UNIT 5 MEASURES OF DISPERSION

Structure

5.0 OBJECTIVES

After going through this Unit, you will be able to:

- explain the concept of dispersion;
- compute numerical quantities that measure the dispersion of a set of data;
- explain Chebychev's inequality;
- compute the coefficient of variation; and
- find a measure for concentration of certain distribution of data.

5.1 INTRODUCTION

In Unit 4 we discussed various measures of central tendency, viz., arithmetic mean, median, mode, geometric mean and harmonic mean. However, in many situations these measures do not represent the distribution of data. For example, look into the following three sets of data:

Set A: 2, 5, 17, 17, 44.

Set B: 17, 17, 17, 17, 17.

Set C: 13, 14, 17, 17, 24.

In all the sets the numerical value of the mean, median and mode are the same, that is, 17. Still all three sets are so different! While in Set B all the observations are equal, in Set A they are so dispersed. Definitely we need another measure which will account for such dispersion of data.

In this Unit you will learn to deal with the concepts and techniques involved in reaching conclusions (making inferences) about a body of data in regard to their distribution over the range of variation of the variable.

5.2 CONCEPT OF DISPERSION

The word dispersion is used to denote the degree of heterogeneity in the data. It is an important characteristic indicating the extent to which observations vary amongst themselves. The dispersion of a given set of observations will be zero when all of them are equal (as in Set B given above). The wider the discrepancy from one observation to another, the larger would be the dispersion. (Thus dispersion in Set A should be larger than that in Set C.) A measure of dispersion is designed to state numerically the extent to which individual observations vary on the average.

There are quite a few measures of dispersion. We discuss them below.

5.2.1 Range

Of all measures of dispersions, range is the simplest. It is defined as *the difference between the largest and the smallest observations*. Thus for the data given at Set A the range is 44 – 2 = 42. Similarly, for Set B the range is 17 – 17 = 0 and for Set C it is 11. Now let us look into some grouped data. For Table 4.2 data (look back to the previous Unit), the range is Rs. 406.5 – Rs. 262.5 = Rs. 144. Notice that, for grouped data, largest and the smallest observations are not identifiable. Hence we take *the difference between two extreme boundaries of the classes*.

It is intuitive that, because of central tendency, if one selects a small sample, observations are more likely to be around its mode than away from it. Less likely or extreme values will be included in the sample when its size is large. This, in other words, implies that range will increase with increase in sample size. Also, it is known that in repeated sampling with same sample size, range varies considerably making it a less suitable measure for comparisons. However, range is a measure which is easy to understand and can be computed quickly.

5.2.2 Inter-quartile Range

Range as a measure of dispersion does not reflect a frequency distribution well, as it depends on the two extreme values. Even one very large or small observation, away from general pattern of other observations in the data set, makes the range very large. For example, in Set A, the range is found to be excessively large (44 – 2 = 42) because of the presence of very large one observation, that is 44. To avoid such extreme observations, particularly when there is a strong central tendency, inter-quartile range is useful as a measure of dispersion. It is defined as

Inter-quartile Range = $Q_3 - Q_1 = P_{75} - P_{25}$.

Inter-quartile range is the range of the middle most 50% of the observations. If the observations are compact around median, i.e., a strong mode close to the median exists 'inter-quartile range will be smaller than half of the range. If the data are

flat, having no central tendency, this measure will be large, and its value will be close to half of the range.

Let us look in to the discrete data given in Table 4.1 of the previous Unit. Here, $P_{75} = 4$ and $P_{25} = 3$. Hence, the inter-quartile range of household size is $4 - 3 = 1$. This shows that a strong central tendency exists in the distribution of household size since range was observed to be 7 (since $8 - 1 = 7$).

For Table 4.2 data, P_{25} of the average monthly expenditure on food was seen to be Rs. 325.50; P_{75} computed similarly works out to be Rs. 377.88 and inter-quartile range is Rs. 377.88 - Rs. 325.50 = Rs. 52.38. Compared to Rs. 52.38, the range was observed to be Rs. 146.00 or 2.79 times larger. This shows not so strong central tendency for average monthly household expenditure on food.

