Topic area: Statistics and probability

Sub-topic area: Descriptive statistics

Unit 14

Measures of dispersion

Number of periods: 7

Key unit competence

Extend understanding, analysis and interpretation of data arising from problems and questions in daily life to include the standard deviation.

Learning objectives

Knowledge and understanding	Skills	Attitudes and values
• Define the variance, standard deviation and the coefficient of variation • Analyse and critically interpret data and infer conclusion.	• Determine the measures of dispersion of a given statistical series. • Apply and explain the standard deviation as the more convenient measure of the variability in the interpretation of data • Express the coefficient of variation as a measure of spread of a set of data as a proportion of its mean.	• Appreciate the importance of measures of dispersion in the interpretation of data • Show concern on how to use the standard deviation as a measure of variability of data.

14.1 Introduction

In Junior Secondary, you were introduced to statistics. You learnt about measures of central tendencies. In this unit, we shall learn about measures of dispersion.

Activity 14.1

Discuss in groups the meaning of measures of central tendencies. What are they? Where can we apply them?

Definitions

A measure of central tendency; also called average, is values about which the distribution of data is approximately balanced. There are three types of measure of central tendency namely the mean, the median and the mode.

***Mean*:** is the sum of data values divided by the number of values in the data

Mode: is the value that occurs most often in the data.

Median: is the middle value when the data is arranged in order of magnitude

The mean

The mean value of a set of data is the sum of all the items in the set of data divided by the number of items.

For discrete raw data mean = $\frac{\text{sum of items}}{\text{number of items}}$ i.e. $\overline{x} = \frac{\Sigma x}{n}$ where n is the number of items.

For example the mean of the numbers 8,10,11,13,15,16,19 and 22 is given by

$$\overline{x} = \frac{8 + 10 + 11 + 13 + 15 + 16 + 19 + 22}{8} = \frac{144}{8} = 14.25$$

For data in ungrouped frequency distribution

$\overline{x} = \frac{\Sigma fx}{\Sigma f}$

Example 14.1

The marks of 20 students in a mathematics test were recorded as follows:

Marks (x)	40	51	56	62	70	75	78
Frequency(f)	2	1	3	5	4	3	2

Find the mean.

Solution

x	**f**	**fx**
40	2	80
51	1	51
56	3	168
62	5	310
70	4	280
75	3	225
78	2	156
	$\Sigma f = 20$	$\Sigma fx = 1270$

$\sum f = 20$ and $\sum fx = 1270$

$\therefore$ mean $= \bar{x} = \frac{\sum fx}{f} = \frac{1270}{20} = 63.5$

The median

The median of data is the middle value when all values are arranged in order of the size.

When the number of the items is odd then the median is the item in the middle. If and when the number of items is even, the median is the mean of the two numbers in the middle.

Example 14.2

Find the median of :

a) 6 2 8 12 3 5 20 15 3

b) 9 4 5 6 8 10 2 7

Solution

a) Arranging the numbers in ascending order we have 2 3 3 5 6 8 12 15 20. These are 9 numbers, and the fifth is the middle of them thus the median is that number. Alternative formula:

Median $= \frac{x_{n+1}}{2}$ where n = number of items

So for our case we have the median = 6

b) Arranging the numbers in ascending order we have 2 4 5 6 7 8 9 10. They are 8 numbers. The median is the mean of the two middle numbers or simply the 4^{th} and the 5^{th} number.

We have $\frac{6+7}{2} = 6.5$

In general we have the formula $Me = x_{\frac{n}{2}} + x_{\frac{n}{2}+1}$ or simply the $\left(\frac{n+1}{2}\right)^{th}$ value.

Example 14.3

The table below shows the marks of 62 students in a test. Find the median.

marks	Number of students
40	2
41	4
42	6
43	9
44	10
45	12
46	8
47	7
48	2
49	1
50	1

Solution

Using the cumulative frequencies

Marks	Cummulative number
40	2
≤41	6
≤42	12
≤43	21
≤44	31
≤45	43
≤46	51
≤47	58
≤48	60
≤49	61
≤50	62

The total number of students is 62. In this case n = 62. Thus the median value is $\left(\frac{n+1}{2}\right)^{th}$ value $= \left(\frac{62+1}{2}\right)^{th} = 31.5^{th}$ value.

The median mark is the 31.5^{th} value which is the mean of 31 and 32. Starting with the lowest mark of 40 and move up the frequencies until you reach the 31^{st} and the 32^{nd} share in the distribution.

