3.2 Measures of Spread/ Dispersion

Spread is the degree of scatter or variation of the variable about the central value. Examples of these measures includes: the range, Inter-Quartile range, Quartile Deviation also called semi Inter-Quartile range, Mean Absolute Deviation, Variance and standard deviation.

Inter-Quartile range and Semi Inter-Quartile Range

Inter-Quartile range (IQR) is the difference between the upper and lower quartiles. Half of this difference is called Quartile Deviation or the semi Inter-Quartile range (SIQR) Ie

$$\text{IQR} = Q_3 - Q_1 \quad \text{and} \quad \text{SIQR} = \tfrac{1}{2}(Q_3 - Q_1)$$

Mean Absolute Deviation (MAD)

It is the average of the absolute deviations from the mean and it's given by

$\text{MAD} = \dfrac{\sum |x - \bar{x}|}{n}$ for ungrouped data but for grouped data $\text{MAD} = \dfrac{\sum f|x - \bar{x}|}{n}$

Example 1 Find the quartile deviation and the mean absolute deviation for the data.

3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13

Solution

Sorted data: 3, 5, 6, 6, 7, 9, 10, 12, 13, 13, 15

Recall $Q_1 = 6$ and $Q_3 = 13$ ie from earlier calculations.

Thus $\text{SIQR} = \tfrac{1}{2}(Q_3 - Q_1) = \tfrac{1}{2}(13 - 6) = 3.5$

$$\bar{x} = \frac{3+5+6+6+7+9+10+12+13+13+15}{11} = 9$$

$$\text{MAD} = \frac{\sum |x - \bar{x}|}{n} = \frac{|3-9| + |5-9| + |6-9| + \ldots + |15-9|}{11} = \frac{6+4+3+\ldots+6}{11} = \frac{36}{11} \approx 3.2727$$

Variance and Standard Deviation

Ignoring the negative sign in order to compute MAD is not the only option we have to deal with deviations. We can square the deviations and then average. The average of the squared deviations from the mean is called the variance denoted s^2 and its given by

$s^2 = \frac{1}{n}\sum (x - \bar{x})^2$ A little algebraic simplification of this formular gives $s^2 = \frac{1}{n}\sum x^2 - \bar{x}^2$

For grouped data $s^2 = \frac{1}{n}\sum f(x - \bar{x})^2 = \frac{1}{n}\sum fx^2 - \bar{x}^2$ where n is the sum of frequencies.

To reverse the squaring on the units we find the square root of the variance.Standard Deviation denoted s is the square root of variance.

Example 1 Find the variance and standard deviation for the data.

3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13

Solution

$$\bar{x} = \frac{3+5+6+6+7+9+10+12+13+13+15}{11} = 9$$

$$S^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{(3-9)^2 + (5-9)^2 + (6-9)^2 + \ldots + (15-9)^2}{11} = \frac{36+16+9+\ldots+36}{11} = \frac{143}{11} = 13$$

Standard deviation $s = \sqrt{\text{variance}} = = \sqrt{13} \approx 3.60555$.

Example 2 Find the standard deviation of the data: 2, 4, 8, 7, 9, 4, 6, 10, 8, and 5.

Solution

Mean $\bar{x} = \dfrac{\sum x}{n} = \dfrac{2+4+8+\ldots+5}{10} = \dfrac{63}{10} = 6.3$ and $\sum x^2 = 2^2 + 4^2 + 8^2 + \ldots + 5^2 = 455$

Standard deviation $s = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} = \sqrt{45.5 - 6.3^2} \approx 2.4104$.

