10 More bivariate analysis

Try Section A after you have completed Exercise 10C.
Try Section B after you have completed Exercise 10F.

Section A

Spearman's rank correlation coefficient

The results of many sports competitions such as diving, gymnastics, dance and boxing are based on judges' scores. Often the judges all give different scores. You can use **Spearman's rank correlation coefficient** to compare two sets of data, to see if there is any connection or relationship between the data. This helps you to **test** the degree of agreement between the judges.

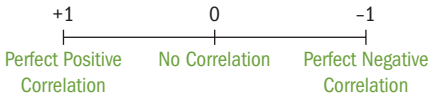
Here we are not interested in examining the possibility of a linear relationship between the scores. We are looking to see if there is a correlation between the **ranks** of contestants as judged by two judges.

Spearman's rank correlation coefficient (r_s)

$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$ where d is the difference between the ranks and n is the number of ranks.

What does this tell us?

- r_s always lies in the range −1 to +1. If it lies close to either of these two values a strong correlation exists between the two variables.
- When $r_s = 1$ there is a perfect positive correlation. The two judges are in perfect agreement.
- When $r_s = 0$ there is no correlation between the judges' ranks.
- When $r_s = -1$ there is perfect negative correlation, the judges' ranks are completely opposite.



This table will help you interpret Spearman's rank correlation coefficient.

r_s	Correlation
$r_s = 1$	perfect positive correlation
$0.7 \leq r_s < 1$	strong positive correlation
$0.4 \leq r_s < 0.7$	moderate positive correlation
$0 < r_s < 0.4$	weak positive correlation
$r_s = 0$	no correlation
$0 > r_s > -0.4$	weak negative correlation
$-0.4 \geq r_s > -0.7$	moderate correlation
$-0.7 \geq r_s < -1$	strong negative correlation
$r_s = -1$	perfect negative correlation

Example 1

In a dance competition two judges rank the eight competitors as shown in the table.

Dancer	Ruby	Finn	Tulisa	Pia	Cher	Leon	Patti	Jake
Rank (Judge A)	2	5	3	6	1	4	7	8
Rank (Judge B)	4	3	2	6	1	8	5	7

a Calculate Spearman's rank correlation coefficient for these results.
b Comment on the degree of agreement between the two judges.

Answers

a

Dancer	Rank (Judge A)	Rank (Judge B)	Difference d (A – B)	Difference^2 d^2
Ruby	2	4	–2	$(-2)^2 = 4$
Finn	5	3	2	$(2)^2 = 4$
Tulisa	3	2	1	$(1)^2 = 1$
Pia	6	6	0	$(0)^2 = 0$
Cher	1	1	0	$(0)^2 = 0$
Leon	4	8	–4	$(-4)^2 = 16$
Patti	7	5	2	$(2)^2 = 4$
Jake	8	7	1	$(1)^2 = 1$
				$\Sigma d^2 = 30$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(30)}{8(8^2-1)} = 1 - \frac{180}{504} \approx 0.643$$

b There is a moderate positive correlation between the judges.

Make a 5 column table with the first 3 column headings from the data given.

The fourth column is for the difference (d) where d = Judge A's rank – Judge B's rank.

The fifth column is for the difference squared (d^2).

Use your GDC for the calculation.

Exercise 1

1 **a** Calculate Spearman's rank correlation coefficient for the data below, which compares the ranks in mathematics and science for 10 students.

Student	A	B	C	D	E	F	G	H	I	J
Rank in Mathematics	1	3	7	5	4	6	2	10	9	8
Rank in Science	3	1	4	5	6	9	7	8	10	2

b What does the sign of r_s tell you?

2 The following table shows the ranked annual income per person and the infant death rate for a sample of 11 countries.

Country	A	B	C	D	E	F	G	H	I	J	K
Rank in income	1	10	4	8	7	2	3	5	6	11	9
Rank in infant death rate	10	4	8	5	3	11	9	6	7	1	2

Rank correlation is a useful tool in geography and social studies.

a Calculate Spearman's rank correlation coefficient for the data.
b What does the r_s value tell you about the income and infant death rate?

3 A farmer wonders if he has winter hens and summer hens. That is, he wants to know if the hens that lay the most eggs in winter also lay the most in summer. He counted the number of eggs laid by 11 hens in January and July.

Hen	Eggs in January	Eggs in July
A	105	99
B	60	66
C	76	91
D	45	67
E	10	17
F	114	132
G	34	66
H	9	12
I	44	70
J	52	83
K	29	72

Here the data is for the number of eggs laid. You will have to rank the hens.

a Calculate Spearman's rank correlation coefficient for the data.
b What does the r_s value tell the farmer?

Section B

Coefficient of determination, r^2

You have probably noticed the r^2 value on your GDC. This is the coefficient of determination.

The coefficient of determination is the percentage of the variation that can be explained by the regression equation.

Every sample has some variation in it unless all the values are identical. The total variation is made up of two parts: the part that can be explained by the regression equation and the part that can't be explained by the regression equation.

Total variation = Explained variation + Unexplained variation

The **ratio** of the explained variation to the total variation is a measure of how good the regression line is. If the regression line passed through every point on the scatter plot exactly, it would be able to explain all of the variation. The further the line is from the points, the less it is able to explain.

The coefficient of determination

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

The coefficient correlation is the square root of the coefficient of determination

$$r = \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}}$$

An r^2 value of 0.95 means that 95% of the variation in y can be explained by the variation in x. The higher the value of r^2, the more confidence you can have in the equation of the regression line, so the more accurate it is for prediction.

