University of South Wales
Prifysgol De Cymru

MS4S08 – Applied Statistics for Data Science

Additional Statistics Exercises & Solutions

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

University of South Wales
Prifysgol De Cymru

Exercises 1

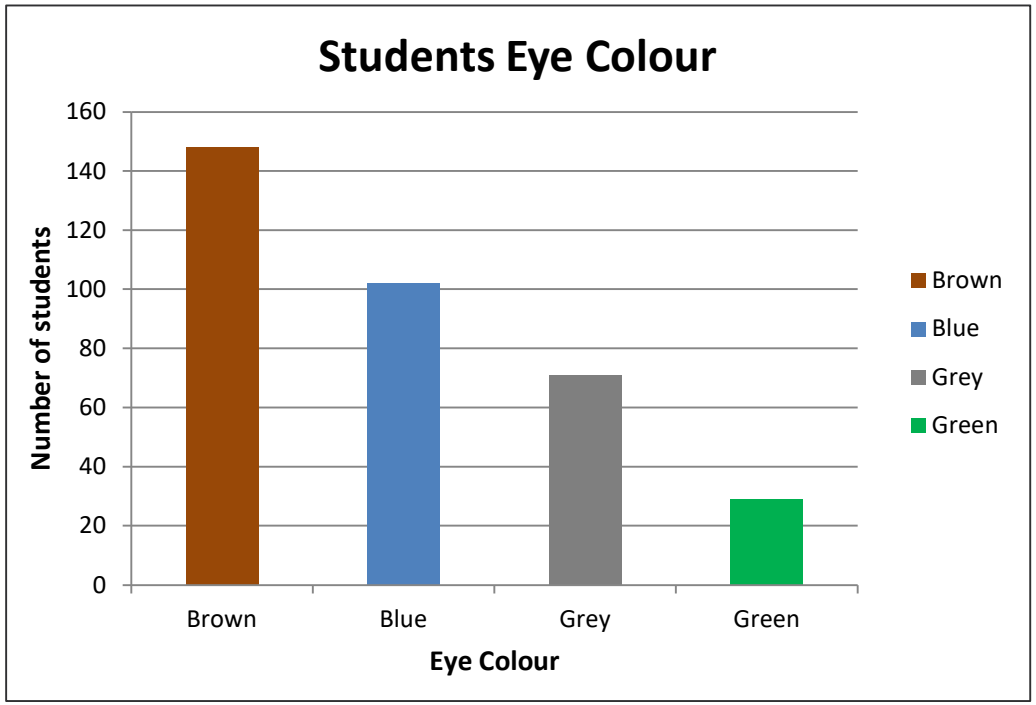
It is time to work through some exercises yourselves.

Please feel free to ask any questions.

University of South Wales
Prifysgol De Cymru

Exercises 1

1. 350 students have eye colour as follows: Brown 148; Blue 102; Grey 71; Green 29. Draw a bar chart to represent these data.



University of South Wales
Prifysgol De Cymru

Exercises 1

2. The 50 digits below are the output from a computer random number generator:

7	5	8	4	7	4	1	2	1	0
5	3	7	5	0	2	4	4	1	4
2	7	6	2	7	0	7	5	1	9
0	9	1	4	3	0	6	6	4	0
5	5	6	1	1	0	8	3	9	8

Form a frequency table and draw a bar chart to represent these data.

