Measures of Dispersion

Safayet Hossain
Lecturer
Department of Statistics
Comilla University

Chapter Summery

- Measures of Dispersion
- Definition of Dispersion
- Objectives of Measures of Dispersion
- Characteristics of a good measure of dispersion
- Types of measures of dispersion
- Relation between Measures of Dispersion
- Problem Solving

INTRODUCTION

- The Measures of central tendency gives us a birds eye view of the entire data they are called averages of the first order,
- it serve to locate the centre of the distribution but they do not reveal how the items are spread out on either side of the central value.
- The measure of the scattering of items in a distribution about the average is called dispersion.

Introduction

- So far we have looked at ways of summarising data by showing some sort of average (central tendency).
- But it is often useful to show how much these figures differ from the average.
- This measure is called **dispersion**.

WHAT IS MEASURE OF DISPERSION?

It has two terms:

- Measure: It means a "specific method of estimation"
- Dispersion: (also known as scatter, spread, variation) the term means " difference or deviation of a certain values from their central value"

DISPERSION

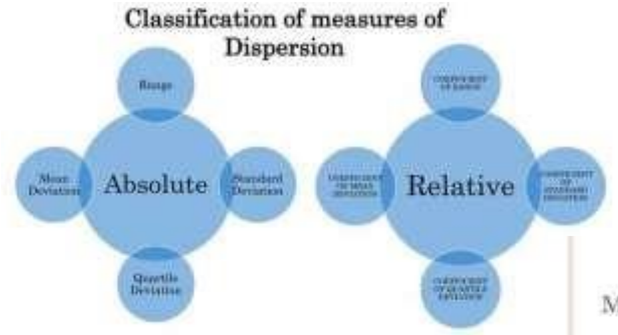
- Dispersion refers to the variations of the items among themselves / around an average.
- Greater the variation amongst different items of a series, the more will be the dispersion.
- As per Bowley, *"Dispersion is a measure of the variation of the items".*

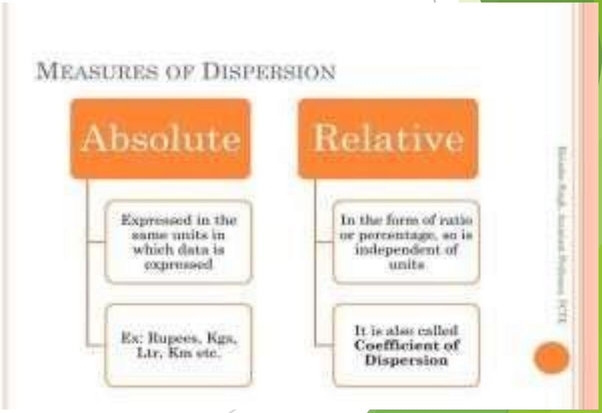
Birinder Singh, Assistant Professor, PCTE

Purpose of Measuring Dispersion

- A measure of dispersion appears to serve two purposes.
- First, it is one of the most important quantities used to characterize a frequency distribution.
- Second, it affords a basis of comparison between two or more frequency distributions.
- The study of dispersion bears its importance from the fact that various distributions may have exactly the same averages, but substantial differences in their variability.

- *Measures of dispersion* are descriptive statistics that describe how similar a set of scores are to each other
  - The more similar the scores are to each other, the lower the measure of dispersion will be
  - The less similar the scores are to each other, the higher the measure of dispersion will be
  - In general, the more spread out a distribution is, the larger the measure of dispersion will be


Classification of measures of Dispersion
Range
Absolute
Mean Deviation
Standard Deviation
Quartile Deviation
Relative


MEASURES OF DISPERSION
Absolute
Relative
Expressed in the same units in which data is expressed
Ex: Rupees, Kgs, Ltr, Km etc.
In the form of ratio or percentage, so is independent of units
It is also called Coefficient of Dispersion

RANGE (R)

- It is the simplest measures of dispersion
- It is defined as the difference between the largest and smallest values in the series

$$R = L - S$$

R = Range, L = Largest Value, S = Smallest Value

- Coefficient of Range = $\frac{L-S}{L+S}$

Range For Ungrouped data:

Let us consider a set of observations $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots\ldots, x_n$ and X_H is maximum and X_L is minimum.

$$\text{Then Range} = X_H - X_L.$$

Example:

Find out the range of the set of observations, -7, -2, -4, 0, 8.