Example 3 Estimate the mean, and standard deviation for the frequency table below:

Class	5-9	10-14	15-19	20-24	25-29	30-34	35-39
freq	5	12	32	40	16	9	6

Solution

Mid pts (x)	7	12	17	22	27	32	37	Total
Freq (f)	5	12	32	40	16	9	6	120
xf	35	144	544	880	432	288	222	2545
fx^2	245	1728	9248	19360	11664	9216	8214	59675

Mean $\bar{x} = \dfrac{\sum fx}{n} = \dfrac{2545}{120} \approx 21.2083$ and $\sum fx^2 = 59675$

Standard deviation $s = \sqrt{\frac{1}{n}\sum fx^2 - \bar{x}^2} = \sqrt{\dfrac{59675}{120} - 21.2083^2} \approx 6.8919$.

Exercise

1) Find the quartile deviation, the mean absolute deviation and the standard deviation of the data: a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 5, 5, 7, 8, 8 and 9
2) 2) The number of goals scored in 20 hockey matches is shown in the table.

No of goals	1	2	3	4	5
No of matches	2	5	6	3	4

Estimate the quartile deviation, the mean absolute deviation and the standard deviation of the number of goals cored

3) consider the frequency table below and estimate quartile deviation, the mean absolute deviation and the standard deviation

Class	8-12	13-17	18-22	23-27	28-32	33-37
Freq	3	10	12	9	5	1

4) The table shows the heights of 30 students in a class calculate an estimate of the quartile deviation, the mean absolute deviation and the standard deviation of the height.

Height (cm)	140<x<144	144<x<148	148<x<152	152<x<156	156<x<160	160<x<164
No of students	4	5	8	7	5	1

5) The grouped frequency table gives information about the distance each of 150 people travel to work.

Height (cm)	0<d<5	5<d<10	10<d<15	15<d<20	20<d<25	25<d<30
No of students	4	5	8	7	5	1

Calculate an estimate for the quartile deviation and the standard deviation of the distance travelled to work by the people.

Properties of Measures of Spread

a) They are not affected by change of origin. Adding or subtracting a constant from each and every observation in a data set does not affect any measures of spread. That is New measure = old measure

b) They are affected by change of scale. Multiplying each and every observation in a data set by a constant value scales up all the measures of spread by the same value except in the case of variance which is scaled up by a square of the same constant.
ie New measure $= K$(old measure) but New variance $= k^2 \times$ old variance

Example: Consider the three sets of data A, B and C below
Set A: 65, 53, 42, 52, 53 Range=23, $MAD_A = 4.8$ and Variance $_A = 66.5$
Set B: 15, 3, -8, 2, 3 Range=23, $MAD_B = 4.8$ and Variance $_B = 66.5$
Set C: 45, 9, -24, 6, 9 Range=69, $MAD_C = 14.4$ and Variance $_C = 598.5$

- Notice that set B is obtained by subtracting 50 from each and every observation in set A and clearly $MAD_B = MAD_A$ and Variance $_B =$ Variance $_A$ Therefore there is no effect on the change of origin ie New measure = old measure..
- Effectively set C is obtained by multiplying each and every observation in set B by 3 and clearly $MAD_C = 3 \times MAD_B$ and Variance $_C = 3^2 \times$ Variance $_B$ Thus
  New measure $= K$(old measure) and New Variance $_C = k^2 \times$ old Variance $_B$

Mean and Standard Deviation Using a Calculator

- When on, press mode key to get;
  COMP SD REG
  1 2 3
- Press 2 to select SD for statistical data.
- Enter data one by one pressing m+ after every value entered. **T**he screen will be showing the number of observations that are fully entered.
- Pressing shift then 1 gives
  $\sum x^2$ $\sum x$ n
  1 2 3

Typing 1 then = gives the value of $\sum x^2$
Similarly typing 2 then = gives the value of $\sum x$

- Pressing shift then 2 gives
  $\bar{x}$ $x\sigma_n$ $x\sigma_{n-1}$
  1 2 3
  Which are the mean uncorrected standard deviation and the corrected standard-deviation

Example using your calculator, obtain the mean and standard deviation of the following data: 31, 52, 29, 60, 58
Solution
Entering data 31M+ 52 M+ 29 M+ 60 M+ 58 M+
$\bar{x} = 46$ and $s = 14.91643$
Question Redo the above example using the data: 235, 693, 484, 118, 470