University of South Wales
Prifysgol De Cymru

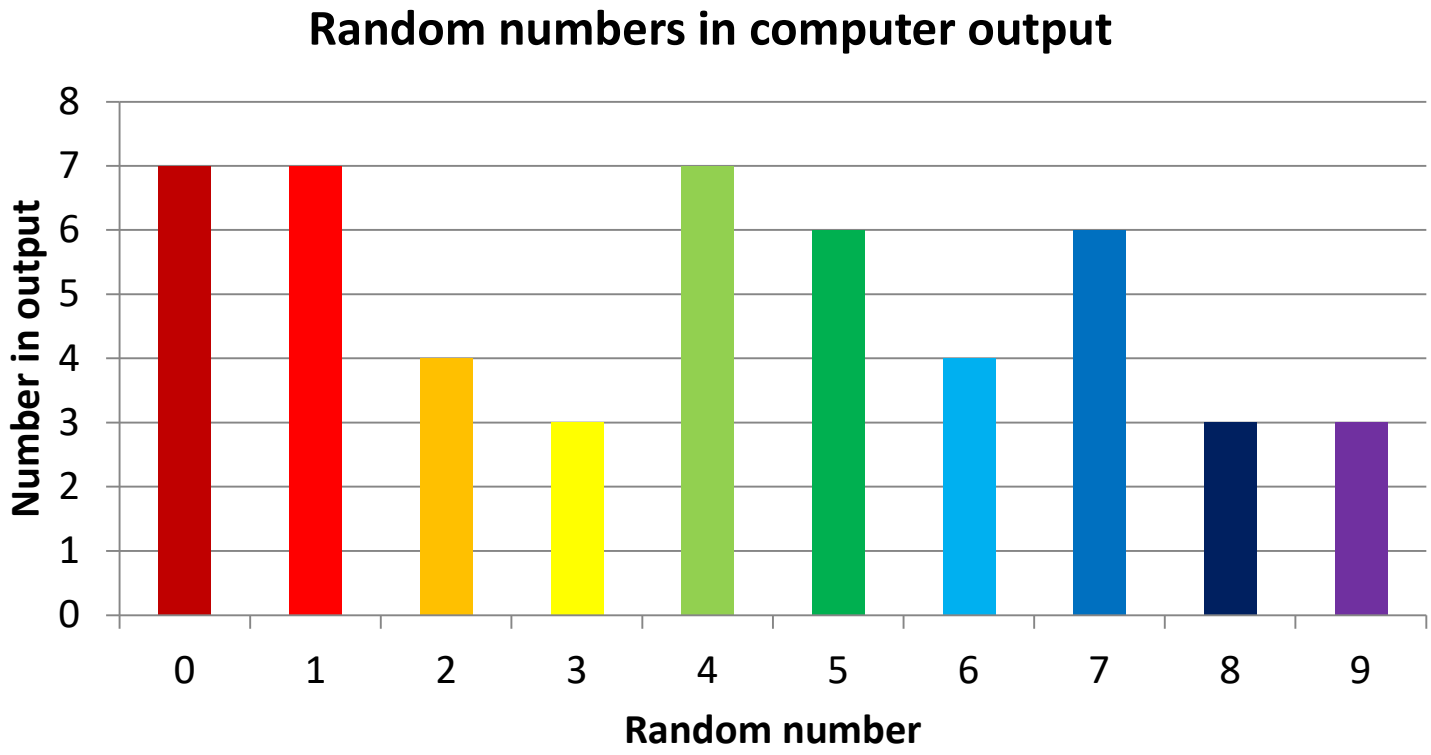
Exercises 1

We form the table as follows:

Random number	Tally	Frequency
0	~~1111~~ 11	7
1	~~1111~~ 11	7
2	1111	4
3	111	3
4	~~1111~~ 11	7
5	~~1111~~ 1	6
6	1111	4
7	~~1111~~ 1	6
8	111	3
9	111	3
Total		50

University of South Wales
Prifysgol De Cymru

Exercises 1



University of South Wales
Prifysgol De Cymru

Exercise 2

It is time to work through some exercises yourselves.

Please feel free to ask any questions.

University of South Wales
Prifysgol De Cymru

Exercise 2

1. The distances, to the nearest mile, travelled from home to their workplace for 40 employees are shown below:

14	16	12	11	12	17	8	33
19	20	12	15	24	27	37	11
30	23	13	28	10	28	35	15
27	18	19	29	20	36	8	16
19	30	25	21	10	22	9	17

Construct an ordered stem-and-leaf diagram. Hence construct a double-stem diagram and draw a histogram to represent the data.

