CHAPTER 4

MEASURES OF VARIATION OR DISPERSION

- An average, such as the mean or median only locates the centre of the data. It is valuable from that standpoint, but an average does not tell us anything about the spread (or scatter) of the data.
- If your nature guide told you that the river ahead averaged 3 feet in depth, would you cross it without additional information?
- Probably not! You would want to know something about the variation in the depth. Is the maximum depth of the river 3.25 feet and the minimum 2.75 feet? If that is the case, would you probably agree to cross?
- What if you learned the river depth ranged from 0.5 feet to 5.5 feet? Your decision would probably be not to cross. Before making a decision to cross the river, you want information on both the typical depth and variation in depth of the river. The degree to which numerical data tend to spread about an average value is called dispersion or variation.

EXAMPLE 1

A testing lab wishes to test two brands of outdoor paints to see how long each will last before fading. The fading lab marks 6 gallons of each of the paints to test. Since the chemical agents were added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown below. Find the mean of each group

BRAND A

10

60

50

30

40

20

The mean for Brand A is

$\boldsymbol{\mu} = \frac{\sum \boldsymbol{x}}{\boldsymbol{N}} =$

$\boldsymbol{\mu} =$

BRAND B

35

45

30

35

40

25

The mean for Brand B is

$\boldsymbol{\mu} = \frac{\sum \boldsymbol{x}}{\boldsymbol{N}} =$

$\boldsymbol{\mu} =$

SIGNIFICANCE OF VARIATION

1. Measuring variability determines the reliability of an average by pointing out as to how far an average is representative of the entire data.
2. Another purpose of measuring variability is to determine the nature and cause of variation in order to control the variation itself.
3. Measures of variation enable comparisons of two or more distribution with regard to their variability.
4. Measuring variability is of great importance to advanced statistical analysis. For example, sampling or statistical inference is essentially a problem in measuring variability.

MEASURES OF VARIATION

1. Range
2. Mean Deviation
3. Standard Deviation
4. Variance

RANGE

It is the simplest measure of dispersion. And it is the difference between the highest and the lowest values in the data set.

Range = Highest Value – Lowest Value

Calculate the Range for Brand A and Brand B in example 1.

BRAND A	BRAND B
10	35
60	45
50	30
30	35
40	40
20	25

MEAN DEVIATION (Ungrouped Data)

It is the arithmetic mean of the absolute values of deviations from the arithmetic mean.

Mean Deviation (MD) = $\frac{\sum|x-\bar{x}|}{n}$

x = the value of each observation

$\bar{x}$ = the arithmetic mean of the values

n = the number of observations

| | = absolute value

Example 4.1

The number of patients seen in the emergency room at St. Luke's Memorial Hospital for a sample of 5 days last year weeks: 103, 97, 101, 106 and 103. Determine the mean deviation and interpret.

Day	**No of cases (x)**	**x - $\bar{x}$**	**$\lvert x - \bar{x} \rvert$**
1	103		
2	97		
3	101		
4	106		
5	103		

MEAN DEVIATION (Grouped Data)

The mean absolute deviation of a grouped data about the mean is the weighted mean of the absolute values of deviation..

$$\textbf{Mean Deviation (MD)} = \frac{\sum f|x-\bar{x}|}{\sum f}$$

MEAN DEVIATION (Grouped Data)

Steps

1. Determine the mean for the frequency
2. Determine the absolute deviation of the midpoint of each class from the mean $|\boldsymbol{x} - \bar{\boldsymbol{x}}|$
3. Multiply $|\boldsymbol{x} - \bar{\boldsymbol{x}}|$ by $\boldsymbol{f}$
4. Compute the arithmetic mean of the absolute deviation

Example 4.2

Compute the mean deviation for the frequency distribution for the data below. It shows the annual income of 50 employees.