2+4+6+9+10=31 so the 31^{st} student obtained 44 marks and the 32^{nd} student obtained 45 marks

So the median is $\frac{44+45}{2} = 44.5$. Alternatively. by using the formula

Median $= \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{31} + x_{32}}{2} = \frac{44+45}{2} = 44.5$

The mode

The mode of a set of data is the value of the higher frequency in the distribution of marks

In Example 14.3 the mode is 45 because it has the highest frequency, 12.

Grouped data

Grouped data is commonly used in continuous distribution data that takes any value in a given range is called **continuous data**.

Such data has values which are only approximations, such as height, weight, mass, time, age and temperature.

The frequency distribution below shows the lengths of metal rods measured to the nearest millimetre.

Length	20–24	25–29	30–34	35–39	40–44
Frequency	5	9	13	11	6

The interval 25–29 means that the length is equal to or greater than 24.5 mm and than 29.5 mm, written as 24.5 mm< l <29.5 mm.

Class boundary	**Class width**
19.5 – 24.5	5
24.5 – 29.5	5
29.5 – 34.5	5
34.5 – 39.5	5
39.5 – 44.5	5

The mean of grouped data

In order to find the mean of the grouped data:

I Find the mid-point of each interval.

II Multiply the mid-point (x) by the frequency (f) of each interval to find f.x.

III Find the sum denoted by Σfx and divide by Σf to obtain the mean.

Example 14.4

Find the mean of the following distribution

Mass (kg)	10–19	20–29	30–39	40–49	50–59	60–69
Frequency	5	7	15	12	8	3

Solution

Mass(kg)	**Mid-point**	**Frequency**	**f.x**
10–19	14.5	5	72.5
20–29	24.5	7	171.5
30–39	34.5	15	517.5
40–49	44.5	12	534.5
50–59	54.5	8	436.0
60–69	64.5	3	193.5

$\Sigma f = 50$ and $\Sigma fx = 1925$ and so the mean Mean $\frac{\Sigma fx}{\Sigma f} = \frac{1925}{50} = 38.5$kg.

The mean from an assumed mean

When data is grouped in classes of equal width, we use the assumed mean in order to reduce the numerical size of the value of n.

Determine the mid-point of each class interval and the classes of the central value of x which is usually the modal value. This value is referred to as an **assumed value** or **working mean**.

Subtract the assumed mean from each value of x and where necessary divide the difference so obtained by the class width. This process is a new set of values, say y, that is:

$$y = \frac{x - \text{assumed mean}}{\text{class width}}$$

The mean of y denoted as y is given by the equation $\bar{y} = \frac{\Sigma fy}{\Sigma f}$ and the mean of x denoted by $\bar{x}$ is given by the formula $\bar{x}$ = assumed mean + $\bar{y}$.class width

Example 14.5

Calculate the mean of the following data using the assumed mean

Mass(kg)	Frequency
10–19	3
20–29	7
30–39	12
40–49	18
50–59	12
60–69	6

Let assumed mean be 44.5

Solution

Class	mid-point	f	$y = \frac{x - 44.5}{10}$	f. y
10–19	14.5	3	–3	–9
20–29	24.5	7	–2	–14
30–39	34.5	12	–1	–12
40–49	44.5	18	0	0
50–59	54.5	12	1	12
60–69	64.5	6	2	12

Σ fy = –11

Σ f = 58, $\bar{y} = \frac{\Sigma fy}{\Sigma f} = \frac{-11}{58} = -0.19$

$\bar{x}$ = 44.5 + (– 0.19 × 10) = 42.6 The mean of the given data is 42.6.

14.2 Measures of dispersion

Activity 14.2

In groups of three, research on the meaning and types of measures of dispersion. Why are they useful? Discuss your findings with the rest of the class.

A measure of dispersion is the degrees of spread of observation in data. The common measure of dispersion are range inter-quartiles, range and the standard deviation (the square root of the variance)

Range

The range is defined as the difference between the largest value in the set of data and the smallest value in the set of data, $X_L - X_S$

Example 14.6

What is the range of the following data?

4 8 1 6 6 2 9 3 6 9

Solution

The largest score (X_L) is 9; the smallest score (X_S) is 1;

the range is $X_L - X_S = 9 - 1 = 8$

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:
  For example 1 1 1 1 9 and 1 3 5 7 9.

