University of South Wales
Prifysgol De Cymru

MS4S08 – Applied Statistics for Data Science

Introduction to Statistics

Dr Penny Holborn
penny.holborn@southwales.ac.uk

University of South Wales
Prifysgol De Cymru

Material

Topics

- Types of data (discreet, continuous, representation of data)
- Measures of average (mean, median, mode, quartiles)
- Measures of spread (range, IQR, standard deviation, skewness)

University of South Wales
Prifysgol De Cymru

Introduction

We frequently open newspapers and see statements such as the following:

- The average fuel economy of new cars sold in the UK is 37.5 mpg.
- The average starting salary for graduates with degrees in Sports Science is £17,375 per year.
- 48% of voters approve of the job the Prime Minister is doing.
- The median family income is £22,788.

These numerical facts or measures are called ***statistics***.

The stages of a typical statistical investigation are as follows…

University of South Wales
Prifysgol De Cymru

Introduction

Statistics provides a bases for investigations in many fields of knowledge such as:

- Social, physical and biological sciences
- Engineering
- Education
- Business
- Medicine
- Law
- Sport

The stages of a typical statistical investigation are as follows…

University of South Wales
Prifysgol De Cymru

Introduction

1. **The collection of relevant and appropriate data**. Valid conclusions can only be drawn from correctly collected data.

2. **The tabulation and summary of the data using suitable graphical or numerical techniques.** This provides insights into the underlying structure of a data set. Appending an appropriate summary to a formal report is more helpful than appending pages of raw data.

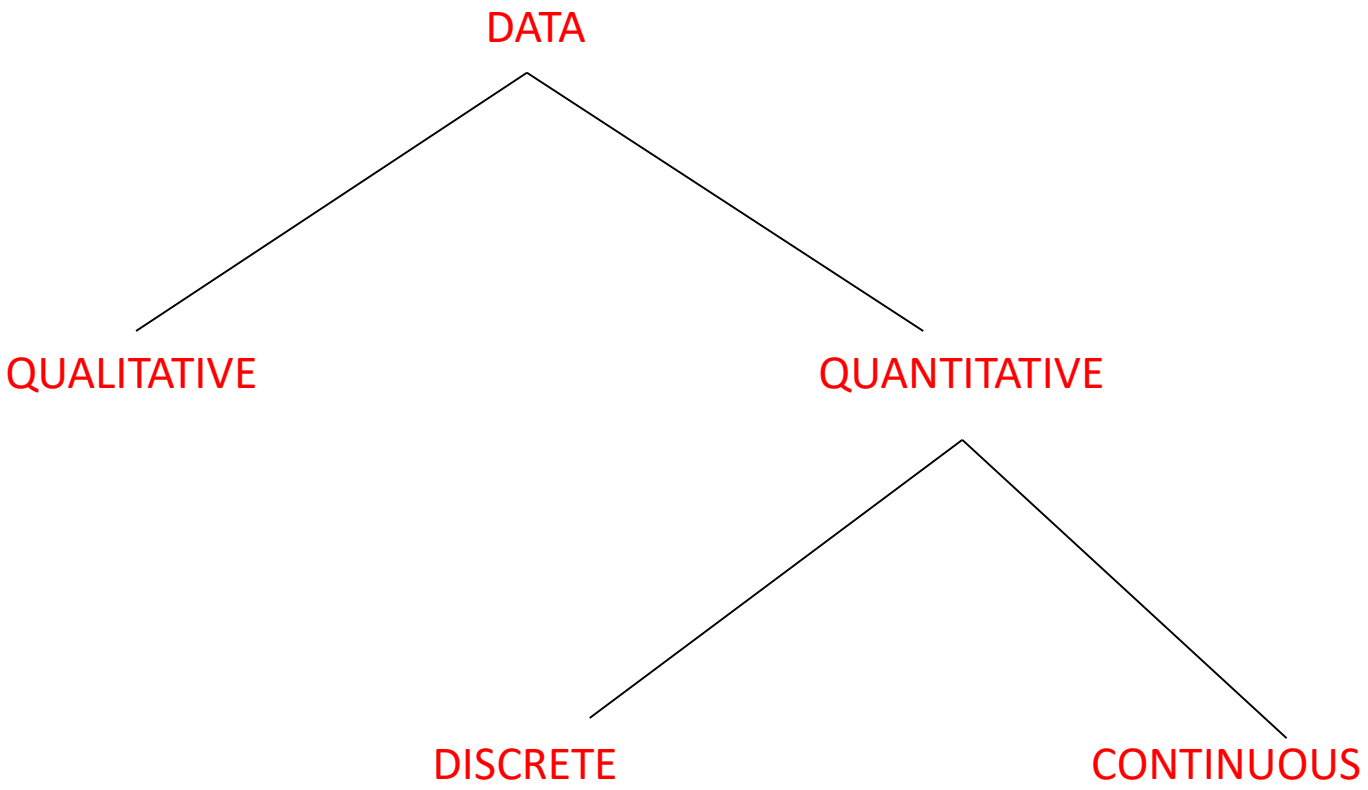
3. **Data analysis.** This involves selecting the appropriate analysis and performing the necessary calculations. Computer packages are very useful when faced by detailed calculations.

