 is 2 (since there is no class above it). The cumulative frequencies are shown in the last column.

For the table, the following two forms of the frequency distribution may also be used.

Cumulative frequency 'more than'

Marks	Number of students
above 90	2
above 80	13
above 70	18
above 60	20
above 50	28
above 40	30
above 30	34
above 20	44
above 10	48
above 0	50

Cumulative frequency 'less than'

Marks	Number of students
under 20	6
under 30	16
under 40	20
under 50	22
under 60	30
under 70	32
under 80	37
under 90	48
under 100	50

Measures of Central Tendency

The following are the five measures of central tendency.

1. Arithmetic mean
2. Geometric mean
3. Harmonic mean
4. Median
5. Mode

Arithmetic Mean (AM)

If $X_1, X_2, X_3 \ldots X_n$ are the given n observations, then their AM, usually denoted by $\overline{X}$, is given by

$$\overline{X} = \frac{X_1 + X_2 + X_{3+\ldots\ldots+} X_n}{n} = \frac{\sum X}{n}$$

Weighted Arithmetic Mean

In the calculation of simple average, each item of the series is considered equally important but there may be cases where all items may not have equal importance and some of them may be comparatively more important than others. In such cases, proper weightage is to be given to various items — the weights attached to each item being proportional to the importance of the item in the distribution. Let $W_1, W_2, W_3, \ldots, W_n$ be the weights attached to variable values $X_1, X_2, X_3, \ldots\ldots, X_n$ respectively. Then the weighted arithmetic mean, usually denoted by $\overline{X}_W$ is given by $\overline{X}_W = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$

Sometimes, to calculate AM, we employ a well-defined short-cut method which is explained in the example below. In this method, we assume any value of the variable as 'assumed mean' and then find the actual mean using $\overline{X} = A + h\left(\frac{1}{N}\sum f_i u_i\right)$ where A: Assumed mean, h: Class size and $u_i = \frac{x_i - A}{h}$

Ex. 1 Two variables X and Y, assume the values $X_1 = 2$, $X_2 = -5$, $X_3 = 4$, $X_4 = -8$ and $Y_1 = -3$, $Y_2 = -8$, $Y_3 = 10$, $Y_4 = 6$, respectively. Calculate

(1) ΣX (2) ΣY (3) ΣXY
(4) ΣX^2 (5) ΣY^2
(6) $(\Sigma X)(\Sigma Y)$ (7) ΣXY^2
(8) $\Sigma(X + Y)(X - Y)$

Sol. (1) $\Sigma X = -7$
(2) $\Sigma Y = 5$
(3) $\Sigma XY = 26$
(4) $\Sigma X^2 = 109$
(5) $\Sigma Y^2 = 209$
(6) $(\Sigma X)(\Sigma Y) = -35$ (using (1) and (2))
(7) $\Sigma XY^2 = -190$
(8) $\Sigma(X + Y)(X - Y) = \Sigma(X^2 - Y^2) = 109 - 209 = -100$, (using (4) and (e))

Ex. 2 Find the mean wage from the data given below.

Wage (Amount in Rs.)	800	820	860	900	920	980	1000
Number of workers	7	14	19	25	20	10	5

Sol. Let the assumed mean be A = 900, h = 20

x_i	f_i	x_i - A	$u_i = (x_1 - A)/h$	$f_i u_i$
800	7	–100	–5	–35
820	14	–80	–4	–56
860	19	–40	–2	–38
900	25	0	0	0
920	20	20	1	20
980	10	80	4	40
1000	5	100	5	25
	$\sum f_i = 100$			$\sum f_i u_i = -44$

Here A = 900, h = 20.

$$\therefore \text{Mean} = \bar{X} = A + h\left(\frac{1}{N}\sum_{i=1}^{n} f_i\, u_i\right)$$

$$= 900 + 20\left(-\frac{44}{100}\right) = 891.2$$

Hence, mean wage = Rs. 891.2.
With the help of this method, calculation of multiplying two inconvenient numbers is avoided.

Median

The median is that value of the variable which divides the group in to two equal parts. One part comprises all the values greater than and the other part comprises all the values less than the median.

Calculation of Median

For individual observations

Step 1: Arrange the observations $x_1, x_2, ..., x_n$ in ascending or descending order of magnitude.
Step 2: Determine the total number of observations, say, n
Step 3: If n is odd, then median is the value of $\left(\frac{n+1}{2}\right)^{th}$ observation. If n is even, then median is the AM of the values of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

For discrete frequency distribution.
Step 1: Find the cumulative frequencies (c.f.)
Step 2: Find $\frac{N}{2}$, where $N = \sum_{i=1}^{n} f_i$
Step 3: See the cumulative frequency (c.f.) just greater than $\frac{N}{2}$ and determine the corresponding value of the variable, which is the median.

For grouped or continuous frequency distribution.
Step 1: Obtain the frequency distribution.
Step 2: Prepare the cumulative frequency column and obtain $N = \Sigma f_i$ Find $\frac{N}{2}$.