Inter-quartile range

The other measure of dispersion is the difference between two percentiles, usually the 25^{th} and the 75^{th} percentiles. For numerical data arranged in ascending order, the quartiles are values derived from the data which divide the data into four equal parts. If there are n observations, the first quartile (or lower quartile) Q1 is the $\frac{1}{4}(n+1)^{th}$ data, the second quartile Q2 (the median) is the $\frac{1}{2}(n+1)^{th}$ data and the third quartile (or upper quartile) Q3 is the $\frac{3}{4}(n+1)^{th}$ data. When the $\frac{1}{4}(n+1)^{th}$ is not a whole number, it is sometimes thought necessary to take the (weighted) average of two observations, as is done for the median. However, unless n is very small, an observation that is nearest will normally suffice.

The inter-quartile range is the central 50% of a distribution when it is arranged in order of size. It is given by the formula $Q_3 - Q_1$ where Q_1 is the lower quartile and Q_3 the upper quartile.

Semi inter-quartile range

The semi-interquartile range (or SIR) is defined as the difference of the first and third quartiles divided by two

- The first quartile is the 25th percentile
- The third quartile is the 75th percentile
- SIR = $\frac{(Q_3 - Q_1)}{2}$

The semi inter-quartile range also called the quartile deviation is the half inter-quartile. That is $\frac{Q_3 - Q1}{2}$.

It is not always true that the quartile deviation is $Q_3 - Q_1$ or $Q_2 - Q_1$

Example 14.7

For the data below

5, 3, 7, 6, 3, 4, 2, 7, 5, 6, 5, 4, 9, 8, 3, 6, 5

Find the

(a) first (lower) quartile

(b) second quartile (the median)

(c) third (upper) quartile

(d) inter-quartile range (IQR)

(e) quartile deviation.

Solution

The ordered data set is:

2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9

There are 17 data, so n=17

a) The first quartile is $\frac{1}{4}(17+1)^{th}$ data = $(4.5)^{th}$ data= $\frac{3-4}{2} = 3.5$

b) The second quartile is $\frac{1}{2}(17+1)^{th}$ data = 9^{th} data = 5

c) The upper quartile is $\frac{3}{4}(17+1)^{th}$ data = $(13.5)^{th}$ data $= \frac{6-7}{2} = 6.5$

d) IQR = $Q_3 - Q_1 = 6.5 - 3.5 = 3$

e) The quartile deviation is $\frac{Q_3 - Q1}{2} = \frac{6.5 - 3.5}{2} = \frac{3}{2} = 1.5$

Variability of data

Each of these sets of numbers has a mean of 7 but the spread of each set is different:

(a) 7, 7, 7, 7, 7

(b) 4, 6, 6.5, 7.2, 11.3

(c) –193, –46, 28, 69, 177

There is no variability in set (a), but the numbers in set (c) are obviously much more spread out than those in set (b).

There are various ways of measuring the variability or spread of a distribution, two of which are described here.

The range is based entirely on the extreme values of the distribution.

In (a) the range = 7– 7 = 0

In (b) the range =11.3 – 4=7.3

In (c) the range =177 – (–193) =370

Note that there are also ranges based on particular observations within the data and these are **percentile** and **quartile ranges**.

The standard deviation and the variance

The standard deviation, s , is a very important and useful measure of spread. It gives a measure of the deviations of the readings from the mean, $\overline{x}$. It is calculated using all the values in the distribution.

To calculate standard deviation, s:

(i) For each reading of x calculate $x - \overline{x}$, its deviation from the mean

(ii) Square this deviation to give $(x - \overline{x})^2$ and note that, irrespective of whether the deviation was positive or negative, this is now positive

(iii) Find $\sum(x - \overline{x})^2$, the sum of all these values,

(iv) Find the average by dividing the sum by n, the number of readings; this gives $\frac{\sum (x - \overline{x})^2}{2}$ and is known as **variance**

(v) Finally, take the positive square root of the variance to obtain the standard deviation, s.

The standard deviation, s, of a set of n numbers, with mean $\bar{x}$, is given by

$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}}$. Each of the following three sets of numbers has mean 7, i.e. $\bar{x} = 7$

(a) For the set 7, 7, 7, 7, 7 since $x - \bar{x} =$ for every reading, $s = 0$, indicating that there is no deviation from the mean.