5.2.3 Mean Deviation

While range depends on the two extreme observations, inter-quartile range depends on the two extreme observations among the middle most 50 percent of the observations. Thus, one talks only about the percentage of observations between minimum, P_{25} and maximum, P_{75}. Thus both range and inter-quartile range do not depend upon all the observations in the sample. Hence while computing range or inter-quartile range we do not say anything about the distribution of observations within the group.

Among many possibilities to quantify spread or dispersion of observations, one possibility is to use the deviation of observations from some central value. Since mean is the most commonly used measure of central tendency, it is often taken as the central value with reference to which the deviations are computed. These deviations are then suitably combined to get a measure of dispersion.

Mean deviation treats every single observation with equal weight, in the form of arithmetic mean of deviations based on each observation.

For observations $X_1, X_2, \ldots X_n$, if one takes deviation as simple difference, then for the i^{th} observation the deviation is $X_i - \overline{X}$ where $\overline{X}$ is the mean. Mean of these deviations is

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}X_i - \frac{\overline{X}}{n}\sum_{i=1}^{n}1 = \overline{X} - \overline{X} = 0.$$

Since simple differences do not lead to any measure, absolute differences are used to define mean deviation.

$$\text{Mean Deviation} = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|, \text{ where}$$

the two vertical bars indicate that the sign of the difference within the two bars is to be taken as positive. For example, $|2 - 4| = 2$ (and not -2).

For frequency data, discrete or continuous type, the formula becomes

$$\text{Mean Deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i |X_i - \bar{X}|,$$

where $N = \sum_{i=1}^{n} f_i$ and X_i's are distinct observations and f_i is the frequency of X_i in the discrete case and X_i is the mid-point of ith class and f_i is its frequency for the continuous case. The need for such a measure is illustrated below.

Following summary values have been computed for two data sets.

	Data Set I	**Data Set II**
Number of observations	7	7
P_{25}	7	7
Median = P_{50}	12	12
P_{75}	17	17
Range	20	20
Inter-quartile range	10	10
Mean	12	12

Thus, based on the above measures only, and not looking at the data sets I and II, it would appear that two persons separately may have worked out on the same data set. However, the two data sets may have been as given below.

Data set I :	3	7	8	12	14	17	23
Data set II :	2	7	11	12	13	17	22

One may construct much more different looking data sets having identical values for the above type of measures. This comparison indicates that more measures are needed and mean deviation is one such. This is not to imply that the above measures and mean deviation together completely describe a data set.

For data set I

Mean deviation =

$$\frac{1}{7}\,(|3 -12| + |7 -12| + |8 -12| + |12 -12| + |14 -12| + |17 -12| + |23 -12|)$$

$$= \frac{9+5+4+0+2+5+11}{7} = \frac{36}{7} = 5.14$$

For data set II

Mean deviation =

$$\frac{1}{7}\,(|2 -12| + |7 -12| + |11 -12| + |12 -12| + |13 -12| + |17 -12| + |22 -12|)$$

$$= \frac{10+5+1+0+1+5+10}{7} = \frac{32}{7} = 4.57.$$

Thus, observations in data set I are more dispersed from mean than that of data set II.

Let us now compute mean deviation of household size and household average monthly food expenditure.

For household size data of Table 4.1, mean = $\overline{X} = 3.74$. Mean deviation is now computed as

$$\text{Mean deviation} = \frac{1}{N}\sum_{i=1}^{n} f_i |X_i - \overline{X}|$$

$$= \frac{1}{100}\left(3|1-3.74| + 16|2-3.74| + \cdots + 2|8-3.74|\right) = \frac{109.12}{100} = 1.0912.$$

For Table 4.2 distribution on average household expenditure on food, mean = $\overline{X}$ = Rs. 348.66.

The mean deviation =

$$\frac{1}{100}\left(2|274.5-348.66| + \cdots + 15|394.5-348.66|\right) = \frac{2510.88}{100} = 25.11.$$

So far we have considered mean deviation from mean. The mean deviation from median or from mode can also be defined in a similar way.