(Note that ‘8’ can be written as’08’ etc.)

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the stem-and-leaf diagram, we note that the range is from 8 to 37, so the stems will be 0, 1, 2 or 3.

This gives the **unordered** stem-and-leaf diagram :

0	889
1	46211792513058969 07
2	047388790512
3	370560

We then **order** the 'leaves' within each stem value. The **ordered** diagram will be:

0	889
1	00111223455667789 99
2	001234577889
3	003567

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the **double-stem** diagram, we split each stem value so that leaf values 0 to 4 and 5 to 9 appear in separate rows.

This gives the **double-stem** diagram :

0	889
1	001112234
1	5566778999
2	001234
2	577889
3	003
3	567

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the **histogram,** we place the data into classes, or intervals.

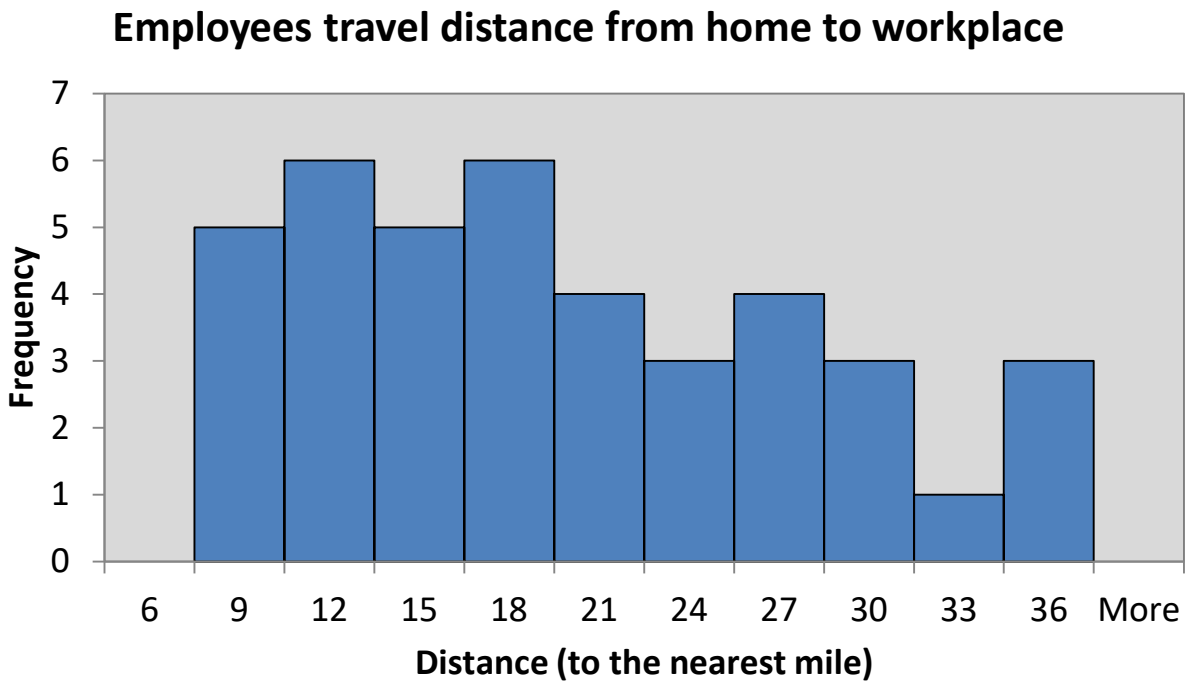
We have:
Smallest value = 8
Largest value = 37
Range = 37 - 8 = 29
Total interval width = 29 + 1 = 30

One possibility is:
8-10
11-13
…….
32-34
35-37
That is, there are 3 rounded value on each interval and 10 intervals in total.

