

Chapter Four

4. Measures of Dispersion (Variation) and Shape

4.1. Introduction

Variation (dispersion) is the scatter or spread of observations /values/ in a distribution. The average or central value is of little use unless the degree of variation, which occurs about it, is given. If the scatter about the measure of central tendency is very large, the average is not a typical value. Therefore it is necessary to develop a quantitative measure of the dispersion (or variation) of the values about the average.

- ❖ Measures of variation are statistical measures, which provide ways of measuring the extent to which the data are dispersed or spread out.

4.2. Objectives of Measuring Variation

Measures of variation are needed for the following basic objectives.

- ❖ To judge the reliability of a measure of central tendency
- ❖ To compare two or more sets of data with regard to their variability
- ❖ To control variability itself
- ❖ To make further statistical analysis

Properties of a good measure of dispersion

A good measure of dispersion should:

- be rigidly defined by a mathematical formula,
- be simple to understand and easy to calculate,
- be unique,
- calculated based on all observations in the series,
- not be affected by some extreme values existing in the series,
- have sampling stability property, and
- be capable of further algebraic treatment as well as further statistical analysis.

4.3. Types of Measures of Dispersion

4.3.1. Absolute Measures of Dispersion

The measures of dispersion which are expressed in terms of the original unit of a series are termed as absolute measures. Such measures are not

suitable for comparing the variability of two distributions which are expressed in different units of measurement and different average size.

4.3.2. Relative Measures of Variation

Relative measures of dispersions are a ratio or percentage of a measure of absolute dispersion to an appropriate measure of central tendency and are thus pure numbers independent of the units of measurement.

- For comparing the variability of two distributions (even if they are measured in the same unit), we compute the relative measure of dispersion instead of absolute measures of dispersion.

I. The Range and Relative Range

Range (R) is defined as the difference between the largest and the smallest observation in a given set of data. That is, $R = x_{\max} - x_{\min}$ where $x_{\max}$ and $x_{\min}$ are the largest and the smallest observations in the series respectively.

In case grouped data, range is found by taking the difference between the class mark of the last class and that of the first class. That is, $R = M_{\text{last}} - M_{\text{first}}$ where M_{last} and M_{first} are the class marks of the last class and that of the first class respectively.

A relative range (RR), also known as *coefficient of range*, is given by

$$RR = \frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{R}{x_{\max} + x_{\min}} \dots \dots \dots \text{for ungrouped data}$$

$$RR = \frac{M_{\text{last}} - M_{\text{first}}}{M_{\text{last}} + M_{\text{first}}} = \frac{R}{M_{\text{last}} + M_{\text{first}}} \dots \dots \dots \text{for grouped data}$$

Properties of Range and Relative Range

- Range and relative range are easy to calculate and simple to understand.
- Both cannot be computed for grouped data with open ended classes.
- They do not tell us anything about the distribution of values in the series.

Example 1: Find the range and relative range for the monthly salary of ten workers in a certain paint factory given below.

462 480 534 624 498 552 606 588 516 570

Solution:

$$x_{\max} = 624 \text{ birr} \quad x_{\min} = 462 \text{ birr}$$

$$R = x_{\max} - x_{\min} = 624 \text{ birr} - 462 \text{ birr} = 162 \text{ birr}$$

$$RR = \frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{624 \text{ birr} - 462 \text{ birr}}{624 \text{ birr} + 462 \text{ birr}} = \frac{162 \text{ birr}}{1086 \text{ birr}} = 0.149$$

Example 2: Find the values of the range and relative range for the following frequency distribution: which shows the distribution of the maximum loads supported by a certain number of cables.

Maximum load (in kilo-Newton)	Number of cables
93 - 97	2
98 - 102	5
103 - 107	12
108 - 112	17
113 - 117	14
118 - 122	6
123 - 127	3
128 - 132	1

Solution:

$$M_{\text{first}} = 95 \text{ kN} \quad M_{\text{last}} = 130 \text{ kN}$$

$$R = M_{\text{last}} - M_{\text{first}} = 130 \text{ kN} - 95 \text{ kN} = 35 \text{ kN}$$

$$RR = \frac{M_{\text{last}} - M_{\text{first}}}{M_{\text{last}} + M_{\text{first}}} = \frac{130 \text{ kN} - 95 \text{ kN}}{130 \text{ kN} + 95 \text{ kN}} = \frac{35 \text{ kN}}{225 \text{ kN}} = 0.156$$

II. The Mean Deviation and Coefficient of Mean Deviation

The *mean deviation (MD)* measures the average deviation of a set of observations about their central value, generally the mean or the median, ignoring the plus/minus sign of the deviations.

