4

Measures of Variation

4.1. Introduction

In the third chapter, we concentrated on a central value (measures of central tendency), which gives an idea of the whole mass that is a complete set of values. However the information so obtained is neither exhaustive nor comprehensive, as the mean does not lead us to know whether the observations are close to each other or far apart. Median is a positional average and has nothing to do with the variability of the observations in a data set. Mode is the largest occurring value independent of the other values in the set. This leads us to conclude that a measure of central tendency is not enough to have a clear idea about the data unless all observations are the same. Moreover two or more data sets may have the same mean and/or median but they may be quite different. So MCT alone do not provide enough information about the nature of the data. The table below displays the price of a certain commodity in four cities. Find the mean and median prices of the four cities and interpret it.

City A	30	30	30
City B	29	30	31
City C	15	30	45
City D	5	30	55

All the four data sets have mean 30 and median is also 30. But by inspection it is apparent that the four data sets differ remarkably from one another. So measures of central tendency alone do not provide enough information about the nature of the data. Thus, to have a clear picture of the data, one needs to have a measure of dispersion or variability among observations in the data set.

Variation or dispersion may be defined as the extent of scatteredness of value around the measures of central tendency. Thus, a measure of dispersion tells us the extent to which the values of a variable vary about the measure of central tendency.

4.2. Objectives of Measures of Variation

1. **To have an idea about the reliability of the measures of central tendency.** If the degree of scatterdness is large, an average is less reliable. If the value of the variation is small, it indicates that a central value is a good representative of all the values in the data set.

2. **To compare two or more sets of data with regard to their variability.** Two or more data sets can be compared by calculating the same measure of variation having the same units of measurement. A set with smaller value posses less variability or is more uniform (or more consistent).

3. **To provide information about the structure of the data.** A value of a measure of variation gives an idea about the spread of the observation. Further, one can summarise about the limits of the expansion of the values in the data set.

4. **To pave way to the use of other statistical measures.** Measures of variation especially variance and standard deviation lead to many statistical techniques like correlation, regression, analysis of variance,... etc.

4.3. Types of Measures of Variation

- **Absolute Measures of Variation**: A measure of variation is said to be an absolute form when it shows the actual amount of variation of an item from a measure of central tendency and are expressed in concrete units in which the data have been expressed.

- **Relative Measures of Variation**: A relative measure of variation is the quotient obtained by dividing the absolute measure by a quantity in respect to which absolute deviation has been computed. It is a pure number and used for making comparisons between different distributions.

Absolute Measures	**Relative Measures**
Range	Coefficient of Range
Quartile Deviation	Coefficient of Quartile Deviation
Mean Deviation	Coefficient of Mean Deviation
Variance	Coefficient of Variation
Standard Deviation	Standard Scores

Before giving the details of these measures of dispersion, it is worthwhile to point out that a measure of dispersion (variation) is to be judged on the basis of all those properties of good measures of central tendency. Hence, their repetition is superfluous.

4.3.1. Range and Relative Range

Range is the simplest and crudest/rough measure of dispersion. It is defined as the difference between the largest and the smallest values in the data.

- For raw data: $R = L - S$
- For grouped data: $R = UCL_{last} - LCL_{first}$

Coefficient of Range:

- For raw data: $CR = \frac{L-S}{L+S}$
- For grouped data: $CR = \frac{UCL_{last}-LCL_{first}}{UCL_{last}+LCL_{first}}$

Range hardly satisfies any property of good measure of dispersion as it is based on two extreme values only ignoring the others. It is not also liable to further algebraic treatment. The main advantage in using range is the simplicity of its computation.

4.3.2. Quartile Deviation and Coefficient of Quartile Deviation

Quartile deviation is sometimes known as Semi-Interquartile Range (SIR). The interquartile Range is $Q_3 - Q_1$. Thus,

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

The corresponding relative measure of variation, coefficient of quartile deviation is:

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

QD involves only the middle 50% of the observations by excluding the observations below the lower quartile and the observations above the upper quartile. Note also that it does not take into account all the individual values occurring between Q_1 and Q_2. It means that, no idea about the variation of even the 50% mid values is available from this measure. Anyhow it provides some idea if the values are uniformly distributed between Q_1 and Q_2.

4.3.3. Mean Deviation and Coefficient of Mean Deviation

The measures of variation discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation is a better measure than range and quartile deviation. Mean deviation is the arithmetic mean of the absolute values of the deviation from some measures of central tendency usually the mean and the median of a distribution. Hence we have mean deviation about the mean $MD(\bar{x})$ and mean deviation about the median $MD(\tilde{x})$.