University of South Wales
Prifysgol De Cymru

Exercise 2

The corresponding **histogram** is:



University of South Wales
Prifysgol De Cymru

Exercise 2

2. A sample of 60 access times (each to the nearest millisecond) onto a computer system are as shown:

894	892	907	908	902	911	894	909	890	903
897	914	892	889	906	913	886	896	910	909
901	898	904	901	901	898	912	911	889	908
885	886	881	904	881	894	879	901	902	916
907	904	897	911	917	928	917	909	921	925
885	913	921	920	889	895	902	904	906	901

Construct a grouped frequency table for these data and hence draw a histogram.

University of South Wales
Prifysgol De Cymru

Exercise 2

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

We have:
Smallest value = 879
Largest value = 928
Range = 928 - 879 = 49
Total width of interval = 49 + 1 = 50

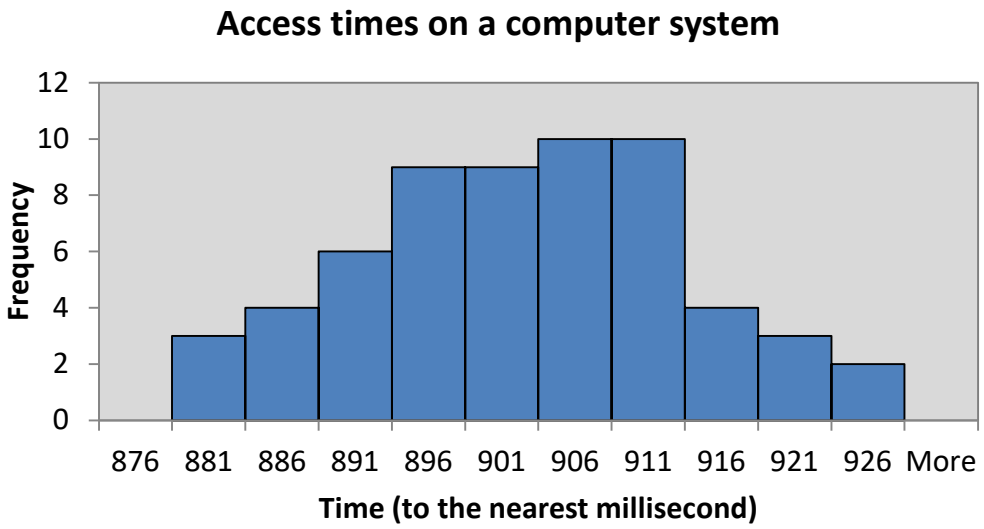
One possibility is:
879 - 883
884 - 888
…….
919 - 923
924 - 928
That is, there are 5 rounded values on each interval and 10 intervals in total.

University of South Wales
Prifysgol De Cymru

Exercise 2

The **grouped frequency table and histogram** is thus:

Class Interval	Class Midpoint	Frequency
879 - 883	881	3
884 - 888	886	4
889 - 893	891	6
894 - 898	896	9
899 - 903	901	9
904 - 908	906	10
909 - 913	911	10
914 - 918	916	4
919 - 923	921	3
924 - 928	926	2
Total		60



University of South Wales
Prifysgol De Cymru

Exercise 2

3. The times taken (to the nearest minute) by 20 students to complete a mathematics test were as follows:

29	43	33	40	20	31	17	29	39	40
21	26	38	15	38	43	32	21	44	37

Construct an ordered double stem-and-leaf diagram to represent these data.
Hence draw a histogram to represent the data.

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the stem-and-leaf diagram, we note that the range is from 15 to 44, so the stems will be 1, 2, 3 or 4.

This gives the **unordered** stem-and-leaf diagram :

1	75
2	909161
3	3198827
4	30034

We then **order** the 'leaves' within each stem value. The **ordered** diagram will be:

1	57
2	011699
3	1237889
4	00334

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the **double-stem** diagram, we split each stem value so that leaf values 0 to 4 and 5 to 9 appear in separate rows.