3.3 Assumed Mean and the Coding Formular

If the observations are too large such that the natural computation of totals is tedious, we can take one of the observations as the working/assumed mean. Let A be any guessed or assumed arithmetic mean and let $d_i = x_i - A$ be the deviations of x_i from A, then

Mean $\bar{x} = A + \frac{1}{n}\sum fd = A + \bar{d}$ and Variance $S^2 = \frac{1}{n}\sum fd^2 - \left(\frac{1}{n}\sum fd\right)^2 = \frac{1}{n}\sum fd^2 - \bar{d}^2$

3.4 Measures of Relative Dispersion:

These measures are used in comparing spreads of two or more sets of observations. These measures are independent of the units of measurement. These are a sort of ratio and are called coefficients.

Suppose that the two distributions to be compared are expressed in the same units and their means are equal or nearly equal. Then their variability can be compared directly by using their standard deviations. However, if their means are widely different or if they are expressed in different units of measurement, we can not use the standard deviations as such for comparing their variability. We have to use the relative measures of dispersion in such situations. Examples of these Measures of relative dispersion includes; Coefficient of quartile deviation, Coefficient of mean deviation and the Coefficient of variation

3.4.1 Coefficient of Quartile Deviation and Coefficient of Mean Deviation

The Coefficient of Quartile Deviation of x CQD(x) is given by $\text{CQD(x)} = \dfrac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\%$

The Coefficient of Mean Deviation CMD(x) is given by $\text{CMD(x)} = \dfrac{\text{MAD}}{\text{Mean}} \times 100\%$

3.4.2 Coefficient of Variation:

Coefficient of variation is the percentage ratio of standard deviation and the arithmetic mean. It is usually expressed in percentage. The coefficient of variation of x denoted C.V(x) is given by the formula

$$\text{C.V(x)} = \frac{S}{\bar{x}} \times 100\%$$

where $\bar{x}$ is the mean and S is the standard deviation of x.

The coefficient has no units ie it's independent of the units of measurements. It is useful in comparing spreads of two or more populations. The smaller the coefficient of variation, the higher the peak and the lower the spread and vice versa.

Note: Standard deviation is absolute measure of dispersion while. Coefficient of variation is relative measure of dispersion.

Example 1 Consider the distribution of the yields (per plot) of two ground nut varieties. For the first variety, the mean and standard deviation are 82 kg and 16 kg respectively. For the second variety, the mean and standard deviation are 55 kg and 8 kg respectively. Then we have, for the first variety $\text{C.V(x)} = \frac{16}{82} \times 100 \approx 19.5\%$

For the second variety $\text{C.V(x)} = \frac{8}{55} \times 100 \approx 14.5\%$

It is apparent that the variability in second variety is less as compared to that in the first variety. But in terms of standard deviation the interpretation could be reverse.

Example 2 Below are the scores of two cricketers in 10 innings. Find who is more „consistent scorer" by Indirect method.

A	204	68	150	30	70	95	60	76	24	19
B	99	190	130	94	80	89	69	85	65	40

Solution:

From a calculator, $\bar{x}_A = 79.6, S_A = 58.2$ $\bar{x}_B = 94.1$ and $S_B = 41.1$

Coefficient of variation for player A is $\text{C.V(x)} = \frac{58.2}{79.6} \times 100 \approx 73.153\%$

Coefficient of variation for player B is $\text{C.V(x)} = \frac{41.1}{94.1} \times 100 \approx 43.7028\%$

Coefficient of variation of A is greater than coefficient of variation of B and hence we conclude that player B is more consistent

Exercise

1) Find the coefficient of quartile deviation, the coefficient of mean deviation and the Coefficient of variation n of x for the following data: a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4
   b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9 c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13
   d) data on marks given by the table below

