DESCRIPTIVE STATISTICS

Dept. of Ag. Stat.

- Numerical measures that are able to describe the features of the data – Averages.
- A single value around which all the values tend to cluster or spread - Central tendency.
- Any arithmetical measure which is intended to represent the center or central value of a set of observations - measure of central tendency.

MEASURES OF CENTRAL TENDENCY

- Mean
- Median
- Mode
- Geometric mean
- Harmonic mean

- Sum of the observation divided by total number of observation.
- Denote the values of the n observations by $x_1, x_2, x_3 \ldots, x_n$;
- $$\bar{x} = \frac{x_1 + x_2 \ldots\ldots\ldots + x_n}{n}$$

MS EXCEL = AVERAGE ()

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

ARITHMETIC MEAN

ARITHMETIC MEAN OF GROUPED DATA

Class midpoints	Frequency
x_1	f_1
x_2	f_2
. .	. .
.	.
x_k	f_k

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{Sum of \{(Mid point of class interval) \times (frequency of the class)\}}{Total\ frequency}$$

PROPERTIES OF A.M

- Sum of the deviations of a set of n observations $x_1, x_2, x_3 \ldots x_n$ from their A.M. is zero.

$$= \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

- A.M. of c$x_1, cx_2, cx_3 \ldots cx_n$ where c is a constant is CX'
- A.M. of $x_1 + c, x_2 + c, x_3$+c…x_n+c is x' +c
- Weighted A.M. = ∑ WiXi / ∑ Wi

MERITS OF A.M.

- Formula is well defined
- Easy to understand and easy to calculate
- Based upon all the observations
- Amenable to further algebraic treatments, provided the sample is randomly obtained.
- Of all averages, arithmetic mean is affected least by fluctuations of sampling

- Cannot be determined by inspection nor it can be located graphically

- Arithmetic mean cannot be obtained if a single observation is missing or lost

- Arithmetic mean is **affected very much by extreme values**

- In extremely asymmetrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location

DEMERITS OF A.M.

MEDIAN

- Median is the middle most item that divides the distribution into two equal parts when the items are arranged in ascending order.

MS EXCEL
= MEDIAN ()

Ungrouped data

- If the number of observations is odd then median is the middle value

- In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms.

MEDIAN FOR DISCRETE DISTRIBUTION

Obtained by considering the cumulative frequencies.

- ✓ Arrange the data in ascending or descending order of magnitude
- ✓ Find out cumulative frequencies
- ✓ Apply formula: Median = Size of $\frac{N+1}{2}$, where $N = \sum f$
- ✓ Look at the cumulative frequency column and find, that total which is either equal to $\frac{N+1}{2}$ or next higher to that and determine the value of the variable corresponding to it, which gives the value of median.

Continuous - data are given in class intervals

- Find $\frac{N+1}{2}$, where $N = \sum f$
- see the (less than) cumulative frequency just greater than $\frac{N}{2}$
- The class corresponding to the **cumulative frequency** just greater than $\frac{N}{2}$ is called the **median class**

CONTINUOUS FREQUENCY DISTRIBUTION:

$$Median = l + \left[\frac{\frac{N}{2} - m}{f}\right] c$$

l - lower limit of median class

f - frequency of the median class

m - cumulative frequency of the class preceding the median class

C - class interval

N - total frequency

- The number of runs scored by 11 players of a cricket team of a school are given. Find median

5	19	42	11	30	21	52	50	0	36	27

- arranged in ascending or descending order of magnitude. Let us arrange the values in ascending order:

0	5	11	19	21	27	30	36	42	50	52

Median = $\frac{N+1}{2}$ value = 6th value

Now the 6th value in the data is 27

Median = 27 runs

WHEN THE NUMBER OF OBSERVATIONS (N) IS ODD:

- Find the median of the following heights of plants in Cms:

6	10	4	3	9	11	18	13

Arrange the given items in ascending order

3	4	6	9	10	11	13	18

- In this data the number of items n = 8, which is even.

  Median = average of $(\frac{n}{2})$ th and $(\frac{n}{2}+1)$ th terms.

  average of 9 and 10   Median = 9.5 Cms.