This gives the **double-stem** diagram :

1	57
2	011
2	699
3	123
3	7889
4	00334

University of South Wales
Prifysgol De Cymru

Exercise 2

To construct the **histogram,** we place the data into classes, or intervals.

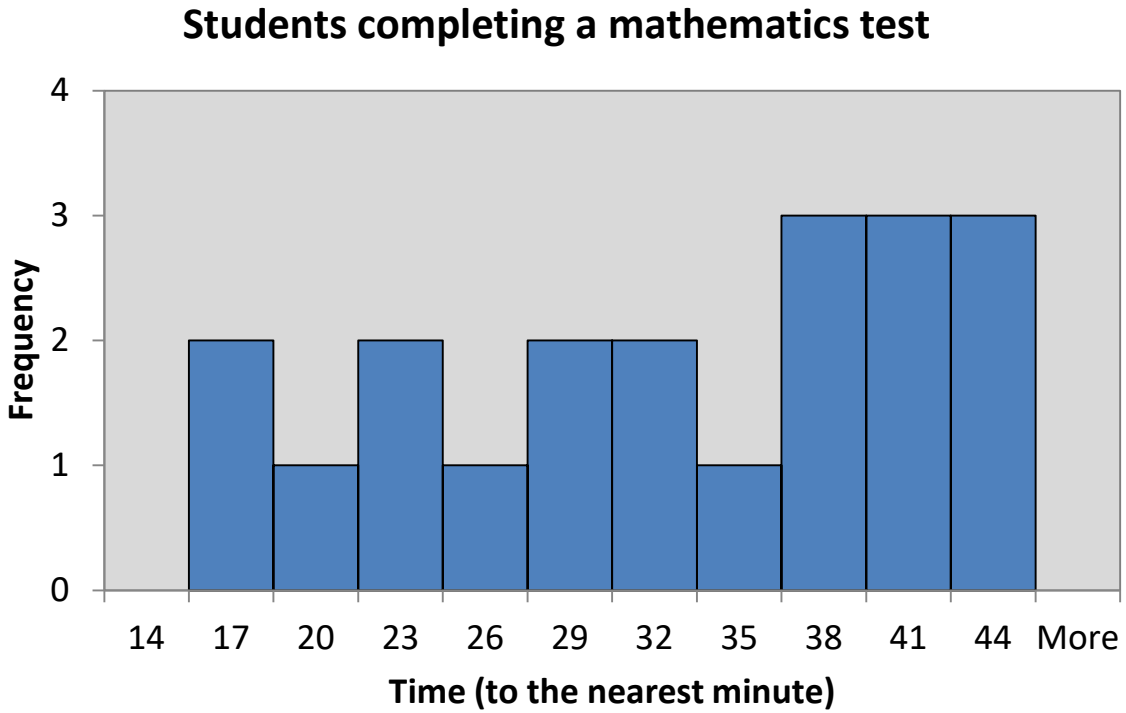
We have:
Smallest value = 15
Largest value = 44
Range = 44 - 15 = 29
Total interval width = 29 + 1 = 30

One possibility is:
15-17
18-20
…….
39-41
42-44
That is, there are 3 rounded value on each interval and 10 intervals in total.

University of South Wales
Prifysgol De Cymru

Exercise 2

The corresponding **histogram** is:



University of South Wales
Prifysgol De Cymru

Exercises 3

It is time to work through some exercises yourselves.

Please feel free to ask any questions.

University of South Wales
Prifysgol De Cymru

Exercises 3

1. The marks of a student in five examinations were 84, 91, 72, 68 and 85.

Find the mean and median mark.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{84 + 91 + 72 + 68 + 85}{5} = \frac{400}{5} = 80$$

The median is the 3rd value (n=5).

The numbers in order are: 68, 72, 84, 85, 91

Therefore the 3rd number is 84.