Solution:

Here, maximum value, $X_H = 8$ and minimum value, $X_L = -7$

Range $= X_H - X_L = 8 - (-7) = 8 + 7 = 15$

Range For Grouped data:

In this case, the range is the difference between the upper boundary of the highest class and the lower boundary of the lowest class.

$$\text{Then Range} = X_U - X_L$$

Where,

X_U= The upper boundary of the highest class.

X_L= The lowest boundary of the highest class.

Example: determine the range from the following frequency distribution.

Salary (TK.)	1700-1800	1800-1900	1900-2000	2000-2100	2100-2200
No. of workers	420	460	500	300	200

Solution:

From the given frequency distribution,

We have,

The upper boundary of the highest class, $X_U = TK. 2200$

And the lowest boundary of the highest class, $X_L = TK. 1700$

Then Range $= X_U - X_L = TK. 2200 - TK. 1700 = TK. 500$

Example:

Compute the coefficient of range from the following data

10, 12, 8, 5, 6, 20

Solution:

$$X_H = Highest\ value = 20,$$

$$X_L = Lowest\ value = 5$$

Coefficient of Range, CR $= \frac{Range}{X_H + X_L} \times 100$;

$$= \frac{20-5}{20+5} \times 100$$

$$= 60\%$$

When To Use the Range

- The range is used when
  - you have ordinal data or
  - you are presenting your results to people with little or no knowledge of statistics
- The range is rarely used in scientific work as it is fairly insensitive
  - It depends on only two scores in the set of data, X_L and X_S
  - Two very different sets of data can have the same range:

    1 1 1 1 9 vs 1 3 5 7 9

Range

MERITS	DEMERITS
Simple to understand	Can't be calculated in open ended distributions
Easy to calculate	Not based on all the observations
Widely used in statistical quality control	Affected by sampling fluctuations
	Affected by extreme values

Birinder Singh, Assistant Professor, PCTE

Practice Problems – Range

Q1: Find the range & Coefficient of Range for the following data: 20, 35, 25, 30, 15

Ans: 20, 0.4

Q2: Find the range & Coefficient of Range:

X	10	20	30	40	50	60	70
F	15	18	25	30	16	10	9

Ans: 60, 0.75

Q3: Find the range & Coefficient of Range:

Size	5-10	10-15	15-20	20-25	25-30
F	4	9	15	30	40

Ans: 25, 5/7

Brinder Singh, Assistant Professor, PCTE

Interquartile Range & Quartile Deviation

- ***Interquartile Range*** is the difference between the upper quartile (Q_3) and the lower quartile (Q_1)
- It covers dispersion of middle 50% of the items of the series
- Symbolically, Interquartile Range = $Q_3 - Q_1$

- ***Quartile Deviation*** is half of the interquartile range. It is also called Semi Interquartile Range
- Symbolically, Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

- ***Coefficient of Quartile Deviation***: It is the relative measure of quartile deviation.
- Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Birinder Singh, Assistant Professor, PCTE

Example

- To calculate the inter-quartile range we must first find the quartiles.
- There are three quartiles, called Q1, Q2 & Q3. We do not need to worry about Q2 (this is just the median).
- Q1 is simply the middle value of the bottom half of the data and Q3 is the middle value of the top half of the data.

Quartile Formula

The Quartile Formula for Q1 = $\frac{1}{4}(n+1)^{th}$ term

The Quartile Formula for Q3 = $\frac{3}{4}(n+1)^{th}$ term

The Quartile Formula for Q2 = Q3 – Q1 (Equivalent to Median)

Main: find quartiles

- Here are the maths test results of 23 male students:

3, 3, 3, 3, 4, 4, 4, 5, 6, 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10

- Lower Quartile: $\frac{n+1}{4} = \frac{23+1}{4} = \frac{24}{4} = 6^{th} = 4$
- Median: $\frac{n+1}{2} = \frac{23+1}{2} = \frac{24}{2} = 12^{th} = 7$
- Upper Quartile: $3 \times LQ = 18^{th} = 9$

#Note: If the result is a fraction value then we will take the next integer observation.