WHEN THE NUMBER OF OBSERVATIONS (N) IS EVEN

Weight of ear head (in g) (X)	No. of ear heads (f)	LCF
40 -60	6	6
60 -80	28	34
80 - 100	35	69 (m)
100 – 120 (median class)	55 (f)	124
120 - 140	30	154
140 -160	15	169
160 - 180	12	181
180 - 200	9	190

$$Median = l + \left[\frac{\frac{N}{2} - m}{f}\right] c$$

(N / 2) = (190 / 2) = 95.

This value lies in between 69 and 124, and less than classes corresponding to these values are 100 and 120, respectively.
Hence the median class is 100 - 120
lower limit of this class is 100.
The cumulative frequency upto 100 is 69
and the frequency of the median class, 100 - 120 is 55.

$$\text{Median} = 100 + \left[\frac{(95 - 69)}{55} \times 20\right] = 100 + \left[\frac{26}{55} \times 20\right]$$

$$= 100 + 9.45 \text{ or } 109.45 \text{ g}$$

- Rigidly defined.
- Easily understood and is easy to calculate. In some cases it can be located merely by inspection.
- Not at all affected by extreme values.
- Can be calculated for distributions with open-end classes

MERITS OF MEDIAN

- In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms
- Not amenable to algebraic treatment
- As compared with mean, it is affected much by fluctuations of sampling.

DEMERITS OF MEDIAN

- Mode is the value which occurs most frequently in a set of observations
- mode is the value of the variable which is predominant in the series.
- In case of discrete frequency distribution mode is the value of x corresponding to maximum frequency

MS EXCEL
= MODE ()

MODE

$$\text{Mode} = l + \left[\frac{(f - f_1)}{2f - f_1 - f_2}\right]C$$

Where l is the lower limit of modal class

C is class interval of the modal class

f the frequency of the modal class

f_1 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively

MODE FOR CONTINUOUS FREQUENCY DISTRIBUTION

Marks	No. of students (f)
10-14	4
15-19	6
20-24	10
25-29	16f_1
30-34	21f
35-39	18f_2
40-44	9
45-49	5

$$\text{Mode} = l + \frac{(f - f_1)}{2f - f_1 - f_2} \text{xC}$$

modal class is 29.5-34.5

Here L = 29.5 , c = 5, $f = 21$, f_1 =16, f_2 = 18

$$\text{Mode} = 30 + \left[\frac{21-16}{2*21-16-18}\right]*5$$

$$= 30 + 3.13$$

$$= 33.63 \text{ m}$$

- Mode is readily comprehensible and easy to calculate.
- Mode is not at all affected by extreme values.
- Open-end classes also do not pose any problem in the location of mode

MERITS OF MODE

- Mode is ill defined. It is not always possible to find a clearly defined mode.
- In some cases, we may come across distributions with two modes. Such distributions are called bi-modal.
- If a distribution has more than two modes, it is said to be multimodal.
- Not based upon all the observations.
- Not capable of further mathematical treatment.
- As compared with mean, mode is affected to a greater extent by fluctuations of sampling.

DEMERITS OF MODE

Mean – Mode = 3 (Mean – Median)

- The positive root of the product of observations. Symbolically,

$$G = (x_1 x_2 x_3 \cdots\cdots\cdots\cdots x_n)^{1/n}$$

- It is also often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment.

GEOMETRIC MEAN

- If the "*n*" non-zero and positive variate -values $x_1, x_2, \ldots\ldots, x_n$ occur $f_1, f_2, \ldots\ldots, f_n$ times, respectively, then the geometric mean of the set of observations is defined by:

$$G = \left[x_1^{f_1}\ x_2^{f_2}\ \cdots\cdots\cdots x_n^{f_n}\right]^{\frac{1}{N}} = \left[\prod_{i=1}^{n} x_i^{f_i}\right]^{\frac{1}{N}}$$

Where $N = \sum_{i=1}^{n} f_i$

GEOMETRIC MEAN OF GROUP DATA

GEOMETRIC MEAN (REVISED EQN.)

Ungroup Data

$$G = \sqrt{(x_1 x_2 x_3 \cdots\cdots\cdots\cdots x_n)}$$

$$G = AntiLog\left(\frac{1}{N}\sum_{i=1}^{n} Log\ x_i\right)$$

Group Data

$$G = \sqrt{(x_1{}^{f_1} x_2{}^{f_2} x_3{}^{f_3} \cdots\cdots\cdots\cdots x_n)}$$

$$G = AntiLog\left(\frac{1}{N}\sum_{i=1}^{n} f_i\ Log\ x_i\right)$$

MS EXCEL
= GEOMEAN ()

- The harmonic mean is a very specific type of average.
- It's generally used when dealing with averages of units, like speed or other rates and ratios.