University of South Wales
Prifysgol De Cymru

Exercises 3

2. Calculate the mean, median and mode of the following sets of data and comment on the relative merits of each measure.

(a) the weekly wages in £s of ten workers, including the manager, in a factory are:
130 132 132 135 135 135 136 138 139 200

(b) the age, in years, of four children in a family are:
6 8 12 14

(c) the colours of fourteen helmets seen on a building site were:
2 red, 1 black, 3 green, 5 yellow, 3 white

(d) the price, pence, of a tin of baked beans in five shops:
14 17 18 19 19

University of South Wales
Prifysgol De Cymru

Exercises 3

(a) $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{130 + 132 + \ldots\ldots + 200}{10} = \frac{1412}{10} = 141.2$

The median is between the 5th and 6th values (n=10).

The numbers are in order, therefore the 5th number is 135 and the 6th number is 135, so the median is 135.

The mode is 135.

We see that the mean is inflated by a large value so may not be a fair representation of the 'average' wages.
In this case the median is equal to the mode and this would seem to be the best way of expressing average wages.

University of South Wales
Prifysgol De Cymru

Exercises 3

(b) $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{6+8+12+14}{4} = \frac{40}{4} = 10$

Median is between the 2nd and 3rd values (n=4).

The numbers are in order, therefore the 2nd number is 8 and the 3rd number is 12, so the median is 10.

The modes of the data are 6, 8, 12 and 14.

For this case, the mean is equal to the median, this highlights that the number of values are equally spaced either side of the mean. Therefore this is a good measure for expressing the average number of children.

As each value appears only once, the mode is not useful here.

University of South Wales
Prifysgol De Cymru

Exercises 3

(c) This is qualitative data therefore we are not able to calculate a mean.

The median identifies the middle, suggesting that there is a way to arrange the items in some order. Some qualitative measures can be arranged in this way, but not all.

We are not able to arrange the variable (colour) in numeric order, therefore we are also not able to calculate the median.

The mode identifies the item that occurs most frequently, so it can be either quantitative or qualitative.

The mode is therefore yellow and this is the only measure we can calculate to express an average for the colour of helmets.

University of South Wales
Prifysgol De Cymru

Exercises 3

(d) $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{14+17+18+19+19}{5} = \frac{87}{5} = 17.4$

The median is the 3rd value (n=5).

The numbers are in order, therefore the 3rd number is 18, so the median is 18.

The mode is 19.

For this example we obtain a similar value for all three measures of the average.

The median is slightly higher than the mean showing that the data is slightly skewed to the right, this is due to the two higher values of 19 and is further highlighted by the mode value being 19.

University of South Wales
Prifysgol De Cymru

Exercises 3

3. The table below gives the numbers of typing errors in a sample of word-processed reports:

No. of errors:	0	1	2	3	4	5	6	7	8
No. of reports:	39	21	21	13	10	7	4	2	1

Calculate the mean, median and mode for the number of errors per report.

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

University of South Wales
Prifysgol De Cymru

Exercises 3

Errors (x)	Freq. (f)	*fx*
0	39	0
1	21	21
2	21	42
3	13	39
4	10	40
5	7	35
6	4	24
7	2	14
8	1	8
Total	**118**	**223**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{223}{118} = 1.89$$

Mean = 1.89

Median is the 59.5th value (n=118) which is 1.

Mode is clearly 0.

This data is clearly skewed and it is difficult to express a meaningful average.

University of South Wales
Prifysgol De Cymru

Exercises 3

4. For Question 2 in Exercises 2, use the grouped frequency table you found to obtain approximations to the mean and median access times.

Class Interval	Class Midpoint	Frequency
879 - 883	881	3
884 - 888	886	4
889 - 893	891	6
894 - 898	896	9
899 - 903	901	9
904 - 908	906	10
909 - 913	911	10
914 - 918	916	4
919 - 923	921	3
924 - 928	926	2
Total		60

University of South Wales
Prifysgol De Cymru

Exercises 3

Class Midpoint (x)	Frequency (f)	*fx*
881	3	2,643
886	4	3,544
891	6	5,346
896	9	8,064
901	9	8,109
906	10	9,060
911	10	9,110
916	4	3,664
921	3	2,763
926	2	1,852
Total	**60**	**54,155**

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{54155}{60} = 902.58$$

Mean = 902.58

Median is the 30.5th value (n=60) which is 901.