CLASS LIMIT	CLASS FREQUENCY
120 – 139	1
140 – 159	4
160 – 179	10
180 – 199	14
200 – 219	12
220 – 239	6
240 – 259	2
260 – 279	1

Class Limit	***f***	***x class midpoint***	***fx***	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
120 – 139	1					
140 – 159	4					
160 – 179	10					
180 – 199	14					
200 – 219	12					
220 – 239	6					
240 – 259	2					
260 – 279	1					
	$\sum f =$		$\sum fx =$			$\sum f\lvert x - \bar{x} \rvert =$

VARIANCE AND STANDARD DEVIATION

Variance

It is the arithmetic mean of the squared deviations from the mean.

Standard Deviation

It is the positive square root of the variance.

Population Variance for ungrouped data

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

σ^2 = population variance

x = the value of an observation in the population

μ = the arithmetic mean of the population

N = the total number of observations in the population

Population Standard Deviation for Ungrouped Data

- $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$

Example 4.3

The ages of all patients in the isolation ward of a hospital are 38, 26, 13, 41 and 22 years.

i. What is the population variance?

ii. What is the population standard deviation?

Solution

Age (x)	$x - \mu$	$(x - \mu)^2$
38		
26		
13		
41		
22		
$\sum x =$		$\sum (x - \mu)^2 =$

Sample Variance

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

Or

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

Sample Standard Deviation

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

OR

$$S = \sqrt{\frac{\sum(x^2) - \frac{(\sum x)^2}{n}}{n - 1}}$$

Example 4.4

- The hourly wages for a sample of part-time employees at Fruit Parkers Inc. are $2, $10, $6, $8 and $9. What is the sample variance and sample standard deviation?

Hourly wage (x)	$x-\bar{x}$	$(x-\bar{x})^2$
2		
10		
6		
8		
9		

Standard Deviation with frequency Distribution

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

$$= \sqrt{\frac{\sum fX^2}{N} - (\mu)^2}$$

Standard Deviation with frequency Distribution

Also for a sample, $\boldsymbol{S} = \sqrt{\dfrac{\sum \boldsymbol{fx^2} - \dfrac{(\sum \boldsymbol{fx})^2}{\boldsymbol{n}}}{\boldsymbol{n-1}}}$

Also for a sample, $\boldsymbol{S} = \sqrt{\dfrac{\boldsymbol{n}(\sum \boldsymbol{fx^2}) - (\sum \boldsymbol{fx})^2}{\boldsymbol{n}(\boldsymbol{n-1})}}$

Compute the standard deviation for the frequency distribution for the data below. It shows the annual income of 50 employees.

CLASS LIMIT	CLASS FREQUENCY
120 – 139	1
140 – 159	4
160 – 179	10
180 – 199	14
200 – 219	12
220 – 239	6
240 – 259	2
260 – 279	1

RELATIVE DISPERSION

- A direct comparison of two or more measures of dispersion – say, the standard deviation for a distribution of annual incomes and the standard deviation of a distribution of absenteeism for this same group of employees – is impossible. Can we say that the standard deviation of 1200 for the income distribution is greater than the standard deviation of 4.5 days for the distribution of absenteeism?
- Obviously not, because we cannot directly compare dollars and days absent from work. In order to make a meaningful comparison of the dispersion in incomes and absenteeism, we need to convert each of these to a relative value – that is a percent

COEFFICIENT OF VARIATION

- The coeffiecient of variation (CV) is a ratio of the Standard Deviation and the mean expressed as a percentage.
- $CV = \frac{S}{\bar{x}} (100)$

COEFFICIENT OF VARIATION

Example

A study of the test scores for an in-plant course in Management Principles and years of service of the employees enrolled in the course resulted in these statistics.

- The mean test score was 200
- The standard deviation was 40
- The mean number of years of service was 20 years
- The standard deviation was 2 years.

Compare the relative dispersions in the distributions using the coefficient of variation.