The mean deviation of a sample of n observations $x_1, x_2, \dots, x_n$ is given as

$$MD = \frac{\sum |x_i - A|}{n} \quad \text{Where } A \text{ is a central measure (the mean or the median)}$$

In case of grouped data, the formula for MD becomes

$$MD = \frac{\sum f_i |m_i - A|}{n} \quad \text{Where } m_i \text{ is the class mark of the } i^{th} \text{ class, } f_i \text{ is the frequency of the } i^{th} \text{ class and } n = \sum f_i.$$

- The mean deviation about the arithmetic mean is, therefore, given by

$$MD = \frac{\sum |x_i - \bar{x}|}{n} \dots \text{for ungrouped data}$$

$$MD = \frac{\sum f_i |m_i - \bar{x}|}{n} \dots \text{for grouped frequency distribution; where } m_i \text{ is the class mark of the } i^{th} \text{ class, } f_i \text{ is the frequency of the } i^{th} \text{ class and } n = \sum f_i$$

- The mean deviation about the median is also given by

$$MD = \frac{\sum |x_i - \tilde{x}|}{n} \dots \text{for ungrouped data}$$

$$MD = \frac{\sum f_i |m_i - \tilde{x}|}{n} \dots \text{for grouped frequency distribution; where } m_i \text{ is the class mark of the } i^{th} \text{ class, } f_i \text{ is the frequency of the } i^{th} \text{ class and } n = \sum f_i.$$

Examples:

1. The following are the number of visit made by ten mothers to the local doctor's surgery. 8, 6, 5, 5, 7, 4, 5, 9, 7, 4
Find mean deviation about mean, median and mode.

Solutions:

First calculate the three averages

$$\bar{X}=6, \tilde{X}=5.5, \hat{X}=5$$

Then take the deviations of each observation from these averages.

X_i	4	4	5	5	5	6	7	7	8	9	total
$ X_i-6 $	2	2	1	1	1	0	1	1	2	3	14
$ X_i-5.5 $	1.5	1.5	0.5	0.5	0.5	0.5	1.5	1.5	2.5	3.5	14
$ X_i-5 $	1	1	0	0	0	1	2	2	3	4	14

$$\Rightarrow M.D(\bar{X}) = \frac{\sum_{i=1}^{10} |X_i - 6|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^{10} |X_i - 5.5|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{10} |X_i - 5|}{10} = \frac{14}{10} = 1.4$$

2. Find mean deviation about mean, median and mode for the following distributions.(exercise)

Class	Frequency
40-	7

44	
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Remark: Mean deviation is always minimum about the median.

The coefficient of mean deviation (CMD) is the ratio of the mean deviation of the observations to their appropriate measure of central tendency: the arithmetic mean or the median.

In general, $CMD = \frac{MD}{A}$ where A is a measure of central tendency: the arithmetic mean or the median.

That is, CMD about the arithmetic mean is given by $CMD = \frac{MD}{\bar{x}}$ where MD is the mean deviation calculated about the arithmetic mean. On the other hand CMD about the median is given by $CMD = \frac{MD}{\tilde{x}}$ in which case MD is calculated about the median of the observations.

Properties of Mean Deviation and coefficient of mean deviation

- It is easy to understand and compute
- It is based on all observations
- It is not affected very much by the values of extreme value(s).

- It is not capable of further mathematical treatments and it is not a very accurate measure of dispersion.

Example: calculate the C.M.D about the mean, median and mode for the data in example 1 above.

Solutions:

$$C.M.D = \frac{M.D}{\text{Average about which deviations are taken}}$$

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}} = \frac{1.4}{6} = 0.233$$

$$C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}} = \frac{1.4}{5.5} = 0.255$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}} = \frac{1.4}{5} = 0.28$$

III. The Variance, the Standard Deviation and Coefficient of Variation

The Variance

Variance is the arithmetic mean of the square of the deviation of observations from their arithmetic mean.

- Population Variance (σ^2)

For ungrouped data

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \dots = \frac{1}{N} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right) \text{ Where } \mu \text{ is the population arithmetic mean and } N \text{ is the total number of observations in the population.}$$

For grouped data

$$\sigma^2 = \frac{\sum f_i (m_i - \mu)^2}{N} = \frac{1}{N} \left(\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{N} \right) \text{ Where } \mu \text{ is the population arithmetic mean, } m_i \text{ is the class mark of the } i^{th} \text{ class, } f_i \text{ is the frequency of the } i^{th} \text{ class and } N = \sum f_i.$$

- Sample Variance (S^2)

For ungrouped data

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \dots = \frac{1}{n-1} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)$$

Where $\bar{x}$ is the sample arithmetic mean and n is the total number of observations in the sample.