4. **Interpreting the results.** This usually involves drawing the correct conclusions from the analysis and stating the conclusions clearly.

University of South Wales
Prifysgol De Cymru

Types of data

We may classify data as follows:



University of South Wales
Prifysgol De Cymru

Types of data

Qualitative data can be placed into categories with respect to some attribute
Examples include:

- Gender
- Location
- Eye colour
- Favorite colour

Quantitative data have numerical values
Examples include:

- Age
- Height
- Exam mark
- Blood pressure

University of South Wales
Prifysgol De Cymru

Types of data

Quantitative data can be further subdivided into two types:

Discrete variables can take only distinct, isolated values, typically whole numbers. They often arise from a counting process:

- Shoe size
- Year of study
- Likert scale (1..5)

Continuous variables can take any value within a sensible range and are normally the result of a measurement process:

- Height
- Temperature
- Distance

University of South Wales
Prifysgol De Cymru

Types of data

In practice, continuous variables may be effectively discrete because of rounding.

Some data can be classified as **qualitative *or* quantitative**.

For example, height is quantitative if measured in centimetres but qualitative if we were to choose to classify individuals as short, medium or tall, according to some specified criterion.

In the next sections, we shall look at ways of representing and summarising data graphically and in tabular form.

University of South Wales
Prifysgol De Cymru

Representation of discrete data

As an aid to summarising discrete quantitative data, we could construct a **frequency table** then use it to produce a bar chart.

Example

The numbers of errors in computer programs written by each of 36 students is as follows:

3	2	1	4	0	2	0	1	5
2	0	0	3	1	1	3	2	4
1	3	2	4	0	2	1	2	1
2	1	0	1	2	3	1	2	4

University of South Wales
Prifysgol De Cymru

Representation of discrete data

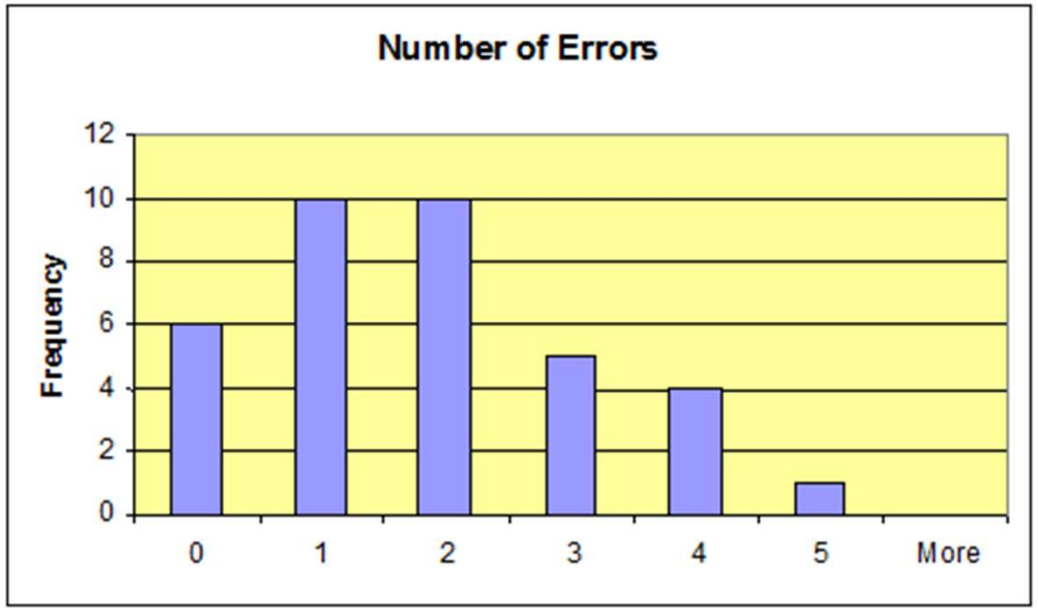
We form the table as follows:

Number of errors	Tally	Frequency
0	𝍸 1	6
1	𝍸 𝍸	10
2	𝍸 𝍸	10
3	𝍸	5
4	1111	4
5	1	1
Total		36

University of South Wales
Prifysgol De Cymru

Representation of discrete data

The corresponding bar chart is:



Note that the bars must be in the given order for quantitative data.