COEFFICIENT OF VARIATION

Example

The variation in the annual incomes of executives is to be compared with the variation of incomes of unskilled employees. For a sample of executives, mean = $500,000 and S = $50,000.

For a sample of unskilled employees, mean = $22,000 and S = $2,200. We are tempted to say that there is more dispersion in the annual incomes of the executive because $50,000> $2,200.

The means are so far apart, however, that we need to convert the statistics to coefficient of variation to make a meaningful comparison of the variation in annual incomes.

COEFFICIENT OF SKEWNESS

- $SK = \frac{3\ (mean - median)}{standard\ deviation}$

COEFFICIENT OF SKEWNESS

Example

The lengths of stay on the cancer floor of a hospital were organized into a frequency distribution. The mean length of stay was 28 days; the median was 25 days and the modal length is 23 days. The standard deviation was computed to be 4.2 days.

a. Is the distribution symmetrical, positively skewed, or negatively skewed?

b. What is the coefficient of skewness? Interpret

CHEBYSHEV'S THEOREM

We have known that a small standard deviation for a set of values indicates that these values are located close to the mean. Conversely, a large standard deviation reveals that the observations are widely scattered about the mean.

The Russian Mathematician P. L. Chebyshev (1821 – 1894) developed a theorem that allows us to determine the minimum proportion of the values that live within a specified number of standard deviations of the mean.

For example, according to Chebyshev‟s theorem, at least three-fourth or 75% of the data will fall within 2 standard deviations of the mean of the set of data.

CHEBYSHEV'S THEOREM

The Chebyshev's Theorem states that for any set of observations (sample or population), the population of values from a set of data that will fall within K standard deviations of the mean will be at least $\mathbf{1 - \frac{1}{K^2}}$, where K is a number greater than 1 (K is not necessarily an integer).

CHEBYSHEV'S THEOREM

For any set of date, at least three-fourth or 75% of the data will fall within 2 standard deviations of the mean of the data set. This result is found by substituting K = 2 in the expression.

$$1 - \frac{1}{K^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0.75 = 75\%$$

CHEBYSHEV'S THEOREM

For any set of date, at least eight-ninth or 88.89% of the data values will fall within 3 standard deviations of the mean of the data set. This result is found by substituting K = 3 in the expression.

$$\mathbf{1 - \frac{1}{K^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 0.8889 = 88.89\%}$$

CHEBYSHEV'S THEOREM

For example, if two variables measured in the same units have the same mean, say 60 and Variable A has a standard deviation of 1.5 and Variable B has a standard deviation of 8, then the data for Variable B will be more spread out than that of Variable A.

EXAMPLE

In a distribution, the mean amount was $51.54 and the standard deviation was imputed to be $7.51. At least what percentage of the distribution lie within plus 3.5 standard deviation and minus 3.5 standard deviation of the mean?

EXAMPLE

The mean price of houses in a certain neighbourhood is GH¢ 60,000 and the standard deviation is GH¢ 10,000. Find the price range for which at least 75% of the houses will sell.

EXAMPLE

A survey of local companies found that the mean amount of travel allowance for executives was $0.25 per mile. The standard deviation was $0.02.

Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between $0.20 and $0.30

THE EMPIRICAL RULE

Chebyshev‟s theorem is concerned with any set of values, that is, the distribution of values can have any shape. However, for a symmetrical, bell-shaped distribution such as the one shown below, we can be more precise in explaining the dispersion about the mean.

These relationships involving the standard deviation and the mean are included in the Empirical Rule, sometime known as the Normal rule.

Empirical Rule: for a symmetrical, bell-shaped distribution, approximately **68% of the observations** will lie within plus and minus **one** standard deviation of the mean; about **95% of the observations** will lie within plus and minus **two** standard deviations and **practically all (99.7%)** will lie plus and minus **three** standard deviations of the mean.

