# 4. Measures of Dispersion Variation

**4.1 <u>Introduction</u>**:

Dispersion is a statistical term that describing the size of the range of values expected for a particular variable. The dispersion of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small when the values though different are close together.

The measures of central tendency, such as the mean, median, and mode are not reveal the whole picture of the distribution of the data set. Two data sets with the same mean may have completely different spreads. The variation among the values of observations for one data set may be much larger or smaller than the other data set. (Note that the words: dispersion, spread and variation have the same meaning). To explain the meaning of dispersion, let us consider the following example. Table **4.1** shows two data sets on the ages (in years) of all patients staying for each of two small hospitals.

**Table: 4.1**.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Hospital 1** | 47 | 38 | 35 | 40 | 36 | 45 | 39 |
| **Hospital 2** | - | 70 | 33 | 18 | 52 | 27 | - |

The mean age of patients in both hospitals is the same, 40 years. If we do not know the ages of individual patients at those two hospitals and we know only that the mean age of the patients at both hospitals is the same, we may conclude that the patients have the similar age distribution. As we can observe, the variation in the patients' ages of the two hospitals is very different. Where, the ages of the patients at the second hospital have much larger variation that the ages of the patients at the first hospital.

Thus, the mean, median, or mode only is not sufficient measure to reveal the shape of the distribution of the data set. Also, we need a measure that can provide some information about the variation among data values. The measures that help us to learn about the spread of the data set are called the *measures of dispersion*. The measures of central tendency and dispersion together give a better picture of the data set than the measures of central tendency alone. We will discuss three measures of dispersion: ***range***, ***variance***, and ***standard deviation***.

**1.2 <u>Measures of Dispersion</u>**:

**4.2.1 <u>Range</u>**: is the difference between the largest and smallest data values in the set.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

If we denote the range by $R$, the largest value by $x_L$ and the smallest value by $x_S$ , we can compute the range as follows:

$$R = x_L - x_S$$

**Example** 1: If have got two set of Population are given in a table

| Section A<br>Population A | section B<br>Population B |
|---|---|
| 10 | 25 |
| 20 | 30 |
| 30 | 35 |
| 40 | 35 |
| 50 | 40 |
| 60 | 45 |

Solution: The largest number in this data set A is 60 and the smallest number is 10. Therefor,

$$\boldsymbol{R_A = x_L - x_S}$$
$$= 60 - 10$$
$$= 50$$

And , The largest number in this data set B is 45 and the smallest number is 25. Therefor

$$\boldsymbol{R_B = x_L - x_S}$$
$$= 45 - 25$$
$$= 20$$

**Example 2**: Table 4.2 gives the total numbers of the insulin pens for four pharmacies in Kufa city. Find the range of the data set.

**Table 4.2**.

| Name of pharmacy | Pharmacy-1 | Pharmacy-2 | Pharmacy-3 | Pharmacy-4 |
|---|---|---|---|---|
| Total numbers (insulin pens) | 53 | 49 | 69 | 267 |

Solution: The largest number in this data set is 267 insulin pens, and the smallest number is 49 insulin pens. Therefor,

$$\text{Range} = \text{Largest value} - \text{Smallest value} \Longrightarrow R = x_L - x_S$$

$$\text{Range} = 267 - 49 = 218 \text{ insulin pens}$$

Thus, the total numbers of the insulin pens in these four pharmacies are spread over a range of 218 insulin pens.

The range has disadvantage of being influenced by outliers. In the above example, if the number of the insulin pens in pharmacy-4 of 267 is dropped, the range will decrease from 218 to 20. Consequently, <u>the range is not good enough measure of dispersion to use for the data set that contains outliers</u>. Another disadvantage of using the range as a measure of dispersion <u>is that its calculation is based on two values only; the largest and smallest. All other values in the data set are ignored when calculating the range</u>. Thus, the range is not very satisfactory measure of dispersion.

**4.2.2 <u>Variance</u>**: Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean.

Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set. Letting $S^2$ stands for the sample variance, the procedure may be written in notational form a follows:

OR, Can be calculated by the average of the squared difference from the Mean.

***In population*** $\boldsymbol{\delta^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$

***In sample*** $\boldsymbol{S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

$$\boldsymbol{S^2 = \frac{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}}$$

*Shortcut formula*

**Example**3 : From the example 1

Solution:

$$\boldsymbol{\delta^2}_{\mathbf{A}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathrm{x_i}-\mu)^2}{\mathbf{N}} = \frac{\mathbf{(10-35)^2+..+(60-35)^2}}{\mathbf{6}} = 291.6$$

$$\boldsymbol{\delta^2}_{\mathbf{B}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathrm{x_i}-\mu)^2}{\mathbf{N}} = \frac{\mathbf{(25-35)^2+..+(45-35)^2}}{\mathbf{6}} = 41.7$$

**Example 4**: Compute the variance of the ages of the patients in the hospital using the data in Table 4.3.

| Random Number | Sample Subject Number | Age |
|---|---|---|
| 137 | 1 | 43 |
| 114 | 2 | 66 |
| 155 | 3 | 61 |
| 183 | 4 | 64 |
| 185 | 5 | 65 |
| 28 | 6 | 38 |
| 85 | 7 | 59 |
| 181 | 8 | 57 |
| 18 | 9 | 57 |
| 164 | 10 | 50 |