(b) For the set 4, 6, 6.5, 7.2, 11.3

$$\sum (x - \bar{x})^2 = (4 - 7)^2 + (6 - 7)^2 + (6.5 - 7)^2 + (7.2 - 7)^2 + (11.3 - 7)^2 = 28.78$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = \sqrt{\frac{28.78}{5}}$$

(c) For the set –193, –46, 28, 69, 177

$$(x - \bar{x})^2 = (-193 - 7)^2 + (-46 - 7)^2 + (28 - 7)^2 + (69 - 7)^2 + (117 - 7)^2$$

$$= 75994$$

$$\sqrt{\frac{\sum(x - \bar{x})^2}{n}} = \sqrt{\frac{75994}{5}} = 123.3$$

Notice that set (c) has a much higher standard deviation than set (b), confirming that it is much more spread about the mean

Remember that
Standard deviation = $\sqrt{\text{variance}}$
Variance = $(\text{standard deviation})^2$

Note

- The standard deviation gives an indication of the lowest and highest values of the data. In most distributions, the bulk of distribution lies within two standard deviations of the mean, i.e. within the interval $\bar{x} \pm 2s$ or $(\bar{x} - 2s, \bar{x} + 2s)$. This helps to give an idea of the spread of the data.
- The units of standard deviation are the same as the units of the data.
- Standard deviations are useful when comparing sets of data; the higher the standard deviation, the greater the variability in the data.

Example 14.8

Two machines A and B are used to pack biscuits. A random sample of ten packets was taken from each machine and the same mass of each packet was measured to the nearest gram and noted. Find the standard deviations of the masses of the packets taken in the sample from each machine. Comment on your answer.

Machine A (mass in g)	196, 198, 198, 199, 200, 200, 201, 201, 202, 205
Machine B (mass in g)	192, 194, 195, 198, 200, 201, 203, 204, 206, 207

Solution

Machine A: $\overline{x} = \frac{\Sigma x}{n} = \frac{2000}{10} = 200$. Machine B: $\overline{x} = \frac{\Sigma x}{n} = \frac{2000}{10} = 200$

Since the mean mass for each machine is 200, $x - \overline{x} = x - 200$

To calculate standard deviation, s; put the data in a table:

Machine A		
x	$x - 200$	$(x - 200)^2$
196	–4	16
198	–2	4
198	–2	4
199	–1	1
200	0	0
200	0	0
201	1	1
201	1	1
202	2	4
205	5	25
		56

$s^2 = \frac{\Sigma(x - 200)^2}{10} = 5.6$

$s = \sqrt{5.6} = 2.37$

Machine A: s.d = 2.37g

Machine B		
x	x – 200	$(x - 200)^2$
192	–8	64
194	–6	36
195	–5	25
198	–2	4
200	0	0
201	1	1
203	3	9
204	4	16
206	6	36
207	7	49
		240

$s^2 = \frac{\Sigma(x - 200)^2}{10} = 24$

$s = \sqrt{24} = 4.980$ g

Machine B: s.d = 4.90 g

Machine A has less variation, indicating that it is more reliable than machine B.

Alternative form of the formula for standard deviation

The formula given above is sometimes difficult to use especially when $\overline{x}$ is not an integer; so an alternative form is often used. This is derived as follows:

$s^2 = \frac{1}{n}\Sigma(x - \overline{x})^2 = \frac{1}{n}\Sigma(x^2 - 2\overline{x}\,x + \overline{x}^2) = \frac{1}{n}(\Sigma x^2 - 2\overline{x}\,\Sigma x + \Sigma\overline{x}^2)$

since, $\frac{\Sigma x}{n} = \overline{x}$

$s^2 = \frac{\Sigma x^2}{n} - 2\overline{x}\,\frac{\Sigma x}{n} + \frac{n\overline{x}^2}{n} = \frac{\Sigma x^2}{n} - 2\overline{x}\,(\overline{x}) + \overline{x}^2 = \frac{\Sigma x^2}{n} - 2\overline{x}^2 + \overline{x}^2 = \frac{\Sigma x^2}{n} - \overline{x}^2$

$s^2 = \frac{\Sigma x^2}{n} - \overline{x}^2$

$s = \sqrt{\frac{\Sigma x^2}{n} - \overline{x}^2}$

Task 14.1

1. For each of the following sets of numbers, calculate the mean and the standard deviation.
   (a) 2, 4, 5, 6, 8
   (b) 6, 8, 9, 11
   (c) 11, 14, 17, 23, 29
2. For each of the following sets of numbers, calculate the mean and standard deviation using one of the methods of the formula for the standard deviation.
   (a) 5, 13, 7, 9, 16, 15
   (b) 4.6, 2.7, 3.1, 0.5, 6.2
   (c) 200, 203, 206, 207, 209

14.3 Coefficient of variation

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ.