Marks Obtained	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	6	12	22	24	16	12	8

2) If the weights of 7 ear-heads of sorghum are 89, 94, 102, 107, 108, 115 and 126 g. Find the arithmetic mean and standard deviation using a calculator hence determine the coefficient of variation of the ear-heads of sorghum
3) The following are the 381soybean plant heights in Cms collected from a particular plot. Using coding formula, Find the mean and Standard deviation of the plants hence determine the coefficient of variation of the 1soybean plant heights:

Plant heights (Cms)	6.8-7.2	7.3-7.7	7.8-8.2	8.3-8.7	8.8-9.2	9.3-9.7	9.8-10.2	10.3-10.7	10.8-11.2	11.3-11.7	11.8-12.2	12.3-12.7
No. of Plants	9	10	11	32	42	58	65	55	37	31	24	7

3.5 Measures of Skewness and Kurtosis

3.5.1 Skewness

Before discussing the concept of skewness, an understanding of the concept of **symmetry** is essential. A plot of frequency against class mark joined with a smooth curve can help us to visually assess the symmetry of a distribution. Usually symmetry is about the central value. Symmetry is said to exist in a distribution if the smoothed frequency polygon of the distribution can be divided into two identical halves wherein each half is a mirror image of the other **Skewness** on the other hand means lack of symmetry and it can be positive or negative. Basically, if the distribution has a tail on the right, (See figure below), then the distribution is positively skewed Eg Most students having vey low marks in an examination. However if the distribution has a tail on the left, then the distribution is negatively skewed. (see figure below). Eg Most students having vey high marks in an examination

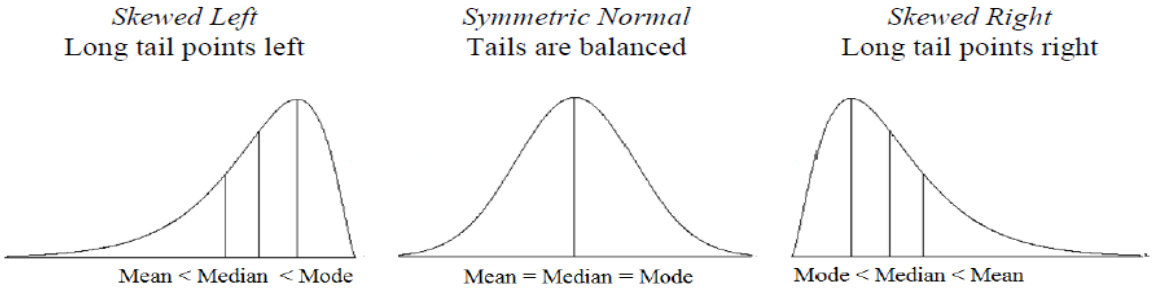


Figure 1. Sketches showing general position of mean, median, and mode in a population.

Measures of Skewness

Generally for any set of values $x_1, x_2, x_3, ...x_n$, the **moment coefficient of skewness** α_3 is given by $\alpha_3 = \dfrac{\sum f(x - \bar{x})^3}{nS^3}$ where S is the standard deviation of X. It's worth noting that if

$\alpha_3 < 0$, the distribution is negatively skewed, if $\alpha_3 > 0$, the distribution is positively skewed and if $\alpha_3 = 0$ the distribution is normal

Other measures of Skewness includes the Karl Pearson coefficient of Skewness SK_p, Bowley's coefficient of Skewness SK_B and Kelley's coefficient of Skewness SK_k.