THE EMPIRICAL RULE

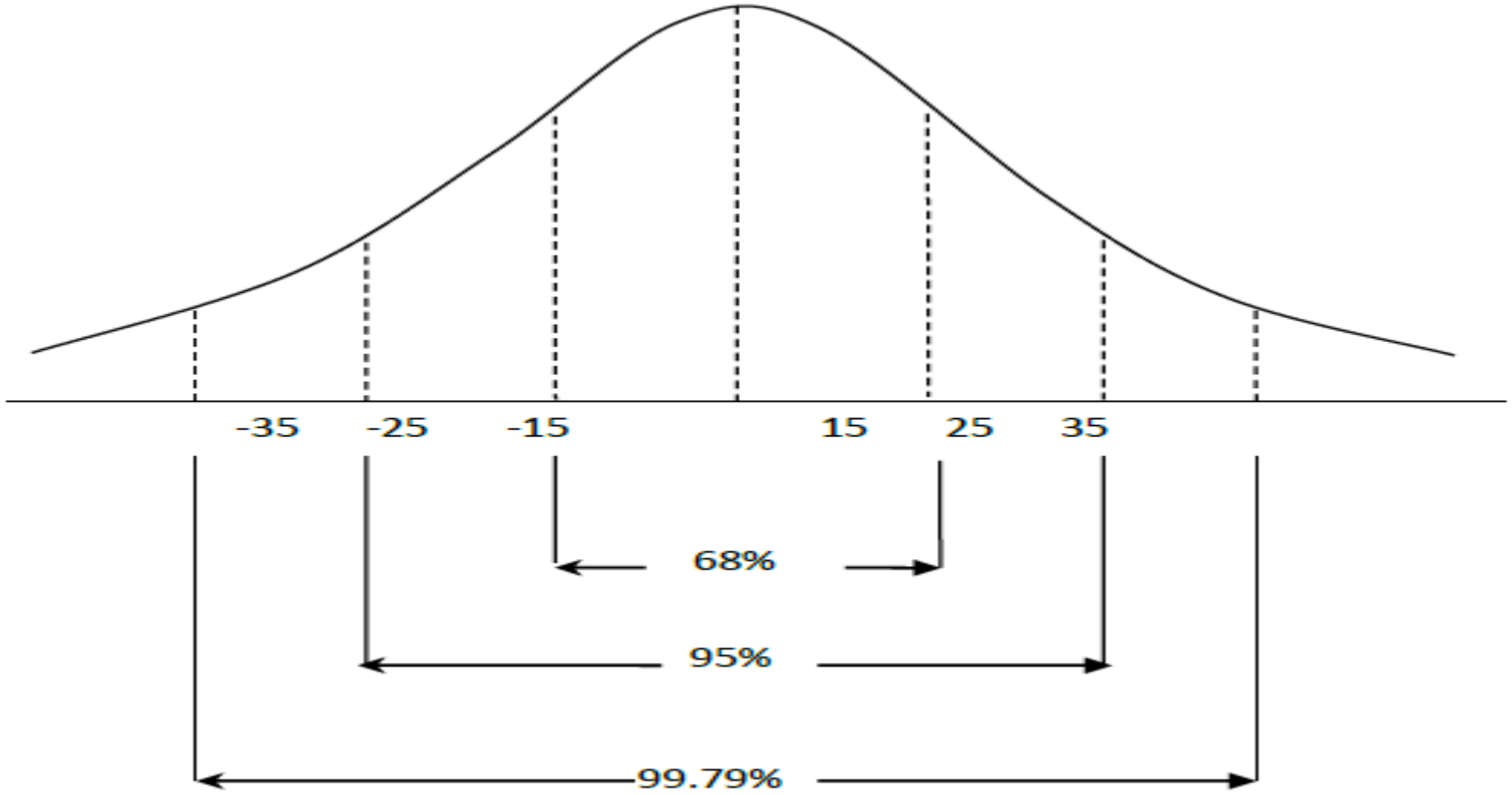

Figure 4.1—A symmetrical bell-shape distribution, shown
The Relationship between Standard Deviation and the Mean