$C_v = \frac{\sigma}{\mu}$

It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a ratio scale, as these are the measurements that can only take non-negative values.

The coefficient of variation (C.V) unlike the previous measures we have studied is a relative measure of dispersion. It is expressed as a percentage rather than in terms of the unit of the particular data. It is useful when comparing the variable of two or more batches of data. Those are expressed in different units of measurement.

$C.V = \frac{\delta}{\bar{x}} \times 100$ where δ is the standard deviation and $\bar{x}$ is the mean.

For example, given that $\delta = 6.26$ and $\bar{x} = 20$ then $CV = \frac{6.26}{20} \times 100 = 31.3\%$

That is for this sample the relative size of the average spread around the mean is 31.3%. The C.V is also very useful when comparing two or more sets of data which are measured in the same units.

The following are more examples.

A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100:

The CV is $100\% \times \frac{0}{100} = 0\%$

A data set of [90, 100, 110] has more variability. Its standard deviation is 8.16 and its average is 100:

The CV is $100\% \times \frac{8.16}{100} = 8.16\%$

A data set of [1, 5, 6, 8, 10, 40, 65, 88] has more variability again. Its standard deviation is 30.78 and its average is 27.875:

The CV is $100\% \times \frac{30.78}{27.875} = 110.4\%$

14.4 Application

Activity 14.3

Carry out research to find out real life applications on measures of dispersion. Discuss your findings with the rest of the class.

Example 14.9

Sona, Karina, Omar, Mustafa and Amie earned scores of 6, 7, 3, 7 and 2 respectively on a standardized test.

Find the mean deviation and standard deviation of their scores.

Solution:

Mean deviation: We must first find the mean of the data set. The mean is $\frac{6+7+3+7+2}{5} = 5$. Then the mean deviation is calculated by

$$\frac{|6-5|+|7-5|+|3-5|+|7-5|+|2-5|}{5} = \frac{1+2+2+2+3}{5} = \frac{10}{5} = 2.$$

The mean deviation is 2.

Standard deviation is

$$= \sqrt{\frac{[(6-5)^2 + (7-5)^2 + (3-5)^2 + (7-5)^2 + (2-5)^2]}{5}}$$

$$= \frac{1+4+4+4+9}{5}$$

$$= \frac{\sqrt{22}}{5} \approx 0.938$$

The standard deviation is approximately 0.938

Task 14.2

1. The marks of a class in a test are as follows

   52, 45, 25, 75, 63, 86, 72, 85, 55, 65, 70, 82, 90, 48, 68, 86, 65, 64, 78, 75, 32, 42. Find the inter-quartile range.
2. Find the standard deviation of the data set 5, 10, 15, 20, 25, 30, 35, 40, 45, 50.
3. Calculate the variance and the standard deviation for the following values: 1, 3, 5, 6, 6, 8, 9, and 10.
4. Ten different teams played football during one season. At the end of the season the top goal scorers from each team scored the following number of goals:

   10, 5, 18, 2, X, 4, 10, 15, 11, 4

   If the mean number of goals scored is 9, what is the:

   a) value of X?
   b) mode?
   c) median?
   d) range?
   e) mean deviation?
   f) standard deviation?
   g) 50^{th} percentile?
   h) percentile of the goal scorer with 11 goals scored?

Summary

1. Mean: is the sum of data values divided by the number of values in the data
2. Mode: is the value that occurs most often in the data.
3. Median: is the middle value when the data is arranged in order of magnitude
4. When data are grouped in classes of equal width, we use the **assumed mean** in order to reduce the numerical size of the value of n.
5. A **measure of dispersion** is the degree of spread of observation in data. The common measures of dispersion are inter-quartiles, range and the standard deviation (the square root of the variance).
6. Range: is the numerical difference between the largest value and the least value of data
7. The **inter-quartile range** is the central 50% of a distribution when it is arranged in order of size.
8. The **coefficient of variation** (C.V) is a relative measure of dispersion. It is expressed as a percentage rather than in terms of the unit of the particular data. It is useful when comparing the variable of two or more batches of data.

   C.V. $= \frac{\delta}{\overline{x}} \times 100$ where δ is the standard deviation and $\overline{x}$ is the mean.