Spearman's Rank Correlation Coefficient (notes take 1 period)

Starter

1. **(Review of last lesson)**
   Is there evidence in the following car production figures to support the view that car production and steel production are negatively correlated? Justify your response.

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cars (000s)	9.8	19	27	32.1	43.8	41.1	37.8	43.1	61.1	78	71.8	55	25.3	17.8	7.6
Steel production	634	690	530	645	730	655	437	380	570	320	300	428	540	640	625

Notes

The product moment correlation coefficient is a good way to test for ***linear*** correlation when the samples come from populations that are normally distributed (i.e. the bell-shaped curve).

When the sample does not conform to the above, we can instead use Spearman's rank correlation coefficient as a means to test ***association*** i.e. data does not need to be linear. However, the data does need to be just increasing or just decreasing (see graphs on p78).

It is named after Charles Spearman, an English psychologist (1863-1945).

Correlation vs. association

Association — any relationship between two variables
Correlation — a linear relationship between two variables

Correlation vs. association

Formula and notation

Spearman's rank correlation coefficient is denoted by r_s and is found by:

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

N.B. The proof of this result is not needed for the course.

Similarly to PMCC, $-1 \le r_s \le 1$ where:

- $r_s = 1$ indicates perfect "agreement"
- $r_s = -1$ indicates perfect "disagreement"

N.B. This formula is given in the formula booklet
There is no way to calculate r_s directly on your calculator

E.g. 1 Five 200-metre runners kept a record of the times achieved in training in dry and wet conditions. The table records the average times.

Competitor	A	B	C	D	E
Dry, D	25.5	25.7	26.2	25.8	23.4
Wet, W	24.4	25.3	24.8	25.4	26.3

(a) Rank these times.
(b) Calculate the value of Spearman's rank correlation coefficient.

Working: (a)

Competitor	A	B	C	D	E
Dry, D	2	3	5	4	1
Wet, W	1	3	2	4	5

(b) The differences in ranks are: $1, 0, 3, 0, 4$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(1^2+0^2+3^2+0^2+4^2)}{5(5^2-1)}$$

$$r_s = 1 - \frac{6 \times 26}{125 - 5}$$

$$r_s = -0.3$$

Hypothesis test using Spearman's rank correlation coefficient

An hypothesis test using Spearman's rank correlation coefficient is very similar to one conducting using PMCC.

1. Calculate r_s.
2. State the null hypothesis: it is always "H_0 : there is no association between the two variables"
3. Decide whether you will carry out a 1—tail or 2—tail test (it is usually given in the question).
4. State the alternative hypothesis, H_1:
   - "H_1 : there is an association between the two variables" — two-tail test
   - "H_1 : there is a positive association between the two variables" — one-tail test
   - "H_1 : there is a negative association between the two variables" — one-tail test
5. Decide on the level of significance (it is usually given in the question)
6. Use the tables to find the critical value based on the number of values, n, and the significance level
7. Compare your r_s value to the critical value:
   - If $|r_s| >$ critical value then reject $H_0 \Rightarrow$ there is evidence to suggest there is an association between the two variables
   - If $|r_s| <$ critical value then do not reject $H_0 \Rightarrow$ there is no evidence to suggest there is an association between the two variables

E.g. 2 Two judges at a horticultural show placed entries in the 'miniature tree' class in the order shown in the table. Test, at the 5% level, if there is evidence of agreement between the judges.

Entry	A	B	C	D	E	F	G
Judge X	3	5	1	2	7	4	6
Judge Y	2	7	3	1	5	4	6

Working: The differences in ranks are: $1, 2, 2, 1, 2, 0, 0$

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(1^2+2^2+2^2+1^2+2^2+0^2+0^2)}{7(7^2-1)}$$

$$r_s = 1 - \frac{6\times 14}{343-7} = 0.75$$

The word '***agreement***' suggests we should test for ***positive association***.
i.e. a $1-$tail test
H_0 : there is no association between the judges' rankings
H_1 : there is an positive association between the judges' rankings
The 5% critical value for a $1-$tail test with 7 values is 0.7143
Since $r_s = 0.75 > 0.7143$, there is evidence to reject H_0
i.e. there is evidence to suggest the judges are in agreement

N.B. When finding the differences, it would not matter if we did $X - Y$ or $Y - X$.

E.g. 3 Ten golfers played the same course under different weather conditions and took the following number of shots for each round.

Golfer	A	B	C	D	E	F	G	H	I	J
Windless day	72	69	78	67	80	79	68	75	74	71
Windy day	72	76	77	70	74	82	73	79	84	81

(a) Rank these rounds, noting that in golf the fewer the shots the better.
(b) Calculate Spearman's rank correlation to for the data.
(c) Is there evidence at the 5% level that there is an association between the golfers relative performance in different conditions?

Dealing with ties (not in the OCR syllabus)
Suppose Judge X in the example above ranked entrant C and D equal top, what ranks would we give them?

1.5 each

When ranks are tied we allocate the average of the ranks.

Video: Spearman's Rank Correlation Coefficient
Video: Spearman's rank
Video: Spearman's rank and hypothesis testing example

Spearman's rank EQ

Solutions to Starter and E.g.s

Exercise
p79 5B Qu 1i, 2i, 3-7, (8 red)

Summary

Spearman's rank correlation: $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$

where d are the differences between the ranks and
n is the number of paired values

$-1 \leq r_s \leq 1$ where:

- $r_s = 1$ indicates perfect "agreement"
- $r_s = -1$ indicates perfect "disagreement"

Hypothesis test using Spearman's rank correlation coefficient

1. Calculate r_s.
2. State the null hypothesis: it is always "H_0 : there is no association between the two variables"
3. Decide whether you will carry out a 1—tail or 2—tail test (it is usually given in the question).
4. State the alternative hypothesis, H_1:
   - "H_1 : there is an association between the two variables" — two-tail test
   - "H_1 : there is a positive association between the two variables" — one-tail test
   - "H_1 : there is a negative association between the two variables" — one-tail test
5. Decide on the level of significance (it is usually given in the question)
6. Use the tables to find the critical value based on the number of values, n, and the significance level
7. Compare your r_s value to the critical value:
   - If $\left|r_s\right| >$ critical value then reject $H_0 \Rightarrow$ there is evidence to suggest there is an association between the two variables
   - If $\left|r_s\right| <$ critical value then do not reject $H_0 \Rightarrow$ there is no evidence to suggest there is an association between the two variables