PRACTICE PROBLEMS – IQR & QD

Q1: Find interquartile range, quartile deviation and coefficient of quartile deviation:

28, 18, 20, 24, 27, 30, 15 *Ans: 10, 5, 0.217*

Q2:

X	10	20	30	40	50	60
F	2	8	20	35	42	20

Ans: 10, 5, 0.11

Q3:

Age	0-20	20-40	40-60	60-80	80-100
Persons	4	10	15	20	11

Ans: 14.33, 0.19

Birinder Singh, Assistant Professor, PCTE

Example:

The Automobile Association checks the prices of gasoline before many holiday weekends. Listed below are the self-service prices for a sample of 8 retail outlets during the May 2004 Memorial Day Weekend.

40, 22, 60, 30, 45, 66, 70, 55

Determine the quartile deviation, Interquartile range and semi-interquartile range.

Mean deviation or Average deviation:

Definition of Mean deviation:

Mean deviation is the mean of absolute deviations of the items from an average like mean, median or mode. Normally, we consider the arithmetic mean as the average.

Notation: For Mean Deviation about Mean: M.D($\bar{x}$)
Mean Deviation about Median: M.D(Me)
Mean Deviation about Mode: M.D(Mo)

1. Mean deviation for ungrouped data:

If $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$ be a set of n observations or values, then the mean deviation is expressed and defined as:

i. Mean deviation about arithmetic mean, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$; $\bar{X} = Arithmetic\ mean$

ii. Mean deviation about median, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n}|x_i-Me|}{n}$; $Me = Median$

iii. Mean deviation about mode, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n}|x_i-Mo|}{n}$; $Mo = Mode$

Measures of Dispersion

Example:

Find out the mean deviation from the following given set 2,3,4,5,6.

Solution: We know,

$$\text{Mean deviation, M.D }(\bar{X}) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

$$\text{Here, } \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+3+4+5+6}{5} = 4$$

$$\text{Mean deviation, M.D }(\bar{X}) = \frac{|2-4|+|3-4|+|4-4|+|5-4|+|6-4|}{5}$$

$$= \frac{6}{5} = 1.2$$

Example:

The number of patients seen in the emergency room at Ibrahim Memorial Hospital for a sample of 5 days last year were: 103, 97, 101, 106 and 103.

Determine the mean deviation and interpret.

Solution: Table for calculation of Mean Deviation

No. of patients (x)	$\lvert x-\bar{x}\rvert = \lvert x-102\rvert$
103	$\lvert x-102\rvert = \lvert 103-102\rvert = 1$
97	$\lvert x-102\rvert = \lvert 97-102\rvert = 5$
101	1
106	4
103	1
$\sum x_i = 510$	$\sum\lvert x_i-\bar{x}\rvert = 12$

Arithmetic Mean,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{510}{5} = 102$$

Mean Deviation about mean, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} = \frac{12}{5} = 2.4$

Interpretation:

The mean deviation is 2.4 patients per day. The no. of patients deviates, on average by 2.4 patients from the mean of 102 patients per day.

1. Mean deviation for grouped data:

If $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$ occur with frequencies $f_1, f_2, f_3, \ldots\ldots\ldots\ldots\ldots, f_n$ respectively then the mean deviation can be written as:

i. Mean deviation about arithmetic mean, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{n}$

ii. Mean deviation about median, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n} f_i |x_i - Me|}{n}$

iii. Mean deviation about mode, M.D $(\bar{X}) = \frac{\sum_{i=1}^{n} f_i |x_i - Mo|}{n}$

Example:

Calculate mean deviation from the following data

Income (TK.)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Persons	6	8	10	12	7	4	3

Solution:

Table for calculation of Mean Deviation from mean

Income (TK.)	Class mid-point (x)	No. of persons (f)	fx	$\|x-\bar{x}\|$	$f\|x-\bar{x}\|$
0-10	5	6	30	26	156
10-20	15	8	120	16	128
20-30	25	10	250	6	60
30-40	35	12	420	4	48
40-50	45	7	315	14	98
50-60	55	4	220	24	96
60-70	65	3	195	34	102
Total		$\sum f_i = 50$	$\sum f_i x_i = 1550$		$\sum f_i \|x_i - \bar{x}\| = 688$

We know, Arithmetic Mean,

$$(\bar{x}) = \frac{\sum_{i=1}^{n} f_i x_i}{n} = \frac{1550}{50} = 31$$

Therefore, Mean deviation about arithmetic mean,

$$\text{M.D}\,(\bar{X}) = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{n} = \frac{688}{50} = 13.76$$

PRACTICE PROBLEMS – MEAN DEVIATION

Q1: Calculate M.D. from Mean & Median & coefficient of Mean Deviation from the following data: 20, 22, 25, 38, 40, 50, 65, 70, 75

Ans: 17.78, 17.22, 0.39,0.43

Q2:

X:	20	30	40	50	60	70
f:	8	12	20	10	6	4

Ans: 10.67, 10.33, 0.26, 0.26

Q3: Calculate M.D. from Mean & its coefficient:

X:	0-10	10-20	20-30	30-40	40-50
f:	5	8	15	16	6

Ans: 9.44, 0.349

Brinder Singh, Assistant Professor, PCTE

Mean Deviation

Merits

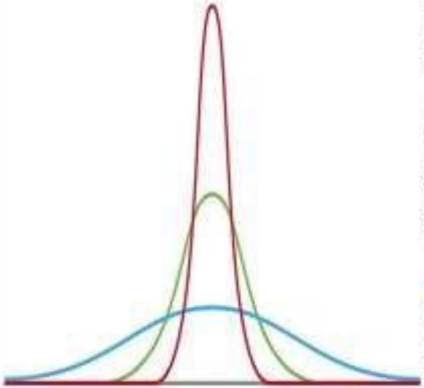
- Simple to understand
- Easy to compute
- Less effected by extreme items
- Useful in fields like Economics, Commerce etc.
- Comparisons about formation of different series can be easily made as deviations are taken from a central value

Demerits

- Ignoring '±' signs are not appropriate
- Not accurate for Mode
- Difficult to calculate if value of Mean or Median comes in fractions
- Not capable of further algebraic treatment
- Not used in statistical conclusions.

Birinder Singh, Assistant Professor, PCTE

WHAT IS VARIANCE?



Variance is a measure that tells us how spread out or scattered the values in a dataset are. It quantifies the average squared difference between each data point and the mean of the dataset.

Imagine a group of people's heights: if the heights vary greatly, the variance **will be high**, indicating a wide range of values. Conversely, if the heights are similar, the variance **will be low**, indicating a narrow range.

Variance helps us understand the dispersion or variability in a dataset, **enabling comparisons between different sets of data** and aiding in decision-making and analysis.

Formula of Variance And Standard Deviation (For Population and Sample)

	Population	Sample
Variance	$\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Standard deviation	$\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

BYJU'S The Learning App

How To Calculate Variance

Data	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
5	−4	16
6	−3	9
8	−1	1
9	0	0
10	1	1
11	2	4
14	5	25

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

STANDARD DEVIATION

- Quartile deviation considers only 50% of the item and ignores the other 50% of items in the series.
- Mean deviation no doubt an improved measure but ignores negative signs without any basis.
- Karl Pearson after observing all these things has given us a more scientific formula for calculating or measuring dispersion. While calculating SD we take deviations of individual observations from their AM and then each squares. The sum of the squares is divided by the number of observations. The square root of this sum is knows as standard deviation.

Advantages and Disadvantages of the standard deviation

Advantages:

- ❑ Lends itself to computation of other stable measures.
- ❑ Average of deviations around the mean.
- ❑ Majority of data within one standard deviation above or below the mean.
- ❑ Not expressed in squared units, so makes more sense descriptively.

Disadvantages:

- ❑ Influenced by extreme scores.