The mean and the median are very similar indicating an even spread of the data either side of the mean.

University of South Wales
Prifysgol De Cymru

Exercises 4

It is time to work through some exercises yourselves.

Please feel free to ask any questions.

University of South Wales
Prifysgol De Cymru

Exercises 4

1. The marks of a student in five examinations were 84, 91, 72, 68 and 85. Find the (population) standard deviation of these marks.

From Q1 Exercise 3 we know $\mu = 80$.

The (population) standard deviation $\sigma = \sqrt{\left(\frac{1}{N}\sum x_i^2\right) - \mu^2}$

$\sum x_i^2 = 84^2 + 91^2 + 72^2 + 68^2 + 85^2 = 32{,}370$

Therefore: $\sigma = \sqrt{\left(\frac{1}{5} * 32370\right) - 80^2} = \sqrt{74} = 8.6023$

University of South Wales
Prifysgol De Cymru

Exercises 4

2. The weekly wages in £s of ten workers, including the manager, in a factory are:
130 132 132 135 135 135 136 138 139 200

Find the inter quartile range for this data.

From Q2a Exercise 3 we know that median =135 (the mean of the 5^{th} and 6^{th} numbers).

There are 5 numbers to the right of the median. Therefore the lower quartile is the 3^{rd} number, that is 132.

Similarly, there are 5 numbers to the right of the median, so the upper quartile is the again the 3^{rd} number, that is 138.

Hence the IQR = $Q_3 - Q_1 = 138 - 132 = 6$.

University of South Wales
Prifysgol De Cymru

Exercises 4

3. The table below gives the numbers of typing errors in a sample of word-processed reports:

No. of errors:	0	1	2	3	4	5	6	7	8
No. of reports:	39	21	21	13	10	7	4	2	1

Calculate the (sample) standard deviation for the number of errors per report, and calculate the inter-quartile range.

University of South Wales
Prifysgol De Cymru

Exercises 4

Errors (x)	Freq. (f)	fx	fx^2
0	39	0	0
1	21	21	21
2	21	42	84
3	13	39	117
4	10	40	160
5	7	35	175
6	4	24	144
7	2	14	98
8	1	8	64
Total	**118**	**223**	**863**

From Q3 Exercise 3 we have:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{223}{118} = 1.89$$

Mean = 1.89

Median is the 59.5th value (n=118) which is 1.

Adding another column we have $\sum fx^2$ = 863.

University of South Wales
Prifysgol De Cymru

Exercises 4

There are 59 numbers to the right of the median. Therefore the lower quartile is the 30th number, that is 0.

Similarly, there are 59 numbers to the right of the median, so the upper quartile is the again the 30th number from the median (89th), that is 3.

Hence the IQR = $Q_3 - Q_1 = 3 - 0 = 3$.

The sample standard deviation $s = \sqrt{\left(\frac{1}{n-1}\sum fx^2\right) - \frac{n}{n-1}\bar{x}^2}$

$$s = \sqrt{\left(\frac{1}{118-1} * 863\right) - \frac{118}{117} * 1.89^2} = \sqrt{3.7734} = 1.9425$$

University of South Wales
Prifysgol De Cymru

Exercises 4

4. A sample of 60 access times (to the nearest millisecond) onto a computer system are as shown:

894	892	907	908	902	911	894	909	890	903
897	914	892	889	906	913	886	896	910	909
901	898	904	901	901	898	912	911	889	908
885	886	881	904	881	894	879	901	902	916
907	904	897	911	917	928	917	909	921	925
885	913	921	920	889	895	902	904	906	901

Use the grouped frequency table you have already found to obtain an approximation to the sample standard deviation.