5.2.4 Variance and Standard Deviation

The most frequently used measures of dispersion are variance and standard deviation. Variance is so commonly used that it is also called dispersion.

Variance is a measure which suitably combines individual deviations from the mean, treating each observation with equal weight as in mean deviation. For variance, however, measure of individual deviation is taken as the *squared difference from the mean*. Since it is more manageable to use the squared difference rather than absolute difference, particularly while doing formal mathematics, use of variance has become more popular. Conventionally variance for a population is denoted by σ^2 (pronounced *sigma-squared*) and variance for a sample is denoted by σ^2. Variance is defined as the mean of the squared deviations of observations from their mean. Variance from raw data is computed by

$$\text{Variance} = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

For frequency data, discrete or continuous type, the formula becomes

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} f_i\left(X_i - \overline{X}\right)^2, \text{ where } N = \sum_{i=1}^{n} f_i$$

In the same scale of measurement, for example, observations with a variance of 2 are less dispersed than observations with variance more than 2. To talk about a distribution in terms of a measure of central tendency and a measure dispersion, it is a practical need to use both measures in the same unit. Mean and mean deviation are in the same unit. Since each deviation has been squared for

Based on variance, an equally or more popular measure of dispersion in the same unit as that of observations is *standard deviation*, abbreviated as s.d. Standard deviation is defined as the *positive square root of variance*, i.e., s.d. = σ. As it is the positive square root of variance, it cannot be negative.

Let us compute the s.d. for household size data of Table 4.1

$$\sigma^2 = \frac{1}{100}\left[3(1-3.74)^2 + 16(2-3.74)^2 + \cdots + 2(8-3.74)^2\right] = \frac{199.24}{100} = 1.9924 \text{ and}$$

$\sigma = 1.4115.$

Similarly for Table 4.2 distribution of average monthly household expenditure on food, variance in Rs.-square is given by

$$\sigma^2 = \frac{1}{100}\left[2(274.50-348.66)^2 + \cdots + 15(394.5-348.66)^2\right] = \frac{95725.437}{100} = 957.25,$$

and s.d. is

σ = Rs. 30.94.

For computational convenience, the formula for variance is written in alternative form as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

or

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}f_i\left(X_i - \overline{X}\right)^2 = \frac{1}{N}\sum_{i=1}^{n}f_iX_i^2 - \overline{X}^2$$

as the case may be. Thus, variance is viewed as

Variance = Mean of Squares – Square of the Mean

Using the above formulae, you may compute the variance for the data given in Tables 4.1 and 4.2 and verify the earlier results.

The computations of variance may be greatly simplified by changing X_i to

$u_i = \frac{X_i - A}{h}$, as was done in the computation of mean in Unit 4.

Note that, since

$$u_i - \overline{u} = \frac{X_i - A}{h} - \frac{\overline{X} - A}{h} = \frac{X_i - \overline{X}}{h}, \text{ we can write}$$

$$X_i - \overline{X} = h(u_i - \overline{u})$$

Hence,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\{h(u_i - \bar{u})\}^2 = h^2\sigma_u^2$$

where σ_x^2 is the variance of X_i and is the variance of u values.

Since the magnitude of u values is smaller, it is easier to compute variance of u values. Then the variance of X values can be easily computed by using the above formula.

Let us compute the variance by applying the above method for the data given in Table 4.2.

If we write $u_i = \frac{X_i - 346.5}{24}$, the u values are

- 3, - 2, - 1, 0, 1, 2 and the respective frequencies are 1, 14, 16, 28, 26, 15.