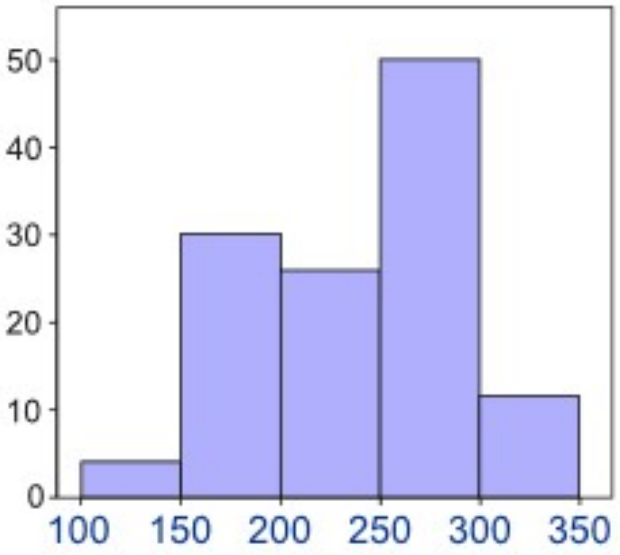
University of South Wales
Prifysgol De Cymru

Representation of continuous data

The most frequently used summaries for continuous data are:

- Grouped frequency tables;
- Histograms;
- Stem-and-leaf diagrams.

Time (minutes)	Frequency
60 up to 90	15
90 up to 120	12
120 up to 150	7
150 up to 180	11
180 up to 210	5



Stem	Leaf
13	6, 9, 9
14	2, 3, 3, 3, 3, 4
14	6, 7, 7, 8, 9
15	1, 3, 4
15	6, 7
16	2, 4

Key:
13 | 6 means 136

University of South Wales
Prifysgol De Cymru

Representation of continuous data

Example

Suppose the weights, to the nearest gram, of 35 packages are:

67	72	75	60	59	62	72
67	75	63	73	67	63	73
68	78	64	74	69	78	69
74	70	74	73	70	73	70
65	71	75	65	76	71	77

To construct a **grouped frequency table**, we place the data into classes, or intervals.

University of South Wales
Prifysgol De Cymru

Representation of continuous data

In our example, we have:

Smallest value = 59
Largest value = 78
Range = 78 – 59 = 19

We divide the range into a number of *class intervals*, all of the same size.

We would normally choose between five and fifteen such intervals.

One possibility is:

59 – 61
62 – 64
65 – 67 etc.

(That is, there are three rounded values on each interval.)

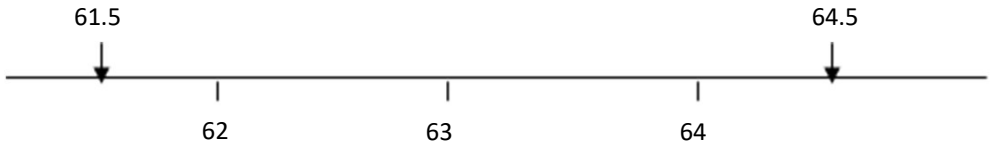
University of South Wales
Prifysgol De Cymru

Representation of continuous data

The **class midpoint** is half the sum of the endpoints.
e.g. For interval 59 to 61, the class mid point is (59 + 61)/2 = 60;

The **class width** is the difference between successive midpoints.
e.g. 63-60 = 3;

Class boundaries require a little thought.



The data has been given to the nearest whole number. The actual boundary is 61.5 since 61.49 would be recorded as 61 and 61.51 would be recorded as 62.

University of South Wales
Prifysgol De Cymru

Representation of continuous data

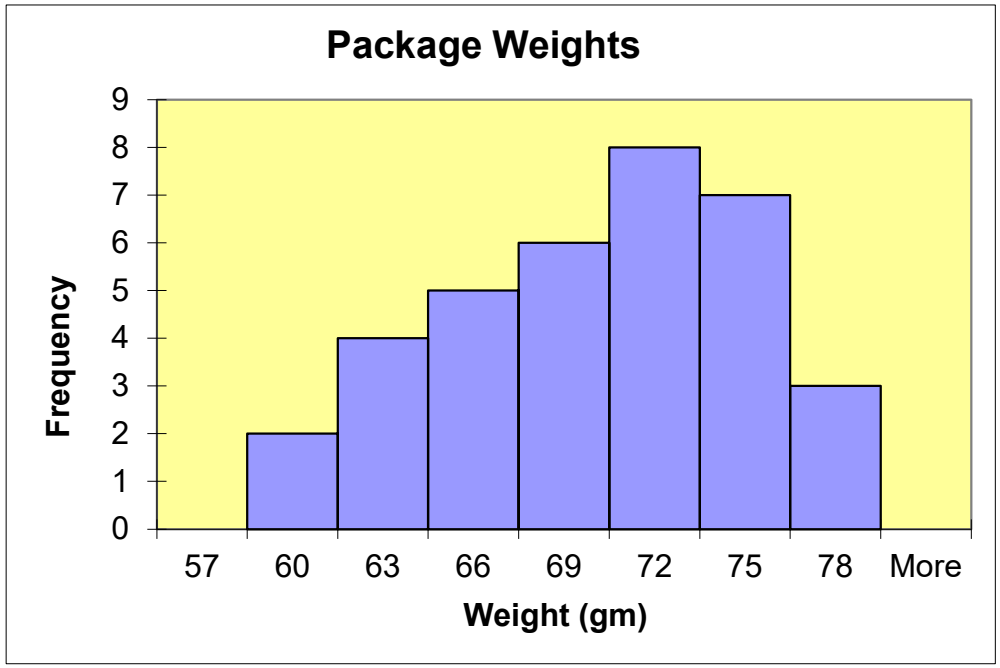
The grouped frequency table is thus:

Class interval	Class midpoint	Tally	Frequency
59 - 61	60	11	2
62 - 64	63	1111	4
65 - 67	66	~~1111~~	5
68 - 70	69	~~1111~~ 1	6
71 - 73	72	~~1111~~ 111	8
74 - 76	75	~~1111~~ 11	7
77 - 79	78	111	3
Total:			35

University of South Wales
Prifysgol De Cymru

Representation of continuous data

A **histogram** drawn from such a table is:



Note that this differs from a bar chart in that there are **no gaps** between the blocks representing the frequencies, reflecting that the data is continuous.

University of South Wales
Prifysgol De Cymru

Measures of average

As an alternative to summarising data by graphical methods, we could try to do so by numerical methods.

If we could find one measure which tells us about the **average** of the data set and another which tells us something about **how spread out** the data set is, then those two measures taken together give us useful information.

There are number of different measures of average.

We shall consider some of the more frequently used including:

- Mean
- Median
- Mode

University of South Wales
Prifysgol De Cymru

Arithmetic mean

Suppose we have a set of *n* numbers denoted by $x_1, x_2, \ldots . x_n$.

The **arithmetic mean** (or the **mean)** is defined by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad = \quad \frac{\textit{Data Total}}{\textit{Total Frequency}}$$

We can write: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or $\bar{x} = \frac{1}{n}\sum x_i$ or even $\bar{x} = \frac{1}{n}\sum x$

University of South Wales
Prifysgol De Cymru

Arithmetic mean

Example

Suppose we have 5 values, 8, 3, 7, 12, 11;

Then the mean is $\bar{x} = \frac{8+3+7+12+11}{5} = 8.2$

Note that the mean does not have to equal one of the data values.

University of South Wales
Prifysgol De Cymru

Arithmetic mean

If we have already formed a frequency distribution from a data set, we can make use of it as follows:

Suppose the value	x_1	occurs	f_1	times
	x_2	occurs	f_2	times
	:		:	
	.		.	
	x_n	occurs	f_n	times

What is the sum for the whole data set? It equals $f_1x_1 + f_2x_2 + \ldots\ldots f_nx_n$.

How many values are there altogether? There are $f_1 + f_2 + \ldots\ldots f_n$.

University of South Wales
Prifysgol De Cymru

Arithmetic mean

So the mean is $\bar{x} = \dfrac{f_1 x_1 + f_2 x_2 + \ldots\ldots + f_n x_n}{f_1 + f_2 + \ldots\ldots + f_n}$

$$= \frac{\textit{Data Total}}{\textit{Total Frequency}}$$

Or, using sigma notation: $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$

University of South Wales
Prifysgol De Cymru

Arithmetic mean

Example

In 10 matches a football team's goal scoring record is as follows:

0 goals scored in 4 games
1 goal scored in 3 games
3 goals scored in 2 games
5 goals scored in 1 game

Find the mean number of goals scored in the 10 matches.

University of South Wales
Prifysgol De Cymru

Arithmetic mean

We can set out the calculations in a table as follows:

Goals (x)	Games (f)	fx
0	4	0
1	3	3
3	2	6
5	1	5
Total	10	14

Hence $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{14}{10}$ = 1.4 goals per game.

University of South Wales
Prifysgol De Cymru

Arithmetic mean

Suppose now that we are considering an entire *population* of size *N*.

The mean is calculated in precisely the same way, but the notation is slightly different.

We define the **population mean** by: $\mu = \frac{1}{N}\sum_{1}^{N} x_i$

Again, the same abbreviated versions of the notation can be acceptable.

University of South Wales
Prifysgol De Cymru

The median

If we write a data set in ascending order (smallest to largest), the **median** is the middle value.

Therefore:

If *n* is odd, the median is defined to be the value in position $\frac{n+1}{2}$;

If *n* is even, the median is the mean of the values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

University of South Wales
Prifysgol De Cymru

The median

Example

Suppose the heights in cm of five children of the same age are 107, 109, 110, 104 and 108.

Placing the data in order gives us 104, 107, 108, 109, 110.

Since $n = 5$, the median is in position 3 and is therefore 108cm.

