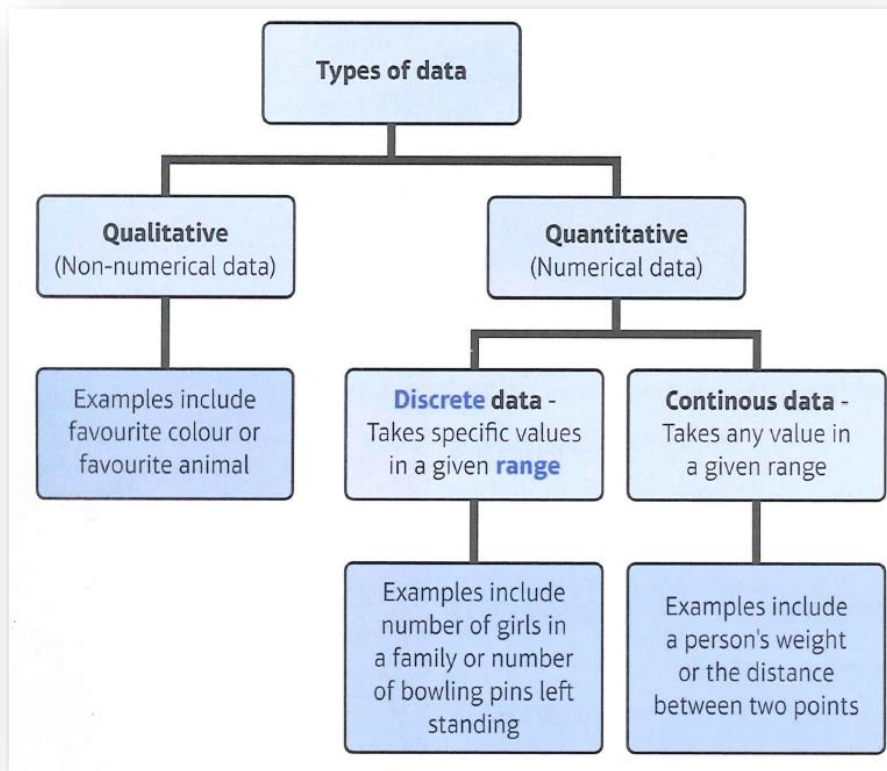


Types of Data:

Qualitative: Non-numeric. Example colours, names, any descriptive data

Quantitative: Numeric data. Any answer that involves numbers

Discrete: When only specific values can be chosen

Continuous: In a given range any value is possible

Example:


State whether each of the following variables is discrete or continuous

- 1) Time taken to finish an exam
- 2) Number of children in a family
- 3) Length of a pencil
- 4) Number of coins in a bag


Frequency (f):

When we have a big number of discrete data, we can use frequency tables. For example the following table shows that we have 25 families with one child and 21 families with 2 children.

Number of children	Frequency
1	25
2	21
3	18
4	9
5	2


Cumulative frequency

x	Number of students, f	Cumulative frequency
35	3	3
36	17	20
37	29	49
38	34	83
39	12	95



$3 + 17 = 20$

$20 + 29 = 49$

Classes

Sometimes we group data together in what we call classes

Table 1	
Marks	Frequency
50-60	5
60-70	8
70-80	15
80-90	6

Table 2	
Marks	Frequency
50-60	5
61-70	8
71-80	15
81-90	6

<i>Classes</i>

In Table 1: It seems that the classes overlap yet this is only a shortcut of saying $50 \leq M < 60$, so a value of 60 is actually in the second class not the first

Class boundaries are 50 and 60

Class width is $60 - 50 = 10$

$$\text{Midpoint} = \frac{50+60}{2} = 55$$

In Table 2: There are gaps between the classes

Class boundaries are 49.5 and 60.5

Class width is $60.5 - 49.5 = 11$. We divide the gap between the classes

$$\text{Midpoint} = \frac{49.5+60.5}{2} = 55$$

Central tendency

Mode/Modal Class: The most repeated value or class

Example 1: 2, (5) 6, (5) 9, 6, (5) 4, 1, 0 mode = 5

Example 2:

Score	Frequency
5	2
(6)	3 ✓
7	2
8	2
9	1
10	1

mode = 6

Example 3:

Class (Rs.)	Tally Marks	Frequency Students
20 - 30		5
30 - 40		8
40 - 50		9
(50 - 60)		10 ✓
60 - 70		6
70 - 80		2
Total		40

modal class = 50 – 60

Median: The middle value when the data values are *put in order*

N.B: To know the mid value use $\frac{n+1}{2}$ where n is the number of values. If the answer is a whole number like 4 then this is the number of the value we want (4th value). If the answer has a half like 6.5 then you need the average of the one before and the one after (Average of 6th and 7th)

Example 1: Some discrete data

Odd number: 5, 7, 1, 3, 6, 8, 2 (7 values)

Put in order: 1, 2, 3, 5, 6, 7, 8

The required value is at $\frac{7+1}{2} = 4^{\text{th}}$ position so Median = 5

Even number: 5, 7, 1, 3, 6, 8, 2, 10 (8 values)

Put in order: 1, 2, 3, 5, 6, 7, 8, 10

The required value is at $\frac{8+1}{2} = 4.5$ so the Median is the average of the 4th and 5th values = 5.5

Example 2: A frequency table

Li Wei records the shirt collar size, x , of the male students in his year. The results are shown in the table.

Shirt collar size	15	15.5	16	16.5	17
Frequency	3	17	29	34	12

For these data, find: The median

$$\sum f = 95. \text{ To get the position of the mid value } \frac{95+1}{2} = 48^{\text{th}}$$

Now we notice by looking at the frequency that $3 + 17 = 20$ which is less than what we need and adding the 29 we reach 49 so the 48th value is under 16. Median = 16

Example 3: Classes and a frequency table

The length x mm, to the nearest mm, of a random sample of pine cones is measured. The data is shown below.

Length of pine cone (mm)	Number of pine cones, f	Cumulative frequency
30–31	2	2
32–33	25	27
34–36	30	57
37–39	13	70

- Find the median class
- Find the median value

The first one is easy we have $\sum f = 70$ so $\frac{70}{2} = 35$ (For continuous data we don't use $\frac{n+1}{2}$) By examining the cumulative frequency, the 35th value is in the class 34 – 36. Median class is 34 – 36

The second one is more precise but we need to know a method called interpolation

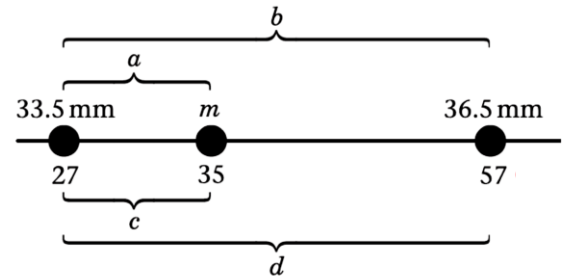
In the upper part we put the class boundaries, and m for Median

In the lower part the cumulative frequency

and the position we are looking for. Then use the famous formula

$$\frac{a}{b} = \frac{c}{d} \quad \frac{m-33.5}{35-27} = \frac{36.5-33.5}{57-27}$$

$$\text{Median (m)} = 34.3$$



Mean: The sum of all observations divided by the number of observations

N.B: If we have observations x_1, x_2, x_3, \dots Then “sum of” is written as $\sum x$ and mean as $(\bar{x}$ for sample) or $(\mu$ for population)

- For discrete data

$$Mean = \frac{\sum x}{n}$$

Example 1: Discrete data

A child at a junior school records the maximum temperature, in °C, for seven days at his school. The results are given below.

15.7 16.1 16.2 47.6 17.4 18.6 16.7

Find the mean.

$$Mean = \frac{\sum x}{n} = \frac{15.7+16.1+16.2+47.6+17.4+18.6+16.7}{7} = 21.2$$

- For a table of data with given frequency our rule is

$$Mean = \frac{\sum fx}{\sum f}$$

Example 2: A frequency table

Li Wei records the shirt collar size, x , of the male students in his year. The results are shown in the table.

Shirt collar size	15	15.5	16	16.5	17
Frequency	3	17	29	34	12

For these data, find: The mean

$$Mean = \frac{\sum xf}{\sum f} = \frac{15 \times 3 + 15.5 \times 17 + 16 \times 29 + 16.5 \times 34 + 17 \times 12}{95} = 16.2$$

- For a table of data with given classes and frequency our rule is the same as the previous one but x is the mid of the class.