THE EMPIRICAL RULE

A sample of the monthly amounts spent for food by a senior citizen living alone approximates a symmetrical, bell-shaped frequency distribution. The sample mean is $150; the standard deviation is $20. Using the empirical rule,

1. About 68 percent of the monthly food expenditure is between what two amounts?

2. About 95 percent of monthly food expenditure is between what two amounts?

3. Almost all of the monthly expenditure is between what two amounts?

THE EMPIRICAL RULE

A sample of the monthly amounts spent for food by a senior citizen living alone approximates a symmetrical, bell-shaped frequency distribution. The sample mean is $150; the standard deviation is $20. Using the empirical rule,

1. About 68 percent of the monthly food expenditure is between what two amounts?

2. About 95 percent of monthly food expenditure is between what two amounts?

3. Almost all of the monthly expenditure is between what two amounts?

OTHER MEASURES OF DISPERSION

- QUARTILES
- DECILES
- PERCENTILES

QUARTILES

The three values, which split a distribution into four equal portions, are known as the quartiles.

- **First Quartile**

The first (lower) quartile Q_1 of a set of data is the value such that 25% of the observations are smaller than Q_1 and 75% are larger than the Q_1.

- **Second Quartile**

The second (middle) quartile Q_2 of a set of data is the median – that is 50% of the observations are smaller than Q_2 and 50% are larger than Q_2.

- **Third Quartile**

The third (upper) quartile Q3 of a set of data is the value such that 75% of all the observations are smaller and 25% of the observations are larger thanQ3.

QUARTILES FOR UNGROUPED DATA

The first, second and third quartiles (Q_1, Q_2, Q_3);

$$Q_1 = \frac{1}{4}(n+1)th$$

$$Q_2 = \frac{2}{4}(n+1)th$$

$$Q_3 = \frac{3}{4}(n+1)th$$

EXAMPLE 4.12

Determine the quartile of the following marks obtained by 10 candidates in an examination

24, 23, 28, 15, 10, 40, 42, 32, 48, 8.

- Steps

1. Arrange the data in an array. n= 10

QUARTILES FOR GROUPED DATA

The first, second and third quartiles (Q_1, Q_2, Q_3);

$$Q_q = L_q + \frac{\frac{q}{4}n - F}{f_q} \times C$$

Q_q = *qth quartile*

L_q = *lower class boundary of the class interval containing the qth quartile*

n = *total frequency*

F = *sum of frequencies (Cumulative frequency) for all class intervals before the class interval containing the qth quartile*

f_q = *frequency of the class interval containing the qth quartile.*

C = *width of the class containing the qth quartile.*

Example

Find the first, second and third quartile for the data below.

It shows the annual income of 50 employees in hundred Ghana cedis.

CLASS LIMIT	CLASS FREQUENCY
120 – 139	1
140 – 159	4
160 – 179	10
180 – 199	14
200 – 219	12
220 – 239	6
240 – 259	2
260 – 279	1

CLASS LIMIT	CLASS BOUNDARIES	f	Cumulative frequency
120 – 139	119.5 – 139.5	1	1
140 – 159	139.5 – 159.5	4	5
160 – 179	159.5 – 179.5	10	15
180 – 199	179.5 – 199.5	14	29
200 – 219	199.5 – 219.5	12	41
220 – 239	219.5 – 239.5	6	47
240 – 259	239.5 – 259.5	2	49
260 – 279	259.5 – 279.5	1	50

DECILES FOR UNGROUPED DATA

The first, second and third…ninth deciles (D_1,D_2, D_3 ... . D_9);

$$D_1 = \frac{1}{10}(n+1)th$$

$$D_2 = \frac{2}{10}(n+1)th$$

$$D_3 = \frac{3}{10}(n+1)th$$

$$D_9 = \frac{9}{10}(n+1)th$$

EXAMPLE

Determine the 4^{th} and 7^{th} decile of the following marks obtained by 10 candidates in an examination

24, 23, 28, 15, 10, 40, 42, 32, 48, 8.