For grouped data

$$S^2 = \frac{\sum f_i (m_i - \bar{x})^2}{n-1} = \dots = \frac{1}{n-1} \left(\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{n} \right)$$

Where $\bar{x}$ is the sample arithmetic mean, m_i is the class mark of the i^{th} class, f_i is the frequency of the i^{th} class and $n = \sum f_i$.

The Standard Deviation

Standard deviation is the positive square root of the variance.

▪ Population Standard Deviation (σ)

$\sigma = \sqrt{\sigma^2}$ where σ^2 is the population variance.

▪ Sample Standard Deviation (S)

$S = \sqrt{S^2}$ where S^2 is the sample standard deviation.

Example 1: compute the variance for the following data

value	3	6	9	12	15	Total
frequency	1	4	10	3	2	20
$f_i x_i$	3	24	90	36	30	183
$x_i - \mu$	-6.15	-3.15	-0.15	2.85	5.85	
$(x_i - \mu)^2$	37.8225	9.9225	0.0225	8.1225	34.2225	
$f_i (x_i - \mu)^2$	37.8225	39.69	0.225	24.3675	68.445	170.55

$$\bar{x} = \frac{\sum f_i X_i}{n} = \frac{183}{20} = 9.15, \text{ where } n = \sum_{i=1}^5 f_i = 20$$

And
$$S^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{170.55}{19} = 8.976$$

Coefficient of Variation

The standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation (CV)*.

Coefficient of variation is used in such problems where we want to compare the variability of two or more different series. Coefficient of variation is the ratio of the standard deviation to the arithmetic mean, usually expressed in percent.

$$CV = \frac{S}{\bar{x}} \times 100$$

. Where S is the standard deviation of the observations.

A distribution having less coefficient of variation is said to be less variable or more consistent or more uniform or more homogeneous.

Example: Last semester, the students of Biology and Chemistry Departments took *Stat 273* course. At the end of the semester, the following information was recorded.

Department	Biolo gy	Chemis try
Mean score	79	64
Standard deviation	23	11

Compare the relative dispersions of the two departments' scores using the appropriate way.

Solution:

Biology Department	Chemistry Department
--------------------	----------------------

$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{23}{79} \times 100 = 29.11\%$	$CV = \frac{S}{\bar{x}} \times 100$ $= \frac{11}{64} \times 100 = 17.19\%$
--	--

Interpretation: Since the CV of Biology Department students is greater than that of Chemistry Department students, we can say that there is more dispersion relative to the mean in the distribution of Biology students' scores compared with that of Chemistry students.

Example: 1. The following table illustrates the frequency distribution of masses of 100 male students in Gander University.

Mass (kg)	60-62	63-65	66-68	69-71	72-74
No. of students	5	18	42	27	8

Find: a) the variance b) the standard deviation c) the coefficient of variation

d) Calculate mean deviation?

Solution:

Mass (kg)	60-62	63-65	66-68	69-71	72-74	total
No. of students(f_i)	5	18	42	27	8	100
class mark(m_i)	61	64	67	70	73	
$f_i m_i$	305	1152	2814	1890	584	6745
$f_i m_i^2$	18605	73728	188538	132300	42632	455803
$ m_i - \bar{x} $	6.45	3.45	0.45	2.55	5.55	
$f_i m_i - \bar{x} $	32.25	62.1	18.9	68.85	44.4	226.5

--	--	--	--	--	--	--

$$\sum_{i=1}^5 f_i m_i = 6745, \quad \sum_{i=1}^5 m_i^2 = 455803, \quad n = \sum_{i=1}^5 f_i = 100$$

$$\text{and } \bar{x} = \frac{\sum_{i=1}^5 f_i m_i}{n} = \frac{6745}{100} = 67.45$$

$$\text{a) } S^2 = \frac{1}{n-1} \left(\sum_{i=1}^5 f_i m_i^2 - \frac{\left(\sum_{i=1}^5 f_i m_i \right)^2}{n} \right) = \frac{1}{99} \left(455803 - \frac{(6745)^2}{100} \right) = 8.61$$

$$\text{b) } S = \sqrt{S^2} = \sqrt{8.61} = 2.93$$

$$\text{c) } CV = \frac{S}{\bar{x}} * 100 = \frac{2.93}{67.45} * 100 = 4.344$$

$$\text{d) } MD = \frac{\sum f_i |m_i - \bar{x}|}{n} = \frac{226.5}{100} = 2.265$$

2. A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1 25 24 23 26 17

City2 22 21 24 22 20

City3 32 27 35 24 28

Which city have the most consistent temperature, based on these data?
(Exercise)