Example 3: Classes and a frequency table

The length, x mm, to the nearest mm, of a random sample of pine cones is measured. The data are shown in the table. Find the mean

Length of pine cone (mm)	30–31	32–33	34–36	37–39
Frequency	2	25	30	13

Hint: To get x which is the mid of the class no need to take the real class width from 29.5 to 31.5 as the mid will stay the same if we say

$$\frac{30+31}{2} = 30.5$$

A good way of doing this is in a table

Class	x (mid of class)	f	$x.f$
30 - 31	30.5	2	61
32 - 33	32.5	25	812.5
34 - 36	35	30	1050
37 -39	38	13	494
Σ		70	2417.5

$$\text{Mean} = \frac{\Sigma xf}{\Sigma f} = \frac{2417.5}{70} = 34.54$$

Combined means:

If set A has size of n_1 and mean \bar{x}_1 and set B has size n_2 and mean \bar{x}_2 then the mean of the combined set of A and B is

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Example:

The mean of a sample of 25 observations is 6.4. The mean of a second sample of 30 observations is 7.2. Calculate the mean of all 55 observations

$$\bar{x} = \frac{25 \times 6.4 + 30 \times 7.2}{25 + 30} =$$

Sometimes you can get the three values mode, median and mean but you have to choose which one of them best describes your data. In general

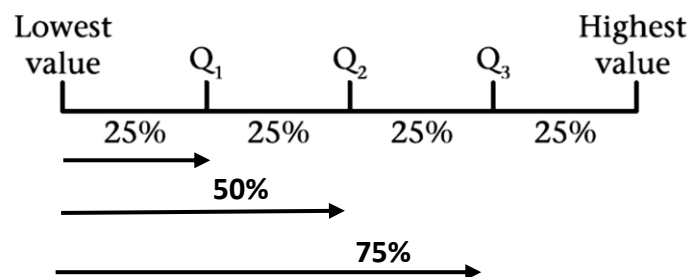
- **Mode** This is used when data are qualitative, or when quantitative with either a single mode or two modes (bimodal). There is no mode if each value occurs just once.
- **Median** This is used for quantitative data. It is usually used when there are extreme values, as they do not affect it as much as they affect the mean.
- **Mean** This is used for quantitative data and uses all the pieces of data. It therefore gives a true measure of the data. However, it is affected by extreme values.

Other measures of location: Quartiles and Percentiles:

Q1 (Lower quartile): Is one quarter of the way through the data (25%)

Q2: Is half of the way through the data (50%) This is the median

Q3 (Higher quartile): Is three quarters of the way through the data (75%)



Percentile: Is like quartile but measured out of 100, so 15th percentile is $\frac{15}{100}$ of the way through the data

For Discrete data

Use these rules to find the upper and lower quartiles for **discrete data**.

- To find the lower quartile for discrete data, divide n by 4. If this is a whole number, the lower quartile is halfway between this data point and the one above. If it is not a whole number, round **up** and pick this data point.
- To find the upper quartile for discrete data, find $\frac{3}{4}$ of n . If this is a whole number, the upper quartile is halfway between this data point and the one above. If it is not a whole number, round **up** and pick this data point.

The data below shows how far (in kilometres) 20 employees live from their place of work.

1	3	3	3	4	4	6	7	7	7
9	10	11	11	12	13	14	16	18	23

Find the median and quartiles for these data.

Median Q_2 : We have 20 values so median lies at $\frac{20+1}{2} = 10.5$

(between 10th and 11th) so between 7 and 9, $\frac{7+9}{2} = 8$

LQ Q_1 : $\frac{20}{4} = 5$ (A whole number so between 5th and 6th) = 4

HQ Q_3 : $\frac{3}{4} \times 20 = 15$ (A whole number so between 15th and 16th) = 12.5

Example 2: Grouped data (Classes) with frequency table

For grouped data we use interpolation.

The length of time (to the nearest minute) spent on the internet each evening by a group of students is shown in the table.

Time spent on the internet (minutes)	30–31	32–33	34–36	37–39
Frequency	2	25	30	13

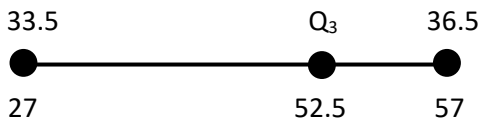
- a Find an estimate for the upper quartile. b Find an estimate for the 10th percentile.