- Steps

1. Arrange the data in an array. n= 10

DECILES FOR GROUPED DATA

The d^{th} deciles for grouped data is given by

$$D_d = L_d + \frac{\frac{d}{10}n - F}{f_d} \times C$$

$D_d = d^{th}$ *decile*

$L_d =$ *lower class boundary of the class interval containing the* d^{th} *decile*

$n =$ *total frequency*

$F =$ *sum of frequencies (Cumulative frequency) for all class intervals before the class interval containing the* d^{th} *decile*

$f_d =$ *frequency of the class interval containing the* d^{th} *decile.*

$C =$ *width of the class containing the* d^{th} *decile.*

Example

Find the first, fifth and ninth decile for the data below.

CLASS LIMIT	CLASS FREQUENCY
120 – 139	1
140 – 159	4
160 – 179	10
180 – 199	14
200 – 219	12
220 – 239	6
240 – 259	2
260 – 279	1

CLASS LIMIT	CLASS BOUNDARIES	f	Cumulative frequency
120 – 139	119.5 – 139.5	1	1
140 – 159	139.5 – 159.5	4	5
160 – 179	159.5 – 179.5	10	15
180 – 199	179.5 – 199.5	14	29
200 – 219	199.5 – 219.5	12	41
220 – 239	219.5 – 239.5	6	47
240 – 259	239.5 – 259.5	2	49
260 – 279	259.5 – 279.5	1	50

PERCENTILES FOR UNGROUPED DATA

The first, second and third…ninety-ninth percentiles($P_1, P_2, P_3 \ldots \ldots P_{99}$);

$$P_1 = \frac{1}{100}(n+1)th$$

$$P_2 = \frac{2}{100}(n+1)th$$

$$P_3 = \frac{3}{100}(n+1)th$$

$$P_{99} = \frac{99}{100}(n+1)th$$

PERCENTILES FOR GROUPED DATA

The p^{th} deciles for grouped data is given by

$$P_p = L_p + \frac{\frac{p}{100}n - F}{f_p} \times C$$

$P_p = p^{th}$ *percentiile*

$L_p =$ *lower class boundary of the class interval containing the* p^{th} *percentile*

$n =$ *total frequency*

$F =$ *sum of frequencies (Cumulative frequency) for all class intervals before the class interval containing the* p^{th} *percentile*

$f_p =$ *frequency of the class interval containing the* p^{th} *percentile.*

$C =$ *width of the class containing the* p^{th} *percentile.*

Example

Find the 10^{th} and 90^{th} percentile for the data below.

CLASS LIMIT	CLASS FREQUENCY
120 – 139	1
140 – 159	4
160 – 179	10
180 – 199	14
200 – 219	12
220 – 239	6
240 – 259	2
260 – 279	1

CLASS LIMIT	CLASS BOUNDARIES	f	Cumulative frequency
120 – 139	119.5 – 139.5	1	1
140 – 159	139.5 – 159.5	4	5
160 – 179	159.5 – 179.5	10	15
180 – 199	179.5 – 199.5	14	29
200 – 219	199.5 – 219.5	12	41
220 – 239	219.5 – 239.5	6	47
240 – 259	239.5 – 259.5	2	49
260 – 279	259.5 – 279.5	1	50

PERCENTILES

The percentile corresponding to a given value of X is computed by the formula

$$Percentile = \frac{(no.\ of\ values\ below\ X)+0.5}{total\ no.of\ values} \times 100\%$$

EXAMPLE

A teacher gives 20-point test to 10 students. The scores are shown below. Find the percentile rank for a score of 12 and 6.

18,15,12,6,8,2,3,5,20,10.