Step 3: See the cumulative frequency just greater than $\frac{N}{2}$ and determine the corresponding class.

This class is known as the median class.

Step 4: Use the formula: Median

$= L + \left(\frac{N/2 - F}{f}\right) \times h$, where, L = lower limit of the median class, f is frequency of the median class, h = width (size) of the median class, F = cumulative frequency of the class preceding the median class, $N = \Sigma f_i$.

Note: The mean deviation from the median for any distribution is minimum.

Ex. 3 Calculate the median for the following distribution:

Class:				
5-10	10-15	15-20	20-25	25-30
30-35	35-40	40-45		
Frequency:				
5	6	15	10	5
4	2	2		

Sol.

Class	Freqency	Cumulative frequency
5 – 10	5	5
10 – 15	6	11
15 – 20	15	26
20 – 25	10	36
25 – 30	5	41
30 – 35	4	45
35 – 40	2	47
40 – 45	2	49
	N = 49	

Here N = 49 $\Rightarrow \frac{N}{2} = \frac{49}{2}$ = 24.5. The cumulative frequency just greater than $\frac{N}{2}$ is 26 and the corresponding class is 15-20.

Thus 15-20 is the median class such that l = 15, f = 15, F = 11, h = 5.

$$\therefore \text{Median} = l + \frac{\frac{N}{2} - F}{f} \times h = 15 + \frac{24.5 - 11}{15} \times 5$$

$$= 15 + \frac{13.5}{15} \times 5 = 19.5$$

Mode

Mode is the value which occurs most frequently in a set of observations.

The mode may or may not exist, and even if it does exist, it may not be unique. A distribution having a unique mode is called 'unimodal' and one having more than one is called 'multimodal'.

Examples

The set 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 has mode 9.

The set 3, 5, 8, 10, 12, 15, 16 has no mode.

The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 has two modes, 4 and 7, and is called bimodal.

Ex. 4

Find the mean, median, and mode for the sets
(1) 3, 5, 2, 6, 5, 9, 5, 2, 8, 6 and
(2) 51.6, 48.7, 50.3, 49.5, 48.9

Sol. (1) Arranged in an array, the numbers are 2, 2, 3, 5, 5, 5, 6, 6, 8 and 9.
Mean = 5.1; Median = Arithmetic mean of two middle numbers = $\frac{1}{2}(5 + 5) = 5$;
Mode = Number most frequently occurring = 5.

(2) Arranged in an array, the numbers are 48.7, 48.9, 49.5, 50.3 and 51.6.
Mean = 49.8; Median = middle number = 49.5; Mode = non existent.

Calculation of Mode

In case of frequency distribution, mode is the value of the variable corresponding to the maximum frequency. In case of continuous frequency distribution, the class corresponding to the maximum frequency is called the modal class and the value of mode is obtained as, Mode $= l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)}$,

where

ℓ = Lower limit of modal class

h = Magnitude of the modal class, f_1= frequency of the modal class

f_0= Frequency of class preceding the modal class

f_2= Frequency of class succeeding the modal class

The above formula can be rephrased as Mode $= L_1 + \left(\frac{D_1}{D_1 + D_2}\right)c$,

where L_1 = Lower class boundary of the modal class (i.e. the class containing the mode)

D_1 = Excess of modal frequency over frequency of next lower class

D_2 = Excess of modal frequency over frequency of next higher class

c = Size of the model class interval.

Ex. 5

Compute the mode of the following distribution:

Class intervals:	0-7	7-14	14-21	21-28	28-35	35-42	42-49
Frequency :	19	25	36	72	51	43	28

Sol. Mode

$$(M_0) = l + \frac{f - f_{-1}}{2f - f_{-1} - f_1} \times i$$

$$= 21 + \frac{72 - 36}{144 - 36 - 51} \times 7$$

$$= 21 + \frac{252}{57} = 21 + 4.42 = 25.42.$$

Relationship Between Mean (M), Median (Md) and Mode(M_0)

In case of symmetrical distribution, Mean = Median = Mode

In case of a 'moderately' asymmetrical distribution, Mode = 3 Median – 2 Mean

Measures of Dispersion

The various measures of dispersion are as follows

1. Range
2. Mean deviation
3. Standard deviation
4. Quartile deviation
5. 10 – 90 percentile range

Range

It is the difference between two extreme observations of a distribution. Let X_{max} be the greatest observation and X_{min} the smallest observation of the variable. Then, Range = $X_{max} - X_{min}$.

Uses

1. Quality control
2. Shares
3. Weather forecast

Ex. 6

Find the range of the sets

(1) 12, 6, 7, 3, 15, 10, 18, 5

(2) 9, 3, 8, 8, 9, 8, 9, 18

Sol. After arranging the terms of (1), we get 3, 5, 6, 7, 10, 12, 15, 18. Therefore, range = 15.