Now suppose that another height, of 112 cm, is added. The revised ordered data set is now 104, 107, 108, 109, 110, 112.

Now $n = 6$ and so the median is the mean of the values in positions 3 and 4, i.e. 108.5cm.

University of South Wales
Prifysgol De Cymru

The mode

The mode is the most frequently occurring value in a data set.

Example

Consider the values 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6.

Value	Frequency
2	2
3	4
4	3
5	2
6	1

The mode is 3 because it occurs most frequently.

University of South Wales
Prifysgol De Cymru

Example

Example

A small business has 25 employees.

12 earn £15,000 pa; 10 earn £20,000 pa; 2 earn £50,000; 1 earns £200,000 pa;

Find the **mean, median and mode** of the earnings.

University of South Wales
Prifysgol De Cymru

Example

First we need to construct a frequency table and calculate *fx.*

Earnings (x)	Freq. (f)	fx
15	12	180
20	10	200
50	2	100
200	1	200
Total	**25**	**680**

We have $\bar{x} = \frac{\sum fx}{\sum f} = \frac{680}{25} = 27.20$ i.e. **Mean earnings** = £27,200 pa

Median is the 13th value (n=25) which is £20,000 pa
Mode is clearly £15,000

University of South Wales
Prifysgol De Cymru

Example

Note

We see that the mean is inflated by a few large values so may not be a fair representation of the ‘average’ earnings.

Similarly the mode is not very useful since most workers are on the lowest wages.

The median would seem to be the best way of expressing average earnings.

University of South Wales
Prifysgol De Cymru

Quartiles

Just as the median divides a set of data into two halves with respect to frequency, the **quartiles** divide the data into quarters.

The **middle quartile Q_2** is the median.

The **lower quartile Q_1** is defined to be the median of those data values to the **left** of the median (not including the median).

The **upper quartile Q_3** is defined to be the median of those data values to the **right** of the median (not including the median).

You should be aware that there are a number of ways of calculating the positions of the quartiles and these may give slightly different answers. We shall consider just one approach.

University of South Wales
Prifysgol De Cymru

Quartiles

Find the quartiles for the data set below:

24 15 14 20 17 23 16 24 25

Arranging the numbers in ascending order, we get:

14 15 16 17 20 23 24 24 25

Since we have n = 9 numbers, the median is the 5th number, that is, **Q2 = 20**.

University of South Wales
Prifysgol De Cymru

Quartiles

There are 4 numbers to the left of the median, 14, 15, 16, 17.

The lower quartile is the mean of the 2nd and 3rd, that is

$$Q_1 = \frac{15+16}{2} = 15.5$$

Similarly, there are 4 numbers to the right of the median, 23, 24, 24, 25.

So the upper quartile is the mean of the 2nd and 3rd, that is

$$Q_3 = \frac{24+24}{2} = 24$$

University of South Wales
Prifysgol De Cymru

Quartiles

Suppose that a value of 26 is added to the data set.

The data set in order is now:

14 15 16 17 20 23 24 24 25 26

Now we have n = 10, so the median is the mean of the 5th number and the 6th number, i.e. $Q_2 = 21.5$.

There are 5 numbers to the left of the median, so the lower quartile is the 3rd, i.e. $Q_1=16$.

Similarly, there are 5 numbers to the right of the median, so $Q_3=24$.

University of South Wales
Prifysgol De Cymru

Quartiles

For grouped data, the procedure for finding the median and quartiles is slightly different.

Q_1 is defined to be the value in position $\frac{n}{4}$;

Q_2 is defined to be the value in position $\frac{n}{2}$;

Q_3 is defined to be the value in position $\frac{3n}{4}$.

One way to obtain the quartiles is to draw a **cumulative frequency graph**.

University of South Wales
Prifysgol De Cymru

Cumulative Frequency Graph

This is a graph with **cumulative frequency** plotted on the Y-axis and **class boundary** plotted on the X-axis.

Successive points are joined by straight lines.

The quartiles and the median are the X values where the Y value equals $\frac{n}{4}$, $\frac{n}{2}$, $\frac{3n}{4}$.

The same approach is used for even and odd values of *n*, and that, just as when calculating the mean, using a grouped frequency table will give only an approximate answer for the median and quartiles.

Different choices of classes will usually give slightly different answers.

University of South Wales
Prifysgol De Cymru

Cumulative Frequency Graph

Example

Consider the weights of 50 boxes of floppy disks, to the nearest kilogram, as shown below:

27	29	31	26	25	28	32	31	30	30
22	23	40	38	32	29	25	32	26	32
30	36	33	36	29	30	23	28	27	39
34	29	31	37	31	30	26	24	28	30
29	35	37	33	26	31	30	38	26	24

University of South Wales
Prifysgol De Cymru

Cumulative Frequency Graph

We construct a grouped frequency table as follows:

Weight (kg)	Frequency
22 – 24	5
25 – 27	9
28 – 30	15
31 – 33	11
34 – 36	4
37 – 39	5
40 – 42	1
Total	**50**

For grouped data, we can identify the modal class (or classes) as the one(s) with the highest frequency. In the above example this is the class '28kg to 30 kg.'