The mean of u values = $\frac{-3\times1 - 2\times14 - 1\times16 + 0\times28 + 1\times26 + 2\times15}{100} = 0.09$

The mean of squares of u values =

$$\frac{9\times1 + 4\times14 + 1\times16 + 0\times28 + 1\times26 + 4\times15}{100} = 1.67$$

Thus $\sigma_u^2 = 1.67 - (0.09)^2 = 1.6619$ and

$\sigma_X^2 = (24)^2.(1.6619) = 957.25.$

Even though change from X to u is for computational ease, it brings up an important issue. Notice that $\sigma_u^2 = 1.6619$ but $\sigma_X^2 = 957.25$, where u was obtained from X by a simple linear transformation, i.e., by change of origin and scale of X values. Typical such natural cases are pounds and kilograms for weight, gallons and litres for liquid volume, etc. Since 1 kg. = 2.2046 lbs., s.d. of 5 kg. when measured in kilograms is same as 11.023 lbs. when measured in pounds; or since 1 litre = 0.22 gallon, s.d. of 5 litres when measured in litres is same as s.d. of 1.1 gallons when measured in gallons. Thus, whereas variance and standard deviation are supposed to measure spread of observations, not much can be made out of these measures due to their dependence on the unit of measurement.

In this context, the single most useful result about the spread of observations based on mean and standard deviation, irrespective of unit of measurement, is due to Chebychev (discussed below in Section 5.3.1).

Check Your Progress 1

1) What is dispersion? What are the common measures of dispersion?

.....

.....

.....

.....

.....

.....

2) In a batch of 10 children the marks obtained by a dull boy are 25 marks below the average marks of other children. Show that the standard deviation of marks for all the children is at least 7.5. If this standard deviation is actually 12.0, find the standard deviation when the dull boy is left out.

.....

.....

.....

.....

.....

.....

3) The following data shows the daily profits (in Rs.) made by a shopkeeper on 15 successive days.

116, 87, 91, 81, 98, 102, 97, 100, 105, 101, 115, 98, 102, 98, 93

Determine the range, the mean deviation about mean and the standard deviation for the data.

.....

.....

.....

.....

.....

.....

4) Compute the arithmetic mean, standard deviation and the mean deviation of the following data.

Scores	4-5	6-7	8-9	10-11	12-13	14-15	Total
f	4	10	20	15	8	3	60

.....

.....

.....

.....

.....

.....

5) The mean and the s.d. of a sample of 100 observations were calculated as 40 and 5.1 respectively by a student who by mistake took one observation as 50 instead of 40. Calculate the correct s.d.

.....

.....

5.3 RELATIONSHIP BETWEEN DISPERSION AND STANDARD DEVIATION

You have earlier learnt that when all the values in a set of data are located near their mean, they exhibit a small amount of dispersion or variation and those set of data in which some values are located far from their mean have a large amount of dispersion. A useful rule that illustrates the relationship between dispersion and standard deviation is given by Chebychev's theorem.

5.3.1 Chebychev's Theorem

For any set of observations and positive constant k (> 1), the proportion of observations lying within k standard deviations of the mean is certain to be at least $1-\frac{1}{k^2}$.

Note that the theorem is not useful for any positive k less than or equal to 1, since $1-\frac{1}{k^2}$ is at the most equal to zero. For other values of k, the minimum proportion can be computed easily. For example, proportion of observations within 1.5 s.d. of the mean is certain to be at least $1-\frac{1}{1.5^2} = 0.556$ or 55.6%. The following figure indicates spread of data based on Chebychev's theorem. For the household size data of Table 4.1, $\bar{X} = 3.74$ and $s = 1.4115$. If we take $k = 2$, we can say that at least $\left[\left(1-\frac{1}{2^2}\right)\times 100\right] = 75\%$ of the households are certain to have their size between $3.74 \pm 2 \times 1.4115$, i.e., between 0.917 and 6.563.