a. Upper quartile Q_3 : at $\frac{3}{4} \times 70 = 52.5^{\text{th}}$

So in the class 34 – 36

Now we use interpolation

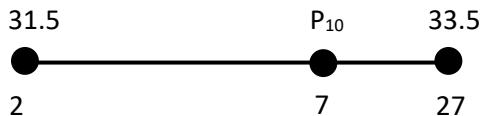
Class	f	c.f
30 - 31	2	2
32 - 33	25	27
34 - 36	30	57
37 - 39	13	70



$$\frac{Q_3 - 33.5}{36.5 - 33.5} = \frac{52.5 - 27}{57 - 27} \quad Q_3 = 36.05$$

b. P_{10} (10th percentile) at $\frac{10}{100} \times 70 = 7^{\text{th}}$

So in class 32 – 33. We use interpolation



$$\frac{P_{10} - 31.5}{33.5 - 31.5} = \frac{7 - 2}{27 - 2} \quad P_{10} = 31.9$$

Measures of spread:

Range: The highest value – the lowest value

Interquartile Range: Upper quartile – Lower quartile = $Q_3 - Q_1$

Interpercentile Range: Difference between the values of two given percentiles

Variance (σ^2):

- For Discrete Data

$$\begin{aligned}\text{Variance} &= \frac{\sum(x-\bar{x})^2}{n} \\ &= \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2\end{aligned}$$

For simplicity we define $S_{xx} = \sum(x - \bar{x})^2$

Which is the same as $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$

$$\text{Variance} = \frac{S_{xx}}{n}$$

Standard deviation (σ): Is the square root of the variance

Example 1: Discrete data

The marks gained in a test by seven randomly selected students are:

3 4 6 2 8 8 5

Find the variance and standard deviation of the marks of the seven students.

$$\sum x = 3 + 4 + 6 + 2 + 8 + 8 + 5 = 36$$

$$\sum x^2 = 9 + 16 + 36 + 4 + 64 + 64 + 25 = 218$$

$$\text{Variance} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{218}{7} - \left(\frac{36}{7}\right)^2 = 4.69$$

$$\text{Standard deviation} = \sqrt{4.69} = 2.17$$

- For a table of data with frequency whether grouping in classes or not

$$\begin{aligned} \text{Variance} &= \frac{\sum f(x-\bar{x})^2}{\sum f} \\ &= \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 \end{aligned}$$

Example 2: A frequency table (With no grouping)

Shamsa records the time spent out of school during the lunch hour to the nearest minute, x , of the students in her year.

Time spent out of school, x (min)	35	36	37	38
Frequency	3	17	29	34

The results are shown in the table.

Calculate the standard deviation of the time spent out of school.

$$\Sigma f = 3 + 17 + 29 + 34 = 83$$

$$\Sigma fx = 3 \times 35 + 17 \times 36 + 29 \times 37 + 34 \times 38 = 3082$$

$$\Sigma fx^2 = 3 \times 35^2 + 17 \times 36^2 + 29 \times 37^2 + 34 \times 38^2 = 114504$$

$$\text{Variance} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f} \right)^2 = \frac{114504}{83} - \left(\frac{3082}{83} \right)^2 = 0.74147$$

$$\text{Standard deviation} = \sqrt{0.74147} = 0.861$$

Example 3: A frequency table (With classes)

Akira recorded the length, in minutes, of each phone call she made for a month. The data are summarised in the table below.

Length of phone call, l (min)	$0 < l \leq 5$	$5 < l \leq 10$	$10 < l \leq 15$	$15 < l \leq 20$	$20 < l \leq 60$	$60 < l \leq 70$
Frequency	4	15	5	2	0	1

Calculate an estimate of the standard deviation of the length of Akira's phone calls.

Class	f	x	fx	fx^2
$0 < l \leq 5$	4	2.5	$4 \times 2.5 = 10$	$4 \times 2.5^2 = 25$
$5 < l \leq 10$	15	7.5	112.5	843.75
$10 < l \leq 15$	5	12.5	62.5	781.25
$15 < l \leq 20$	2	17.5	35	612.5
$20 < l \leq 60$	0	40	0	0
$60 < l \leq 70$	1	65	65	4225
Σ	27		285	6487.5

$$\text{Variance} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f} \right)^2 = \frac{6487.5}{27} - \left(\frac{285}{27} \right)^2 = 128.858$$

$$\text{Standard deviation} = \sqrt{128.858} = 11.4$$

Coding: This is a way of simplifying data to make working easier

N.B.: *When data is coded, different statistics change in different ways. For instance, the mean always changes but the standard deviation does not change by adding or subtracting numbers while it does by multiplying or dividing.*

If data are coded using the formula $y = \frac{x-a}{b}$ where x is the original data and y , the coded