Coefficient of Variation (CV):

Coefficient of variation is the most commonly used measure of relative measures of dispersion. It is 100 times of a ratio of the standard deviation to the arithmetic mean. It is denoted by C.V. and written as:

$$C.V = \frac{\sigma}{\bar{x}} \times 100; \ \bar{x} \neq 0$$

Example:

Compute the coefficient of variation from the following data:

Monthly Income	2501-5000	5001-7500	7501-10000	10001-12500	12501-15000
No. of Families	65	130	215	100	70

Ques 1:Which data is more consistent?
Hints: Less CV more consistent.

x_i	x_i^2
3	9
5	25
6	36
4	16
3	9
5	25
4	16
$\Sigma x_i = 30$	$\Sigma x_i^2 = 136$

Number of guavas, $n = 7$

$$\text{Mean } \bar{x}_1 = \frac{30}{7} = 4.29$$

$$\text{Standard deviation } \sigma_1 = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

$$\sigma_1 = \sqrt{\frac{136}{7} - \left(\frac{30}{7}\right)^2} = \sqrt{19.43 - 18.40} \simeq 1.01$$

Coefficient of variation for guavas

$$C.V_1 = \frac{\sigma_1}{\bar{x}_1} \times 100\% = \frac{1.01}{4.29} \times 100\% = 23.54\%$$

x_i	x_i^2
1	1
3	9
7	49
9	81
2	4
6	36
2	4
$\Sigma x_i = 30$	$\Sigma x_i^2 = 184$

Number of oranges $n = 7$

$$\text{Mean } \bar{x}_2 = \frac{30}{7} = 4.29$$

$$\text{Standard deviation } \sigma_2 = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

$$\sigma_2 = \sqrt{\frac{184}{7} - \left(\frac{30}{7}\right)^2} = \sqrt{26.29 - 18.40} = 2.81$$

Coefficient of variation for oranges

$$C.V_2 = \frac{\sigma_2}{\bar{x}_2} \times 100\% = \frac{2.81}{4.29} \times 100\% = 65.50\%$$

Ques 2:Which data is more consistent?

Hints: Less CV more consistent.

Example:

The run-scores of two cricketers for 10 innings are given below:

Cricketers-A	114	45	0	31	75	102	198	8	0	7
Cricketers-B	15	25	18	30	11	4	23	21	31	22

Who of the two is a more consistent batsman?

Solution:

In order to find out who batsman is more consistent, we have to calculate the coefficient of variation for each batsman.

Table for calculation of C.V.

Cricketer-A		Cricketer-B	
Score (x)	x^2	Score (y)	y^2
114	12996	15	225
45	2025	25	625
0	0	18	324
31	961	30	900
75	5625	11	121
102	10404	4	16
198	39204	23	529
8	64	21	441
0	0	31	961
7	49	22	484
$\sum x = 580$	$\sum x^2 = 71328$	$\sum y = 200$	$\sum y^2 = 4626$

We know, Coefficient of Variation, $C.V = \frac{\sigma}{\bar{x}} \times 100$

For Cricketer-A

Arithmetic Mean,

$$\bar{x} = \frac{\sum x}{n} = \frac{580}{10} = 58$$

Standard Deviation,

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

$$= \sqrt{\frac{71328}{10} - \left(\frac{580}{10}\right)^2}$$

$$= 61.39$$

Coefficient of Variation,

$$C.V = \frac{\sigma}{\bar{x}} \times 100 = \frac{61.39}{58} \times 100$$

$$= 105.84\%$$

For Cricketer-B

Arithmetic Mean,

$$\bar{y} = \frac{\sum y}{n} = \frac{500}{10} = 20$$

Standard Deviation,

$$\sigma = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}$$

$$= \sqrt{\frac{4626}{10} - \left(\frac{200}{10}\right)^2}$$

$$= 7.91$$

Coefficient of Variation,

$$C.V = \frac{\sigma}{\bar{y}} \times 100 = \frac{7.91}{20} \times 100$$

$$= 39.56\%$$

Comment:

From the above result, we see that, C.V. (A) = 105.84% and C.V. (B) 39.56%

Since, C.V. (A)> C.V. (B). Therefore, the cricketer-B is a more consistent.