MS EXCEL
= HARMEAN ()

$$\overline{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

HARMONIC MEAN

Measures of dispersion

- Average alone is not sufficient to describe the characteristics of a distribution.
- Dispersion – Degree to which the numerical data tend to spread or scatter about a central value.
- The difference measures used to find the degree of scatter or spread – Measures of Dispersion.

- Range
- Interquartile range
- Quartile deviation
- Mean deviation
- Standard deviation
- Coefficient of variation

- Difference between the largest and smallest values in a set of data
- Useful for: daily temperature fluctuations or share price movement

Range = largest observation - smallest observation

RANGE

- The three parts which divide a series of frequency distribution into four equal parts.
- Q1 – 25% of observation below Q1 and 75% above Q1
- Q2 – 50% below Q2 and 50% above Q2
- Q3 – 75% below Q3 and 25% above Q2
- Position of Q1 = N/4 th observation
- Position of Q2 = N/2 th observation
- Position of Q3 = 3N/4 th observation

QUARTILES

INTERQUARTILE RANGE

- Measures the range of the middle 50% of the values only
- The difference between the upper and lower quartiles

Interquartile range = upper quartile - lower quartile

$$= Q_3 - Q_1$$

- The inter-quartile range is frequently reduced to the measure of semi-interquartile range, known as the quartile deviation (QD), by dividing it by 2. Thus

$$QD = \frac{Q_3 - Q_1}{2}$$

QUARTILE DEVIATION (QD)

- Measures the ‘average’ distance of each observation away from the mean of the data

- Gives an equal weight to each observation

- Generally more sensitive than the range or interquartile range, since a change in any value will affect it

MEAN DEVIATION (MD)

MEAN DEVIATION

The mean of the absolute deviations

- Mean deviation from A.M. (Mean deviation about mean)

$$Mean\ deviation = \frac{\sum \left| x - \bar{x} \right|}{n}$$

TO CALCULATE MEAN DEVIATION

1. Calculate mean of data	**Find** $\bar{x}$
2. Subtract mean from each observation **Record the differences**	**For each** x**, find** $x - \bar{x}$
3. Record absolute value of each residual	**Find** $\lvert x - \bar{x} \rvert$ **for each** x
4. Calculate the mean of the absolute values	$Mean\ deviation = \dfrac{\sum \lvert x - \bar{x} \rvert}{n}$ Add up absolute values and divide by n

STANDARD DEVIATION

The positive square root of the mean-square deviations of the observations from their arithmetic mean.

Also called root mean square deviation

Population

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Sample

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$$

MS EXCEL
= STDEV ()

$$SD = \sqrt{\text{variance}}$$

TO CALCULATE STANDARD DEVIATION

1. Calculate the mean	$\bar{x}$
2. Calculate the residual for each x	$x - \bar{x}$
3. Square the residuals	$(x - \bar{x})^2$
4. Calculate the sum of the squares	$\sum\left(x - \bar{x}\right)^2$
5. Divide the sum in Step 4 by (n-1)	$\dfrac{\sum\left(x - \bar{x}\right)^2}{n-1}$
6. Take the square root of quantity in Step 5	$\sqrt{\dfrac{\sum\left(x - \bar{x}\right)^2}{n-1}}$

STANDARD DEVIATION FOR GROUPED DATA

- SD is :

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{N}}$$

Where

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

- Simplified formula

$$s = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$$

- Mean – Mode = 3 (Mean – Median)
- Quartile Deviation (QD) = 2/3 of Standard Deviation (SD)
- Mean Deviation (MD) = 4/5 of Standard Deviation (SD)
- SD : MD: QD = 4 : 5: 6

IN A MODERATELY SKEWED DISTRIBUTION

COEFFICIENT OF VARIATION

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

where $\bar{X}$ = *the mean of the sample*

σ = *the standard deviation of the population*

Is a measure of relative variability used to

- measure changes that have occurred in a population over time
- compare variability of two populations that are expressed in different units of measurement
- expressed as a percentage rather than in terms of the units of the particular data

COEFFICIENT OF VARIATION