Methods Obtaining Mean Deviation

- For raw data: $MD(\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n}$ and $MD(\tilde{x}) = \frac{\sum |x_i - \tilde{x}|}{n}$
- For grouped data: $MD(\bar{x}) = \frac{\sum f_i |m_i - \bar{x}|}{\sum f_i}$ and $MD(\tilde{x}) = \frac{\sum f_i |m_i - \tilde{x}|}{\sum f_i}$

MD is not much affected by extreme values. Its main drawback is that the algebraic negative signs of the deviations are ignored. MD is minimum when the deviation is taken from median. The coefficient of mean deviations are:

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}}$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}}$$

Examples

1. Consider a sample with data values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the range, coefficient of range, quartile deviation, coefficient of quartile deviation, mean deviation about mean, mean deviation about median, coefficient of mean deviation about mean and coefficient of mean deviation about median.

Solution:

Data: 15, 20, 25, 25, 27, 28, 30, 34

$$R = max - min = 34 - 15 = 19, CR = \frac{max - min}{max + min} = \frac{34 - 15}{34 + 15} = 0.388$$

To find QD and CQD, we have to calculate Q_1 and Q_3 first.

$$\begin{aligned} Q_1 &= \left(\frac{1 \times (8+1)}{4}\right)^{th} value = 2.25^{th} value \\ &= 2^{nd} value + 0.25(3^{rd} value - 2^{nd} value) \\ &= 20 + 0.25 \times (25 - 20) \\ &= 21.25 \end{aligned}$$

$$\begin{aligned} Q_3 &= \left(\frac{3 \times (8+1)}{4}\right)^{th} value = 6.75^{th} value \\ &= 6^{th} value + 0.75(7^{th} value - 6^{th} value) \\ &= 28 + 0.75 \times (30 - 28) \\ &= 29.5 \end{aligned}$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{29.5 - 21.25}{2} = 4.125$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{29.5 - 21.25}{29.5 + 21.25} = \frac{8.25}{50.75} = 0.163$$

Beside to this to compute $MD(\bar{x}), MD(\tilde{x}), CMD(\bar{x})$ and $CMD(\tilde{x})$ we should obtain $\bar{x}$ and $\tilde{x}$.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{204}{8} = 25.5;\ \tilde{x} = 26$$

$$\begin{aligned} MD(\bar{x}) &= \frac{\sum |x_i - \bar{x}|}{n} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_8 - \bar{x}|}{8} \\ &= \frac{|15 - 25.5| + |20 - 25.5| + ... + |34 - 25.5|}{8} \\ &= \frac{34}{8} = 4.25 \end{aligned}$$

$$\begin{aligned} MD(\tilde{x}) &= \frac{\sum |x_i - \tilde{x}|}{n} = \frac{|x_1 - \tilde{x}| + |x_2 - \tilde{x}| + ... + |x_8 - \tilde{x}|}{8} \\ &= \frac{|15 - 26| + |20 - 26| + ... + |34 - 26|}{8} \\ &= \frac{32}{8} = 4 \end{aligned}$$

Thus,

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} = \frac{4.25}{25.5} = 0.1667$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}} = \frac{4}{26} = 0.154$$

2. Calculate the R, QD and CQD for the following frequency distribution.

Class limits	10-14	15-19	20-24	25-29	30-34	35-38
Frequency	8	10	22	35	15	10

Solution:

Previously, we have obtained the following quantities for the students score data:

$$\bar{x} = 25.64, \tilde{x} = 26.1, Q_1 = 20, Q_3 = 31.07$$

Class	m_i	f_i	$\lvert m_i - \bar{x}\rvert$	$f_i\lvert m_i - \bar{x}\rvert$	$\lvert m_i - \tilde{x}\rvert$	$f_i\lvert m_i - \tilde{x}\rvert$
10.5-14.5	12.5	4	13.14	52.56	13.6	54.4
14.5-18.5	16.5	7	9.14	63.98	9.6	67.2
18.5-22.5	20.5	8	5.14	41.12	5.6	44.8
22.5-26.5	24.5	10	1.14	11.40	1.6	16.0
26.5-30.5	28.5	12	2.86	34.32	2.4	28.8
30.5-34.5	32.5	7	6.86	48.02	6.4	44.8
34.5-38.5	36.5	8	10.86	86.88	10.4	83.2
Total		56		338.28		339.2