After arranging the terms of (2), we get 3, 8, 8, 8, 9, 9, 9, 18. Therefore, range = 15.

Note that range is not a very good measure of dispersion because, like in this example, the range is same but the variation is greater in the first series than the second series.

To counter this, semi-interquartile range and 10 – 90 percentile range were designed to improve on the deficiencies of the range.

Mean Deviation (MD)

If $X_1, X_2, X_3 \ldots X_n$ are n given observations, then the mean deviation (MD) about A, is given by

$$MD = \frac{1}{n}\Sigma|X_i - A| = \frac{1}{n}\Sigma|d_i|, \text{ where } d_i = X_i - A.$$

Modulus removes the effects of negative deviations and therefore gives us the absolute value of the deviation.

In case of frequency distribution, mean deviation about A is given by

$$\frac{1}{n}\Sigma f_i|X_i - A| = \frac{1}{n}\Sigma f_i|d_i| = \overline{|X - \bar{X}|}.$$

Ex. 7

Determine the percentage of the students heights in the following table of mean deviation of the heights of 100 male students, that fall within the ranges

(1) $\bar{X} \pm MD$ (2) $\bar{X} \pm 2\,MD$

(3) $\bar{X} \pm 3\,MD$

Height (inches)	Number of students
60 - 62	5
63 - 65	18
66 - 68	42
69 - 71	27
72 - 74	8
	Total = 100

Sol.

Height (in)	Class Mark (X)	$\lvert X - \bar{X}\rvert = \lvert X - 76.45\rvert$	Frequency (f)	$f\lvert X - \bar{X}\rvert$
60 - 62	61	6.45	5	32.25
63 - 65	64	3.45	18	62.1
66 - 68	67	0.45	42	18.9
69 - 71	70	2.55	27	68.85
72 - 74	73	5.55	8	44.4
			$N = \Sigma f = 100$	$\Sigma f\lvert X - \bar{X}\rvert = 226.50$

The range from 65.19 in. to 69.71 inches is $\bar{X} \pm MD = 67.45 \pm 2.26$. This range includes all the individuals in the third class + $\frac{1}{3}$(65.5 – 65.19) of the students in the second class + $\frac{1}{3}$(69.71 – 68.5) of the students in the fourth class (since the class-interval size is 3 inches, the upper class boundary of the second class is 65.5 inches, and the lower class boundary of the fourth class is 68.5 inches). The number of students in the range $\bar{X} \pm MD$ is $42 + \frac{0.31}{3}(18) + \frac{1.21}{3}(27) \approx 55$, which is 55% of total.

(2) The range from 62.93 to 71.97 inches is $\bar{X} \pm 2\,MD = 67.45 \pm 3(2.26) = 67.45 \pm 6.78$. The number of students in the range $\bar{X} \pm 2\,MD$ is

$$18 - \left(\frac{62.93 - 62.5}{3}\right)18 + 42 + 27 + \left(\frac{71.97 - 71.5}{3}\right)8 \approx 86,$$ which is 86% of the total.

(3) The number of students in the range $\bar{X} \pm 3\,MD$ is

$$5 - \left(\frac{60.67 - 59.5}{3}\right)5 + 18 + 42 + 27 + \left(\frac{74.5 - 74.23}{3}\right)8 \approx 97,$$ which is 97% of the total.

Standard Deviation

If $X_1, X_2, \ldots X_N$ is the set of N observations, then its standard deviation is given by

$\sigma = \sqrt{\frac{1}{N}\Sigma (X_i - \overline{X})^2}$ where $\overline{X}$ is the AM alternative formula for standard deviations is

$$\sigma = \sqrt{\frac{\Sigma f_i X_i^2}{N} - \left(\frac{\Sigma f_i X_i}{N}\right)^2} = \sqrt{\overline{X^2} - \overline{X}^2}$$

The root mean square deviation about any point a is defined as $S = \sqrt{\frac{\sum_{i=1}^{N}(X_i - a)^2}{N}}$, where a is the average besides the arithmetic mean. Of all such deviations, the minimum is that for which $a = \overline{X}$. This is so because the sum of the squares of the deviation of a set of numbers X_i from any number a is minimum if and only if $a = \overline{X}$. This property provides an important reason for defining the standard deviation as above.

Note that root mean square deviation about origin is also called the quadratic mean and is given by $S = \sqrt{\frac{1}{N}\sum x_i^2}$.

Also note that σ is independent of origin but not of scale. This means that if each of the observations are increased by a constant k, σ does not change. But, if multiplied by k, σ changes to kσ. Therefore, if each observation is transformed to $\frac{ax_i + b}{c}$, σ transforms to $\frac{a\sigma}{\overline{X}}$.

Variance

It is the square of the standard deviation. Therefore, variance = $\sigma^2 = \frac{1}{N}\Sigma (X_i - \overline{X})^2$. Generally, s^2 represents sample variance and σ^2 represents population variance. Sample variance means variance of a sample drawn out from a population.