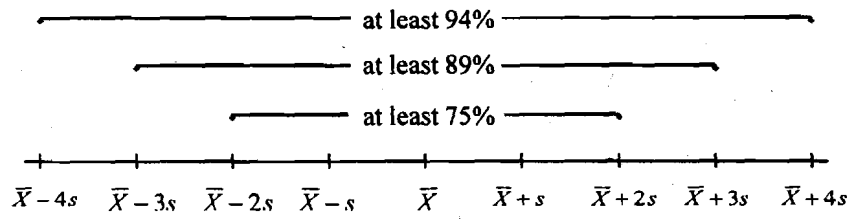

Fig. 5.1

For the Table 4.2 distribution of average monthly household expenditure on food $\bar{X}$ = Rs. 348.66 and s = Rs. 30.94, at least 55.6% (for $k = 1.5$) of households are certain to have monthly average food expenditure between Rs. 302.25 and Rs. 395.07. You can find the relevance of this theorem when we study normal distribution later in Unit 15.

5.3.2 Shape of Distribution

For methodological studies in many situations, a distribution is adequately described by measures of central tendency and dispersion. Yet other measures are also in use to describe distributions in practical situations, particularly for economic variables such as income, consumption, economic assets, etc., which are non-negative. Two such measures are *coefficient of variation* and *concentration ratio*. These measures will be viewed here essentially as measures of inequality in the distribution of economic variables.

5.3.3 Coefficient of Variation

Let us propose to compare economic status of households in two villages. The summary figures of monthly calorie intake of households are given below for the two villages.

	Villages	
	A	**B**
Number of Households (n)	817	561
Mean calorie intake ($\bar{X}$)	2417	2235
s. d. of calorie intake (σ)	418	232

The problem is to identify the village that has more inequality as far as calorie intake is concerned. Village A has higher mean calorie intake but has larger s.d. and larger number of households compared to village B. Village A may actually have more number of poorer households than in village B. Therefore, in village A, inequality between households may be more than that in village B. One index which measures the quantum of such disparity is called the coefficient of variation, abbreviated as c.v. It is defined as percentage standard deviation per unit of mean, i.e.,

$$\text{c.v.} = \frac{\sigma}{\bar{X}} \times 100$$

Since σ and $\bar{X}$ have the same unit of measurement, c.v. is unit free and is not affected by the choice of unit of measurement.

For village A, c.v. $= \frac{418}{2417} \times 100 = 17.29$ and for village B,

$$\text{c.v.} = \frac{232}{2235} \times 100 = 10.38.$$

Since the coefficient of variation in village A is greater than the coefficient of variation in village B, the inequalities are greater in village A compared to village B.

To compare the extent of inequalities, we compute

$\frac{17.29 - 10.38}{10.38} \times 100 = 66.57$ which implies that compared to village B, 66.57% more inequality exists in village A.

5.3.4 Concentration Ratio

Above was a comparison of inequality between two villages, without quantifying the level of inequality within each village. If a distribution has a long right tail, it

shows that a few have a large share. In other words, a majority of population has a very small share. Let us consider the distribution of income of a hypothetical economy. Suppose there are three classes of people in the economy — the upper class, the middle class and the lower class. Let 10%, 30% and 60% be the share of population in these three classes respectively. Suppose the lower class receives only 20% of the national income, the middle class 30% and the upper class the rest, i.e., the remaining 50%. We can now present the data in a percentage cumulative frequency distribution form. Thus, the lowest 60% of the population receives only 20% of the income, the lowest 90% receive 50% (= 20 +30) of the income and obviously, 100% of the population receive 100% of the income. If we take a graph paper where the percent cumulative frequency is plotted on the horizontal axis and percent cumulative total income is plotted on the vertical axis and we plot the point (0, 0), (60, 20), (90, 50) and (100, 100), then the curve joining these points is what we call the ***curve of concentration*** or ***Lorenz curve***. The straight line joining the points (0, 0) and (100, 100) give the line of *equal distribution* or the *equitable line*. The equitable line is that one which shows that the proportion of share is exactly the same as the proportion of population who are supposed to share. The area between the line of equal distribution and the curve of concentration, called the *area of concentration* is an indicator of the degree of concentration; the larger the area the greater is the concentration.