The **Karl Pearson's coefficient of Skewness** is based upon the ***divergence of mean from mob*** in a skewed distribution. Recall the empirical relation between mean, median and mode which states that, for a moderately symmetrical distribution, we have

$$\text{Mean - Mode} = 3\,(\text{Mean - Median})$$

Hence Karl Pearson's coefficient of skewness is defined by;

$$SK_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}},$$

The **Bowley's coefficient of Skewness** is based on quartiles. For a symmetrical distribution, it is seen that Q_1 and Q_3 ***are*** equidistant ftom median.

$$SK_B = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \text{ where } Q_k \text{ is the } K^{th} \text{ quartile.}$$

The **Kelly's coefficient of Skewness** is based on P_{90} and, P_{10} so that only 10% of the observations on each extreme are ignored.. This is an improvement over the Bowley's coefficient which leaves 25% of the observatories on each extreme of the distribution.

$$SK_k = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \text{ where } P_k \text{ is the } K^{th} \text{ percentile.}$$

Interpreting Skewness

If the coefficient of skewness is positive, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If the coefficient of skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer. If the coefficient of skewness = 0, the data are perfectly symmetrical. But a skewness of exactly zero is quite unlikely for real-world data, so *how can you interpret the skewness number*? Bulmer, M. G., *Principles of Statistics* (Dover,1979) — a classic — suggests this rule of thumb: If the coefficient of skewness is:-

- less than −1 or greater than +1, the distribution is *highly skewed*.
- between −1 and $-\frac{1}{2}$. or between $+\frac{1}{2}$. and +1, the distribution is *moderately skewed*.
- between $-\frac{1}{2}$ and $+\frac{1}{2}$.., the distribution is *approximately symmetric.*

Example The following figures relate to the size of capital of 285 companies :

Capital (in *Ks* lacs.)	1-5	6-10	11-15	16-20	21-25	26-30	31-35
No. of companies	20	27	29	38	48	53	70

Compute the Bowley's coefficients of skewness and interpret the results.

Solution

Boundaries	0.5-5.5	5.5-10.5	10.5-15.5	15.5-20.5	20.5-25.5	25.5-30.5	30.5-35.5
CF	20	47	76	114	162	215	285

$$Q_1 = \tfrac{1}{4}(286)^{th} \text{ value} = 71.5^{th} \text{ value} = 10.5 + \left(\frac{71.5 - 47}{29}\right) \times 5 \approx 14.7241$$

$$Q_2 = \tfrac{1}{2}(286)^{th}\text{ value} = 143^{rd}\text{ value} = 20.5 + \left(\frac{143-114}{48}\right)\times 5 \approx 23.5208$$

$$Q_3 = \tfrac{3}{4}(286)^{th}\text{ value} = 214.5^{th}\text{ value} = 25.5 + \left(\frac{214.5-162}{53}\right)\times 5 \approx 30.4528$$

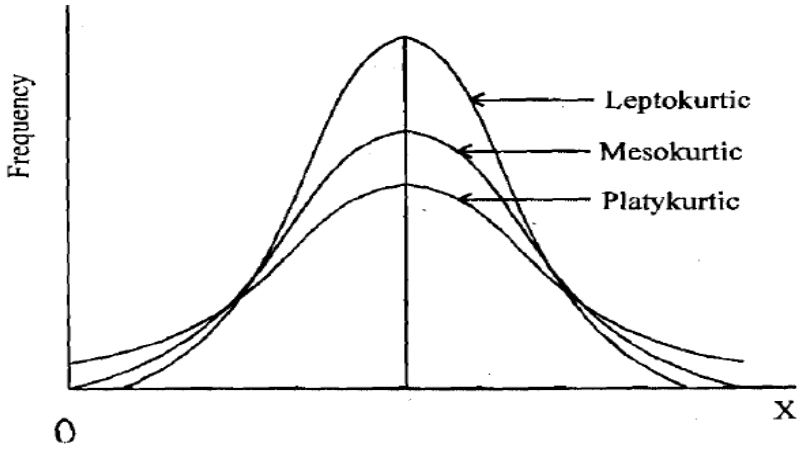
$$SK_p = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{30.4528 - 2\times 23.5208 + 14.7241}{30.4528 - 14.7241_1} \approx -0.11855.$$

This value lies between $-\frac{1}{2}$ and $+\frac{1}{2}$, therefore the distribution is approximately symmetric.