$$R = UCL_{last} - LCL_{first} = 38 - 11 = 27$$

$$CR = \frac{UCL_{last} - LCL_{first}}{UCL_{last} + LCL_{first}} = \frac{38 - 11}{38 + 11} = \frac{27}{49} = 0.551$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{31.07 - 20}{2} = \frac{11.07}{2} = 5.54$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{31.07 - 20}{31.07 + 20} = \frac{11.07}{51.07} = 0.22$$

$$MD(\bar{x}) = \frac{\sum f_i|m_i - \bar{x}|}{\sum f_i} = \frac{338.28}{56} = 6.04$$

$$MD(\tilde{x}) = \frac{\sum f_i|m_i - \tilde{x}|}{\sum f_i} = \frac{339.2}{56} = 6.06$$

$$CMD(\bar{x}) = \frac{MD(\bar{x})}{\bar{x}} = \frac{6.04}{25.64} = 0.24$$

$$CMD(\tilde{x}) = \frac{MD(\tilde{x})}{\tilde{x}} = \frac{6.06}{26.1} = 0.23$$

4.3.4. Variance and Standard Deviation

Variance and standard deviation are the most superior and widely used measures of dispersions and both measure the average dispersion of the observations around the mean. The variance of a data set is the sum of the squares of the deviation of each observation taken from the mean divided by total number of observations in the data set. The positive square root of variance is called standard deviation.

For a *population* containing N elements, the population standard deviation is denoted by the Greek letter σ (sigma) and hence the population variance is denoted by σ^2.

- For raw data: $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$ and $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$
- For grouped data: $\sigma^2 = \frac{\sum f_i(m_i-\mu)^2}{N}$ and $\sigma = \sqrt{\frac{\sum f_i(m_i-\mu)^2}{N}}$

For a *sample* of n elements, the sample variance and standard deviation denoted by s^2 and s, respectively, are calculated as using the formulae:

- For raw data: $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$ and $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
- For grouped data: $s^2 = \frac{\sum f_i(m_i-\bar{x})^2}{\sum f_i-1}$ and $s = \sqrt{\frac{\sum f_i(m_i-\bar{x})^2}{\sum f_i-1}}$

Examples

1. Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.

   Solution:

   We are expected to compute the sample mean $\bar{x}$ first since the sample variance is a function the sample mean.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{10+20+12+17+16}{5} = \frac{75}{5} = 15$$

$$\begin{aligned} S^2 &= \frac{\sum(x_i-\bar{x})^2}{n-1} \\ &= \frac{(10-15)^2+(20-15)^2+(12-15)^2+(17-15)^2+(16-15)^2}{5-1} \\ &= \frac{64}{4} = 16 \end{aligned}$$

   Hence, $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{16} = 4$.

2. Calculate the variance and standard deviation for the following frequency distribution.

Class limits	10-14	15-19	20-24	25-29	30-34	35-38
Frequency	8	10	22	35	15	10

Solution:

The necessary calculation for calculating variance are as follows.

Class	m_i	f_i	$(m_i - \bar{x})$	$(m_i - \bar{x})^2$	$f_i(m_i - \tilde{x})^2$
10.5-14.5	12.5	4	-13.14	172.6596	690.6384
14.5-18.5	16.5	7	-9.14	83.5396	584.7772
18.5-22.5	20.5	8	-5.14	26.4196	211.3568
22.5-26.5	24.5	10	-1.14	1.2996	12.9960
26.5-30.5	28.5	12	2.86	8.1796	98.1552
30.5-34.5	32.5	7	6.86	47.0596	329.4172
34.5-38.5	36.5	8	10.86	117.9396	943.5168
Total		56		338.28	2870.8576

$$s^2 = \frac{\sum f_i(m_i - \bar{x})^2}{\sum f_i - 1} = \frac{2870.8576}{55} = 52.19$$

Therefore $s = \sqrt{52.19} = 7.22$.

- The main objection of mean deviation, removal of the negative signs, is removed by taking the square of the deviations from the mean. The first main demerit of variance is that its unit is the square of the unit of measurement of the variable values. For example, the sample variance of $2m$, $6m$ and $4m$ is $4m^2$. The interpretation is, on average each value differs from the mean by $4m^2$, which is completely wrong because one thing the unit of measurement of variance is not the same as that of the data set. The other disadvantage of variance is, the variation of the data is exaggerated because the deviation of the each value from the mean is squared. For the given example, the variation of the data is exaggerated from two to four since it is taking the square of the deviations. Variance also gives more weight the extreme values as compared to those which are near to the mean value.