University of South Wales
Prifysgol De Cymru

Cumulative Frequency Graph

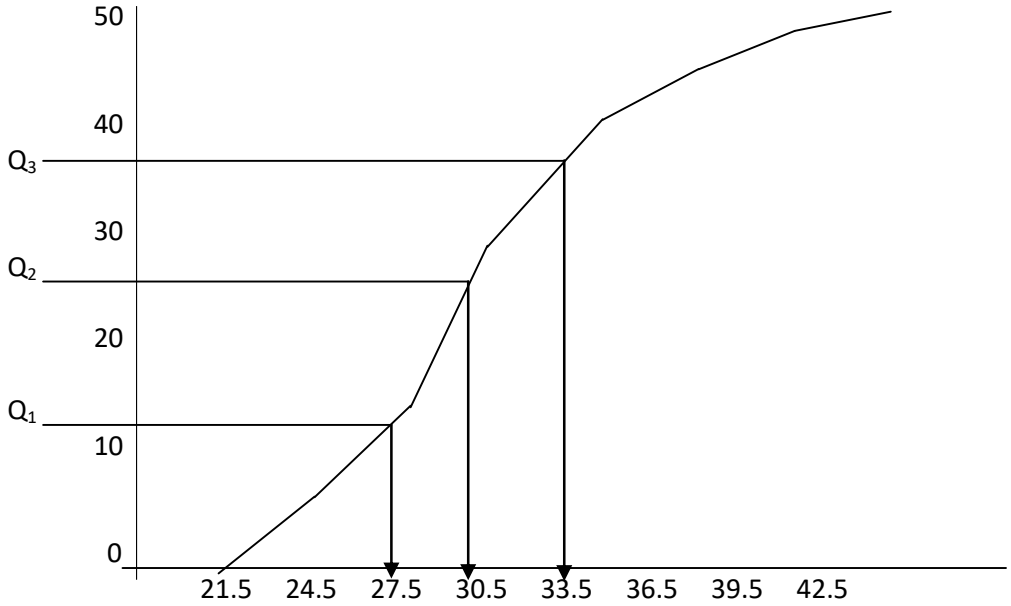
We now construct a table of **cumulative frequencies** using the class boundaries as follows:

Less than	y
21.5	0
24.5	5
27.5	14
30.5	29
33.5	40
36.5	44
39.5	49
42.5	50

University of South Wales
Prifysgol De Cymru

Cumulative Frequency Graph

Next we draw the graph from which we can read off the quartiles:



The total frequency is 50, so we need to find the X values corresponding to
Y = 12.5, Y =25; Y= 37.5.
From the above graph we find the quartiles are 27.0 and 32.8, and the median is 29.7.

University of South Wales
Prifysgol De Cymru

Measures of spread

Having obtained a measure of location or average for a set of data, we now need an additional measure which gives us an indication of how “spread out” a set of data is.

The two sets of data {9, 10, 11} and {5, 10, 15} each have a mean (and median) of 10.

The distinction between them is the spread of the values within each data set.

We shall examine a number of measures which aim to tell us how closely a set of data is grouped around its mean or median.

University of South Wales
Prifysgol De Cymru

Range

The **range** of a data set is the difference between the largest and smallest values.

For example, if the ages in years of a family group were 30, 2, 7, 4, 32 and 10.

The range would be 32 – 2 = 30 years.

However, the range is not a very satisfactory measure of spread since it depends entirely on the smallest and largest values which may represent nothing more than experimental errors.

University of South Wales
Prifysgol De Cymru

The Interquartile range

When discussing measures of average, we introduced the idea of quartiles.

We define the **interquartile range** as:

IQR = $Q_3 - Q_1$.

This is the interval within which half the observations lie.

It therefore gives some information on the dispersion of a data set but, by definition, excludes extreme values.

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

The dispersion, or spread, of a set of data will be small if the data values are tightly grouped about the mean and large if they are widely scattered.

It makes sense, therefore, to consider dispersion in the context of deviations from the mean.

This has one big drawback.

For example, we have seen that the mean of the 5 values {8, 3, 7, 12, 11} is 8.2.

If we calculate the sum of the deviations from the mean we get:

$$(8-8.2)+(3-8.2)+(7-8.2)+(12-8.2)+(11-8.2)$$
$$=-0.2-5.2-1.2+3.8+2.8$$
$$=0$$

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

This is no good as a measure of dispersion since the positive and negative differences simply cancel each other out.

This happens for any data set.

One way of eliminating negative numbers is to square them.

So, consider a **population of *N*** numbers $x_1, x_2, \ldots . x_N$ with mean .