Question: Compute the Karl Pearson's and the Kelly's coefficient of skewness for the above data and interpret the results.

3.5.2 Kurtosis

It measures the peakedness of a distribution. If the values of x are very close to the mean, the peak is very high and the distribution is said to be **Leptokurtic**. On the other hand if the values of x are very far away from the mean, the peak is very low and the distribution is said to be **Pletykurtic.** Finally if x values are at a moderate distance from the mean then the peak is moderate and the distribution is said to be **mesokurtic.**.



Measures of Kurtosis

Generally for a set of values $x_1, x_2, x_3, ...x_n$, the moment coefficient of kurtosis α_4 is given by $\alpha_4 = \dfrac{\sum f(x-\bar{x})^4}{nS^4}$ where $\bar{x}$ and S are the arithmetic mean and standard deviation of X.

Example Find the coefficient of Skewness α_3 and the coefficient of kurtosis α_4 for the data 5, 6, 7, 6, 9, 4, 5

Solution

$\bar{x} = \frac{1}{n}\sum x = \frac{42}{7} = 6$ and Standard deviation $s = \sqrt{\frac{1}{n}\sum (x-\bar{x})^2} = \frac{4}{\sqrt{7}}$

x	5	6	7	6	9	4	5	Total
$(x-\bar{x})^2$	1	0	1	0	9	4	1	16
$(x-\bar{x})^3$	-1	0	1	0	27	-8	-1	18
$(x-\bar{x})^4$	1	0	1	0	81	16	1	100

Coefficient of Skewness $\alpha_3 = \dfrac{\sum (x - \bar{x})^3}{nS^3} = \dfrac{18}{7} \times \left(\dfrac{\sqrt{7}}{4}\right)^3 \approx 0.744118$

Coefficient of kurtosis $\alpha_4 = \dfrac{\sum f(x - \bar{x})^4}{nS^4} = \dfrac{100}{7} \times \left(\dfrac{\sqrt{7}}{4}\right)^4 \approx 2.73438$

Exercise

1. Find the moment coefficient of Skewness and kurtosis for the dat below. a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9 c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13
   d) data on marks given by the table below

Marks Obtained	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	6	12	22	24	16	12	8

   e) Data given by the table below

Marks Obtained	0-10	10-20	20-30	30-40
No. of Students	1	3	4	2

2. Compute the Bowley's coefficient of skewness, the Kelly's coefficient of skewness and the Percentile coefficient of kurtosis for the following data and interpret the results.
   a) 9, 3, 4, 2, 9, 5, 8, 4, 7, 4 b) 1, 2, 2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8 and 9
   c) 3, 6, 9, 10, 7, 12, 13, 15, 6, 5, 13 d) data on heights given by the table below

Heightl (in *inches*.)	58	59	60	61	62	63	64	65
No. of persons	10	18	30	42	35	28	16	8

   e) data on daily expenditure of families given by the table below

Daily Expenditure (Rs)	0-20	20-40	40-60	60-80	80-100
No. of persons	13	25	27	19	16

   f) Data on marks given by the table below

Marks Obtained	0-20	20-40	40-60	60-80	80-100
No. of Students	8	28	35	17	12

3. The following measures were computed for a frequency distribution :
   Mean = 50, coefficient of Variation = 35% and Karl Pearson's Coefficient of Skewness $SK_p = -0.25$. Compute Standard Deviation, Mode and Median of the distribution.

4. Bivariate Data

4.1 Introduction

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

A Bivariate Data is a a set of paired measurements which are of the form

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Examples

i. Marks obtained in two subjects by 60 students in a class.
ii. The series of sales revenue and advertising expenditure of the various branches of a company in a particular year.
iii. The series of ages of husbands and wives in a sample of selected married couples.

In a bivariate data, each pair represents the values of the two variables. Our interest is to find a relationship (if it exists) between the two variables under study.