- Standard deviation is considered to be the best measure of dispersion because the unit of measurement is the same as the data set and the exaggeration made by variance will be eliminated by taking the square root of it. In simple words, it explains the average amount of variation on either sides of the mean. If the standard deviation of the data is

small the values are concentrated near the mean and if it large the values are scattered away from the mean.

Properties of Variance and Standard Deviation

1. If a constant is added (subtracted) to (from) each and every observation, the standard deviation as well as the variance remains the same.

2. If each and every value is multiplied by a nonzero constant k, the standard deviation is multiplied by k and the variance is multiplied by k^2.

3. If there are k different groups having the same units of measurement with sample means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$, number of sample observations $n_1, n_2, ..., n_k$ and sample variances $s_1^2, s_2^2, ..., s_k^2$ respectively, then the variance of all the groups called the *pooled variance* denoted by s_p^2 is given by:

$$s_p^2 = \frac{(n_1 - 1)[s_1^2 + (\bar{x}_1 - \bar{x}_c)^2] + ... + (n_k - 1)[s_k^2 + (\bar{x}_k - \bar{x}_c)^2]}{n_1 + n_2 + ... + n_k - k}$$

$$s_p^2 = \frac{\sum(n_i - 1)[s_i^2 + (\bar{x}_i - \bar{x}_c)^2]}{\sum n_i - k}$$

If $\bar{x}_1 = \bar{x}_2 = ... = \bar{x}_k$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2}{n_1 + n_2 + ... + n_k - k} = \frac{\sum(n_i - 1)s_i^2}{\sum n_i - k}$$

Examples

1. The mean weight of 150 students is 60 kilograms. The mean weight of boys is 70 kg with a standard deviation of 10 kg. For the girls, the mean weight is 55 kg and the standard deviation 15 kg. Then,

   (a) Find the number of boys and girls.

   (b) Find the combined standard deviation.

2. A distribution consists of four parts characterized as follows. Find the mean and standard deviation of the distribution.

Part	No of items	Mean	S.D.
1	50	61	8
2	100	70	9
3	120	50	10
4	30	83	11

3. The arithmetic mean and standard deviation of a series of 20 items were computed as 20 and 5 respectively. While calculating these, an item 13 was misread as 30. Find the correct mean and standard deviation.

4. The following data are some of the particulars of the distribution of weights of boys and girls in a class.

	Boys	Girls
Number	100	50
Mean	60	45
Variance	9	4

   a) Find the mean and variance of the combined series.

   b) If one of the values is misread as 60 instead of 40 what is the correct standard deviation.

4.3.5. Coefficient of Variation

All absolute measures of dispersion have units. If two or more distributions differ in their units of measurement, their variability cannot be compared by any of the absolute measure of variation. Also, the size of the absolute measures of dispersion depends upon the size of the values. That is if the size of the values is larger, the value of the absolute measures will also be larger. Generally absolute measures of variation fail to be appropriate for comparing two or more groups if:

- ⊛ The groups have different units of measurement.
- ⊛ The size of the data between the groups is not the same.

Coefficient of variation is a relative measure of standard deviation. It is the ratio of the standard deviation to the mean and expressed as percent. Hence, it is a unitless measure of variation and also takes into account the size of the means of the distributions.

- For population: $cv = \frac{\sigma}{\mu} \times 100\%$
- For sample: $cv = \frac{s}{\bar{x}} \times 100\%$

■ The distribution having less cv is said to be less variable or more consistent or more uniform. For field experiments, cv , is generally reported. If it is small, it indicates more reliability of experimental findings.

Examples

1. Compare the variability of the following two sample data sets using standard deviation and coefficient of variation.

   A : 2 Meters, 4 Meters, 6 Meters

   B : 1000 Liters, 800 Liters, 900 Liters

2. The average IQ of statistics students is 110 with standard deviation 5 and the average IQ of mathematics students is 106 with standard deviation 4. Which class is less variable in terms of IQ?

4.4. Exercises

1. Find the range, quartile deviation, mean deviation about the mean, mean deviation about the median, mean deviation about the mode, variance, standard deviation and coefficient of variation for the following distribution.

Class	2-4	4-6	6-8	8-10
Frequency	2	5	4	7

2. Explain the rationale for using $n-1$ to compute the sample variance.

3. What is the purpose of coefficient variation?

4. Two persons participated in five shooting competition and were able to hit the target correctly out of fifteen shots as given below.

Competitor A	6	12	12	10	7
Competitor B	12	15	7	7	4

Which competitor is more uniform in shooting performance?