Exercise 2

1 What proportion of the variance of y can be explained when
a $r^2 = 0.85$ **b** $r^2 = 0.75$ **c** $r^2 = 0.4$?

2 What proportion of the variance of y can be explained when
a $r = 0.3$ **b** $r = 0.9$ **c** $r = -0.4$?

3 The table below shows the increase in weight of a tomato crop (kg) for the amount of chemical (g) applied.

Chemical (g)	4.2	6.1	3.9	5.7	7.3	5.9
Increase (kg)	27.1	30.4	25.0	29.7	40.1	28.8

a Use your GDC to find the linear relationship between the amount of chemical (x) and the increase in weight (y).
b Write down the r^2 value.
c A typical report on the results would look like this:
A simple linear regression was performed on six crops of tomatoes to determine if there was a significant relationship between amount of chemical used and the weight of the crop. There was a positive significant relationship between the amount of chemical used and the weight of the crop. Furthermore, x% of the variability in the weight of the crop could be explained by the amount of chemical used.
Write down the value of x.

R^2 from nonlinear regression

The value R^2 quantifies goodness of fit. It is a fraction between 0.0 and 1.0, and has no units. Higher values indicate that the model fits the data better. You can interpret R^2 from nonlinear regression very much like you interpret r^2 from linear regression.
By tradition, statisticians use uppercase (R^2) for the results of nonlinear regression and lowercase (r^2) for the results of linear regression.

Exercise 3

1 A dye is absorbed into a patient's body to help with the X-ray process.
The table below shows the activity level for the first ten minutes.

Time (minutes)	0	1	2	3	4	5	6	7	8	9	10
Activity level	10030	8170	6794	5502	4486	3682	3060	2477	2045	1645	1328

a Sketch the graph of the data over the given interval.
b Write down the r^2 value and explain its significance.
c It is suggested that this would be better with an exponential function.
Use your GDC to find this function.
d Show your R^2 value for the exponential function and explain its significance.
e Validate your equation by substituting in the time and activity level at 3 minutes.
f When will the activity level drop to 500 counts per minute?

2 **Challenge**.

When using Microsoft Excel you may see the RMSE stated on graphs.

The root mean square error (RMSE) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated.

RMSE is a good measure of precision. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

Research and explain the RMSE.

Chapter 10 extension worked solutions

Exercise 1

1 a

Student	Rank in Mathematics	Rank in Science	d	d^2
A	1	3	–2	4
B	3	1	2	4
C	7	4	3	9
D	5	5	0	0
E	4	6	–2	4
F	6	9	–3	9
G	2	7	–5	25
H	10	8	2	4
I	9	10	–1	1
J	8	2	6	36
				$\Sigma d^2 = 96$

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(96)}{10(10^2 - 1)} = 1 - \frac{576}{990} \approx 0.418$$

b There is a moderate positive correlation. The sign of r_s shows the correlation is positive, i.e. there is some agreement between the ranks in maths and in science.

2 a

Country	Rank in income	Rank in infant death rate	d	d^2
A	1	10	–9	81
B	10	4	6	36
C	4	8	–4	16
D	8	5	3	9
E	7	3	4	16
F	2	11	–9	81
G	3	9	–6	36
H	5	6	–1	1
I	6	7	–1	1
J	11	1	10	100
K	9	2	7	49
				$\Sigma d^2 = 426$

$$r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6(426)}{11(11^2 - 1)} = 1 - \frac{2556}{1320} \approx -0.936$$

b This represents strong negative rank correlation between income and infant death rate, i.e. infant death rate tends to fall as income increases.

3

Hen	Eggs in January	Eggs in July	R_1	R_2	d	d^2
A	105	99	2	2	0	0
B	60	66	4	8.5	–4.5	20.25
C	76	91	3	3	0	0
D	45	67	6	7	–1	1
E	10	17	10	10	0	0
F	114	132	1	1	0	0
G	34	66	8	8.5	–0.5	0.25
H	9	12	11	11	0	0
I	44	70	7	6	1	1
J	52	83	5	4	1	1
K	29	72	9	5	4	16
						$\Sigma d^2 = 39.5$

If two numbers are the same, the ranks must be split between them. In this case 66 eggs are covering ranks 8 and 9, so both hens have rank number 8.5.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(39.5)}{11(11^2-1)} = 1 - \frac{237}{1320} \approx 0.820$$

b r_s of 0.8 shows that there is a strong, positive correlation between the two sets of data. So the hens who lay the most eggs in January also lay the most eggs in July.

Exercise 2

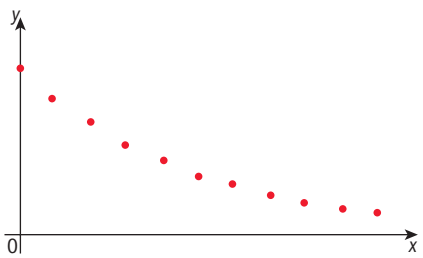
1 **a** 85% **b** 75% **c** 40%

2 **a** 9% **b** 81% **c** 16%

3 **a** $y = 9.87 + 3.68x$
b 0.807
c $x = 80.7$

Exercise 3

1 **a**



b $r^2 = 0.932$. 93.2% of the variability in the activity level could be explained by the time elapsed after the dye was inserted.
c $y = (10\,069)e^{-0.201x}$
d $R^2 = 0.9998$, 99.98% of the variability in the activity level could be explained by the time elapsed after the dye was inserted.
e $y = (10\,069)e^{-0.201(3)} = 5509 \approx 5502$
f $500 = (10\,069)e^{-0.201x}$

$$\frac{500}{10\,069} = e^{-0.201x}$$
$$\ln\frac{500}{10\,069} = \ln e^{-0.201x}$$
$$\ln\frac{500}{10\,069} = -0.201x$$
$$x = \left(\ln\frac{500}{10\,069}\right) \div -0.201$$
$$x = 14.94 \text{ min}$$