The **population variance** is defined as $\sigma^2 = \dfrac{\sum (x_i - \mu)^2}{N}$

This formula can lead to complicated calculations.

So the following simplified formula will be used $\sigma^2 = \left(\dfrac{1}{N} \sum x_i^2 \right) - \mu^2$

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

However, suppose our data are measured in centimeters.

Then the variance will be measured in square centimeters and it is clearly unsatisfactory to have a measure of dispersion whose units differ from those of the original data.

To overcome this, we define the **population standard deviation** $\sigma = \sqrt{\frac{1}{N}\sum(x_i - \mu)^2}$

For ease of calculation, $\sigma = \sqrt{\left(\frac{1}{N}\sum x_i^2\right) - \mu^2}$

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

Example

Consider the set of numbers:

5.2 7.1 6.9 3.1 4.7 6.3 8.0 8.1 5.9

Suppose we wish to calculate the population standard deviation.

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

To aid our calculations, we construct the following table:

x	x^2
5.2	27.04
7.1	50.41
6.9	47.61
3.1	9.61
4.7	22.09
6.3	39.69
8.0	64.00
8.1	65.61
5.9	34.81
55.3	**360.87**

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

Here, we have $N = 9$, $\sum x = 55.3$ and $\sum x^2 = 360.87$

Thus the mean = $\frac{55.3}{9} = 6.1444$

The population variance is = $\frac{360.87}{9} - 6.1444^2 = 2.342469$

Hence the standard deviation is = $\sqrt{2.342469} = 1.5305$

As was the case with the mean, we modify our procedure slightly when we have the data already presented in the form of a **frequency distribution**.

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

Suppose the values $x_1, x_2, \ldots, x_N$ occur with frequencies $f_1, f_2, \ldots, f_N$.

Recall that the mean, $\mu = \dfrac{\sum f_i x_i}{\sum f_i}$.

Then the variance is $\sigma^2 = \dfrac{\sum f_i (x_i - \mu)^2}{\sum f_i} = \dfrac{\sum f_i x_i^2}{\sum f_i} - \mu^2$

and so the standard deviation is $\sigma = \sqrt{\dfrac{\sum f_i (x_i - \mu)^2}{\sum f_i}} = \sqrt{\dfrac{\sum f_i x_i^2}{\sum f_i} - \mu^2}$.

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

Consider again the example we used previously concerning the number of goals scored by a football team. We can expand the table to give:

Goals (x)	Games (f)	fx	fx^2
0	4	0	0
1	3	3	3
3	2	6	18
5	1	5	25
Total	10	14	46

Mean = $\frac{14}{10}$ = 1.4; Variance = $\frac{46}{10} - 1.4^2 = 2.64$ and Standard deviation = $\sqrt{2.64} = 1.6248$

For **grouped data**, the same formula is used where the *x*'s are the class mid-points.

University of South Wales
Prifysgol De Cymru

Variance and standard deviation

Example

Consider again the example concerning the weights of 35 packages. The expanded table is:

Class interval	Class mid-point x	Frequency f	fx	fx^2
59 – 61	60	2	120	7200
62 – 64	63	4	252	15876
65 – 67	66	5	330	21780
68 – 70	69	6	414	28566
71 – 73	72	8	576	41472
74 - 76	75	7	525	39375
77 – 79	78	3	234	18252
	Total	**35**	**2451**	**172521**

Mean $= \frac{2451}{35} \cong 70.03$

Variance $= \frac{172521}{35} - 70.03^2 \cong 25.17$

Standard Deviation $= \sqrt{25.170612} \cong 5.02$

University of South Wales
Prifysgol De Cymru

The sample standard deviation

In practice, it is likely that we are using **sample data** and calculating a **sample standard deviation** to provide an estimate of the corresponding **population standard deviation** .

If we were to calculate the **sample standard deviation**, denoted by **s**, using the same basic formula, it is the case that **s** will tend to underestimate .

It can be shown that the best estimate of the population standard deviation is actually:

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

The corresponding simplified version of this formula is:

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

University of South Wales
Prifysgol De Cymru

The sample standard deviation

Example

Consider again the data used previously. Suppose that the set of observations:

5.2 7.1 6.9 3.1 4.7 6.3 8.0 8.1 5.9

Now represent a set of values or measurements collected from a sample of respondents.

We again use the expanded table (as above) so we still have:

N = 9, $\sum x$ = 55.3 and $\sum x^2$ = 360.87. Thus the sample mean = $\frac{55.3}{9} = 6.1444$ (as before) .