**Table 4.3**.

Solution:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

- First we should calculate the value of the mean $\bar{x}$ using the equation

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{43 + 66 + 61 + \cdots + 50}{10} = 56$$

- Second we substitute the value of the mean in the variance equation as follows:

$$S^2 = \frac{(43-56)^2 + (66-56)^2 + (61-56)^2 + \cdots + (50-56)^2}{9} = \frac{810}{9} = 90$$

another solution by using Shortcut formula

| $x_i$ | $x_i^2$ |
|---|---|
| *43* | *1849* |
| *66* | *4356* |
| *61* | *4096* |
| *64* | *3721* |
| *65* | *4225* |
| *38* | *1444* |
| *59* | *3481* |
| *57* | *3249* |
| *57* | *3249* |
| *50* | *2500* |

$$\boldsymbol{S^2 = \frac{n\sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}}$$

$$= \frac{10(32170) - 313600}{90}$$

*=90*

**4.2.3 <u>Standard Deviation ($S_d$)</u>**:

The standard deviation is the most used measure of dispersion. A measure based on calculating the distance of each data value from the mean.

The standard deviation is obtained by taking the positive square root of the variance. The variance represents squared units ($S^2$) and, therefore, it is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units.

To obtained the measure of dispersion in original units, we simply take consider the square root of the variance ($\sqrt{s^2}$). The result is called the standard deviation and it is given by the following:

$$\textbf{\textit{In population}}\ \boldsymbol{\delta} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

$$\textbf{\textit{In sample}}\ S_d = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

## *The Variance* and Standard Deviation of grouped data

**the formula**

$$S^2 = \frac{n\sum_{i=1}^{n} x_m^2 f - (\sum_{i=1}^{n} x_m f)^2}{n(n-1)}$$

$$S_d = \sqrt{S^2}$$

***Where***

***$x_m$- class mark***

***F - frequency***

Example: Consider the following grouped data on the amount of time (in hours) that 80 college students devoted to leisure activities during a typical school week . Find the variance and standard deviation.

| *Time* | *Number of student* |
|---|---|
| *10-14* | *8* |
| *15-19* | *28* |
| *20-24* | *27* |
| *25-29* | *12* |
| *30-34* | *4* |
| *35-39* | *1* |

Solution

| *Time x* | *Number of student f* | $x_m$ | $x_m^2$ | $x_m^2 f$ | $x_m f$ |
|---|---|---|---|---|---|
| *10-14* | *8* | *12* | *144* | *1152* | *96* |
| *15-19* | *28* | *17* | *289* | *8092* | *476* |
| *20-24* | *27* | *22* | *484* | *13068* | *594* |
| *25-29* | *12* | *27* | *729* | *8748* | *324* |
| *30-34* | *4* | *32* | *1024* | *4096* | *128* |
| *35-39* | *1* | *37* | *1369* | *1369* | *37* |

$$\mathbf{S^2} = \frac{n\sum_{i=1}^{n} x_m^{\ 2} f - (\sum_{i=1}^{n} x_m f)^2}{n(n-1)} = \frac{80(36525) - 1655}{80*79} = 28.95$$

$$S_d = \sqrt{28.95} = 5.38$$

**4.3 <u>Coefficient of Variation (CV)</u>**:

A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows: (standard deviation) / (expected value). The coefficient of variation represents the ratio of the standard deviation to the mean. The CV is widely used in analytical chemistry. The formula of CV is given by:

$$CV = \frac{S_d}{\bar{x}}(100)\%$$

**Example 3**: Suppose two samples of human males yield the following results (as shown in table 4.4):

| Human male | Sample 1 | Sample 2 |
|---|---|---|
| Age | 25 years | 11 years |
| Mean weight | 65 kg | 30 kg |
| Standard deviation | 10 kgs | 10 kgs |

**Table 4.4**

Solution: A comparison of the standard deviations might lead one to conclude that the two samples possess equal variability. If we compute the coefficient of variation, however, we have for the 25-year-olds:

$$CV = \frac{S_d}{\bar{x}}(100)\%$$

$$CV = \frac{10}{65}(100)\% = 0.1538 * 100\% = 15.4\%$$

And for the 11-year-olds

$$CV = \frac{10}{30}(100)\% = 33.3\%$$

If we compare these results, we get quite a different impression. It is clear from this example that variation is much higher in the sample of 11-year-olds that in the sample of 25-year-olds.

**Example 4**: Find the coefficient of variation of 5, 10, 15, 20?

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{50}{4} = 12.5$$

Construct the following table:

| $\boldsymbol{x}$ | $\boldsymbol{x - \bar{x}}$ | $\boldsymbol{(x - \bar{x})^2}$ |
|---|---|---|
| 5 | $-7.5$ | 56.25 |
| 10 | $-2.5$ | 6.25 |
| 15 | 2.5 | 6.25 |
| 20 | 7.5 | 56.25 |

| | | |
|---|---|---|
| $\sum x = 20$ | | $\sum (x-\bar{x})^2 = 125$ |

$$CV = \frac{S_d}{\bar{x}}(100)\%$$

$$S_d = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{125}{3}} = 6.45$$

$$\Rightarrow CV = \frac{6.45}{12.5} * 100\% = 51.64\%$$

**H.W**

1. Some friends have just measured the heights of their dogs (in millimetres):

The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Find out the Mean, the Variance, and the Standard Deviation.

2. Find the coefficient of variation 100, 145, 170, 150?