University of South Wales
Prifysgol De Cymru

Exercises 4

Class Midpoint (x)	Frequency (f)	*fx*	*fx²*
881	3	2,643	2,328,483
886	4	3,544	3,139,984
891	6	5,346	4,763,286
896	9	8,064	7,225,344
901	9	8,109	7,306,209
906	10	9,060	8,208,360
911	10	9,110	8,299,210
916	4	3,664	3,356,224
921	3	2,763	2,544,723
926	2	1,852	1,714,952
Total	**60**	**54,155**	**48,886,775**

From Q4 Exercise 3 we have:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{54155}{60} = 902.58$$

Mean = 902.58

Adding another column we have $\sum fx^2$ = 48,886,775.

$$s = \sqrt{\left(\frac{1}{59} * 48886775\right) - \frac{60}{59} * 902.58^2}$$

$$s = \sqrt{131.1121} = 11.4504$$

University of South Wales
Prifysgol De Cymru

Exercises 5

It is time to work through some exercises yourselves.

Please feel free to ask any questions.

University of South Wales
Prifysgol De Cymru

Exercises 5

1. The marks of a student in five examinations were 84, 91, 72, 68 and 85. Find the (sample) standard deviation of these marks.

From Q1 Exercise 3 & 4 we know $\bar{x} = 80$.

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

$\sum x_i^2 = 84^2 + 91^2 + 72^2 + 68^2 + 85^2 = 32{,}370$

Therefore: $$s = \sqrt{\left(\frac{1}{5-1} * 32370\right) - \frac{5}{5-1} * 80^2} = \sqrt{92.5} = 9.6177$$

Compared to: $\sigma = 8.6023$

University of South Wales
Prifysgol De Cymru

Exercises 5

2. The table below gives the numbers of typing errors in a sample of word-processed reports:

No. of errors:	0	1	2	3	4	5	6	7	8
No. of reports:	39	21	21	13	10	7	4	2	1

Calculate two measures of skewness for this data.

From Q3 Exercise 3 and 4 we know: $\bar{x}$ = 1.89, Q_1 = 0 , Q_2 = 1 , Q_3 = 3 , s =1.9425

Pearson Coefficient of Skewness = $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3*(1.89-1)}{1.9425} = 1.3745$

Quartile Coefficient of Skewness $= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{3 - 2*1 + 0}{3 - 0} = \frac{1}{3}$

University of South Wales
Prifysgol De Cymru

Exercises 5

3. For Q4 of Sheets 3 & 4, use the data in the grouped frequency table you have already found to obtain an approximation to the mean and the sample standard deviation.

A sample of 60 access times (to the nearest millisecond) onto a computer system are as shown:

894	892	907	908	902	911	894	909	890	903
897	914	892	889	906	913	886	896	910	909
901	898	904	901	901	898	912	911	889	908
885	886	881	904	881	894	879	901	902	916
907	904	897	911	917	928	917	909	921	925
885	913	921	920	889	895	902	904	906	901

University of South Wales
Prifysgol De Cymru

Exercises 5

Class Midpoint (x)	Frequency (f)	*fx*	fx^2
881	3	2,643	2,328,483
886	4	3,544	3,139,984
891	6	5,346	4,763,286
896	9	8,064	7,225,344
901	9	8,109	7,306,209
906	10	9,060	8,208,360
911	10	9,110	8,299,210
916	4	3,664	3,356,224
921	3	2,763	2,544,723
926	2	1,852	1,714,952
Total	**60**	**54,155**	**48,886,775**

From Q4 Exercise 3 we have:

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{54155}{60} = 902.58$$

Mean = 902.58

Adding another column we have $\sum fx^2$ = 48,886,775.

$$s = \sqrt{\left(\frac{1}{59} * 48886775\right) - \frac{60}{59} * 902.58^2}$$

$$s = \sqrt{131.1121} = 11.4504$$