But the sample variance = $\frac{360.87}{8} - \frac{9}{8} 6.1444^2 = 2.6352778$

Hence the standard deviation is = $\sqrt{2.6352778} = 1.6234$

University of South Wales
Prifysgol De Cymru

The sample standard deviation

For **grouped data** the simplification formula for the **sample standard deviation** is

$$s = \sqrt{\left(\frac{1}{n-1}\sum fx^2\right) - \frac{n}{n-1}\bar{x}^2} \qquad \text{where} \quad \bar{x} = \frac{\sum fx}{\sum f} \quad .$$

So if, in the previous example, the 35 packages are regarded as a sample from a large batch then the sample standard deviation is given by:

$$\sqrt{\frac{172521}{34} - \frac{35}{34}\,70.028571^2} = 5.0903$$

University of South Wales
Prifysgol De Cymru

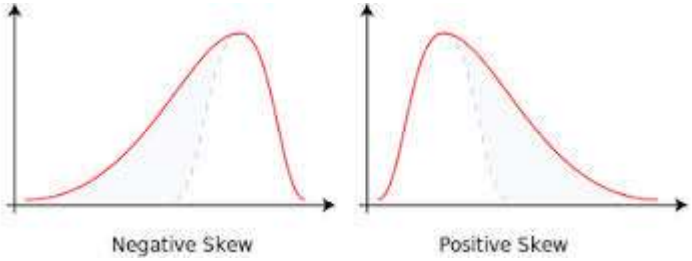
Skewness

This measures how much (or little) symmetry there is in a set of data.

For a very large number of observations, the frequency polygon would look more like a curve than a set of straight line segments.

If this curve has a longer tail to the right than to the left, it is said to be **skewed to the right**, or to have **positive skewness**.

Conversely, if the curve has a longer tail to the left than to the right, it is said to be **skewed to the left**, or to have **negative skewness**.

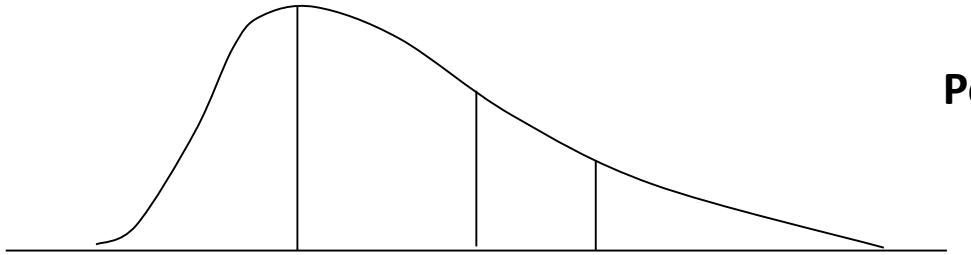


University of South Wales
Prifysgol De Cymru

Skewness

Example

A wages or income distribution will tend to be skewed to the right because of a small number of very large values not compensated for by very small values.



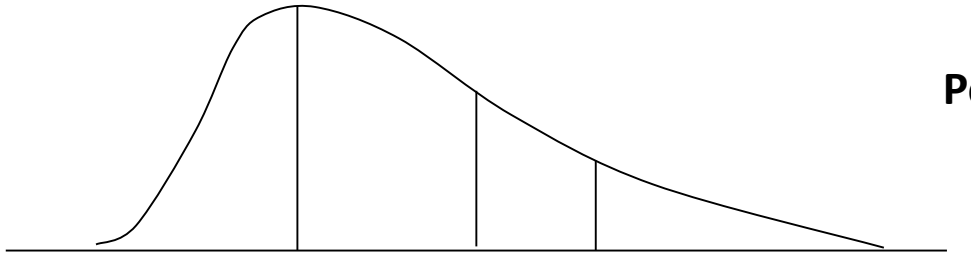
Positive skewness

University of South Wales
Prifysgol De Cymru

Skewness

Example

A wages or income distribution will tend to be skewed to the right because of a small number of very large values not compensated for by very small values.



Positive skewness

The **Pearson Coefficient of Skewness**, is defined by **= 3(mean – median) / standard deviation**

University of South Wales
Prifysgol De Cymru

Pearson's Coefficient of Skewness

Example

Consider again the example which dealt with the weights in grams of 35 packages. The data (in order) are reproduced below:

59	60	62	63	63	64	66
66	67	67	67	68	69	69
70	70	70	71	71	72	72
73	73	73	73	74	74	74
75	75	75	76	77	78	78

We have already found that the mean is 70.03, and the standard deviation is 5.02 (based on the grouped frequency table). The median is 71 (18^{th} value).

The Pearson Coefficient of Skewness is therefore $= \dfrac{3.(70.03 - 71)}{5.02} = -0.58$ (negative skew)

University of South Wales
Prifysgol De Cymru

Quartile Coefficient of Skewness

Another measure of skewness, defined in terms of the quartiles, is the **Quartile Coefficient of Skewness**, given by

$$\text{Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

The above example gives:

Median = Q_2 = 71 (18^{th} value)
Q_1 = 67 (9^{th} value)
Q_3 = 74 (27^{th} value)

$$\text{Skewness} = \frac{74 - 2*71 + 67}{74 - 67} = \frac{-1}{7} = -0.14$$