Then

$$\bar{x} = b\bar{y} + a$$

$$\sigma_x = b\sigma_y$$

Example1: (Discrete data)

a. Find the mean and the standard deviation of the following lengths

x (mm)	110	120	130	140	150
----------	-----	-----	-----	-----	-----

$$\sum x = 110 + 120 + 130 + 140 + 150 = 650$$

$$\sum x^2 = 110^2 + 120^2 + 130^2 + 140^2 + 150^2 = 85500$$

$$\text{Mean } \bar{x} = \frac{\sum x}{n} = \frac{650}{5} = 130$$

$$\text{Variance} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{85500}{5} - \left(\frac{650}{5}\right)^2 = 200$$

$$\text{Standard deviation} = \sqrt{200} = 14.1$$

b. Use the given coding to find the coded mean and coded standard deviation, then use them to find the originals for the above data

i. $y = x - 100$ ii. $y = \frac{x}{10}$ iii. $y = \frac{x - 100}{10}$

i.

x	110	120	130	140	150	Σ
$y = x - 100$	10	20	30	40	50	150
y^2	100	400	900	1600	2500	5500

$$\bar{y} = \frac{\Sigma y}{n} = \frac{150}{5} = 30$$

$$\bar{x} = b\bar{y} + a \quad \bar{x} = (1)30 + 100 = 130$$

$$y = \frac{x - 100}{1}$$

a
b

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{5500}{5} - \left(\frac{150}{5}\right)^2 = 200$$

$$\sigma_x = \sigma_y = \sqrt{200} = 14.1$$

Notice that: Adding or subtracting numbers does not change the standard deviation

ii.

x	110	120	130	140	150	Σ
$y = \frac{x}{10}$	11	12	13	14	15	65
y^2	121	144	169	196	225	855

$$\bar{y} = \frac{\Sigma y}{n} = \frac{65}{5} = 13$$

$$\bar{x} = b\bar{y} + a \quad \bar{x} = (10)13 + 0 = 130$$

$$y = \frac{x - 0}{10}$$

a
b

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{855}{5} - \left(\frac{65}{5}\right)^2 = 2$$

$$\sigma_y = \sqrt{2} = 1.41 \quad \sigma_x = b\sigma_y \quad \sigma_x = 1.41 \times 10 = 14.1$$

iii.

x	110	120	130	140	150	Σ
$y = \frac{x - 100}{10}$	1	2	3	4	5	15
y^2	1	4	9	16	25	55

$$\bar{y} = \frac{\Sigma y}{n} = \frac{15}{5} = 3$$

$$\bar{x} = b\bar{y} + a \quad \bar{x} = (10)3 + 100 = 130$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{55}{5} - \left(\frac{15}{5}\right)^2 = 2$$

$$\sigma_y = \sqrt{2} = 1.41 \quad \sigma_x = b\sigma_y \quad \sigma_x = 1.41 \times 10 = 14.1$$

$$y = \frac{x - 100}{10}$$

Example2: (Using classes and frequency)

Akira recorded the length, in minutes, of each phone call she made for a month, as summarised in the table shown. Use the coding $y = \frac{x-7.5}{5}$ to

calculate an estimate for

a. The mean

b. The standard deviation

Length of phone call	Number of occasions
$0 < l \leq 5$	4
$5 < l \leq 10$	15
$10 < l \leq 15$	5
$15 < l \leq 20$	2
$20 < l \leq 60$	0
$60 < l \leq 70$	1

Class	f	x	$y = \frac{x-7.5}{5}$	fy	fy ²
$0 < l \leq 5$	4	2.5	-1	-4	4
$5 < l \leq 10$	15	7.5	0	0	0
$10 < l \leq 15$	5	12.5	1	5	5
$15 < l \leq 20$	2	17.5	2	4	8
$20 < l \leq 60$	0	40	6.5	0	0
$60 < l \leq 70$	1	65	11.5	11.5	132.25
Σ	27			16.5	149.25

$$a. \quad \bar{y} = \frac{\Sigma fy}{\Sigma f} = \frac{16.5}{27} = 0.6111$$

$$\bar{x} = b\bar{y} + a \quad \text{Mean} = \bar{x} = 5 \times 0.6111 + 7.5 = 10.6$$

$$b. \quad \sigma_y^2 = \frac{\Sigma fy^2}{\Sigma f} - \left(\frac{\Sigma fy}{\Sigma f}\right)^2 = \frac{149.25}{27} - \left(\frac{16.5}{27}\right)^2 = 5.154$$

$$\sigma_y = \sqrt{5.154} = 2.27 \quad \sigma_x = b\sigma_y \quad \sigma_x = 5 \times 2.27 = 11.35$$