Coefficient of Inequality

Let us take the coordinates of the above points in per unit terms instead of percentage terms. Thus, the coordinates of the points, in the above example, can be written as (0, 0), (0.60, 0.20), (0.90, 0.50) and (1.00, 1.00). The coefficient of inequality of income distribution is then defined as the ratio of the area of concentration to total area of the triangle. Since the area of the triangle is 0.5 (since $\frac{1}{2} \times 1 \times 1 = 0.5$), the coefficient of inequality is equal to twice the area of concentration when coordinates of various points are taken in per unit rather than in percentage.

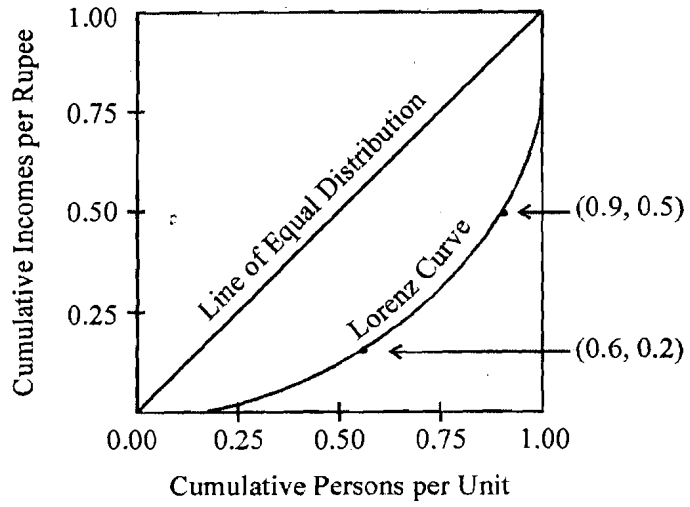

Fig. 5.2

Check Your Progress 2

1) The following figures give the crude birth rate per 1000 people in Switzerland from 1968 to 1980.

Crude birth rate (X): 17.1, 16.5, 15.8, 15.2, 14.3, 13.6, 12.9, 12.3, 11.7, 11.5, 11.3, 11.3, 11.6.

Calculate the Variance, Standard Deviation and Coefficient of Variation.

.....

.....

.....

.....

.....

.....

2) The following table gives the distribution of age of lady teachers of a school as revealed by records.

Age Group (years)	No. of lady teachers
15 - 19	3
20 - 24	13
25 - 29	21
30 - 34	15
35 - 39	5
40 - 44	4
45 - 49	2

Calculate coefficient of variation, and (ii) number of teachers between the age 26 and 33 years.

.....

.....

.....

.....

.....

.....

5.4 LET US SUM UP

In this Unit you learned about the measures of dispersion. The most important measures of dispersion you learned about in this unit are the variance, standard deviation and the concentration ratio. You have also learned to compute variance, standard deviation and coefficient of variation using both ungrouped and grouped data. The coefficient of variation is used to compare the dispersion of two distributions having either different means (even when their variables are measured in same units) or different units of measurement of their variables.

5.5 KEY WORDS

Coefficient of Variation: It is a relative measure of dispersion which is independent of the units of measurement. As opposed to this Standard Deviation is an absolute measure of dispersion.

Mean Deviation: It is the arithmetic mean of absolute deviations (i.e., the differences) from mean or median or mode.

Range: It is the difference between the largest and the smallest observations of a given set of data.

Standard Deviation: It is the positive square root of the variance.

Variance: It is the arithmetic mean of squares of deviations of observations from their arithmetic mean.

5.6 SOME USEFUL BOOKS

Elhance, D. N. and V. Elhance, 1988, *Fundamentals of Statistics,* Kitab Mahal, Allahabad.

Nagar, A. L. and R. K. Dass, 1983, *Basic Statistics,* Oxford University Press, Delhi

Mansfield, E., 1991, *Statistics for Business and Economics: Methods and Applications,* W.W. Norton and Co.

Yule, G U. and M. G Kendall, 1991, *An Introduction to the Theory of Statistics,* Universal Books, Delhi.

5.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Do it yourself.
2) 9.9
3) 35, 6.46, 8.85
4) 9.23, 2.49, 2.03
5) 5.0

Check Your Progress 2

1) 4.085, 2.021, 15.004%
2) 23.47%, 25 (rounded figure).