Correlation Cheat Sheet

In this chapter you will learn about the product moment correlation coefficient and the Spearman's rank correlation coefficient. You will also be familiar with how to carry out hypothesis tests to assess correlation.

Product Moment Correlation Coefficient (PMCC)

The product moment correlation coefficient (PMCC), or r, tells you whether there is a linear correlation between two continuous variables. It is a suitable measure when both variables are normally distributed. It can be calculated using the formula below, which can be found in the formula booklet.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$$

The value of r can range from -1 to 1. The sign shows whether the correlation is positive or negative, and the number tells you how strong the correlation is. Here are some examples of interpretation of r:

r	Interpretation
-1	Perfect negative linear correlation
-0.8	Strong negative linear correlation
0	No correlation
0.4	Weak positive linear correlation
1	Perfect positive linear correlation

Note: When interpreting correlation always refer back to the context of the question.

Sometimes data can come is huge and messy numbers, and they can be coded to make calculations less complicated. Coding the data does not affect correlation as long as the coding is linear.

Example 1: A teacher wants to know whether there is a correlation between the marks from the physics exam and the marks from the chemistry exam. The marks for 10 students are recorded in the table below.

Physics (p)	73	89	56	55	85	98	88	63	47	93
Chemistry (c)	68	91	40	62	79	94	76	71	52	78

You can use $\Sigma p = 747$, $\Sigma c = 711$, $\Sigma c^2 = 53031$ and $\Sigma pc = 55499$ and $S_{pp} = 3010.1$

i) Find the product moment correlation coefficient.

Find S_{cc}.	$S_{cc} = \Sigma c^2 - \frac{(\Sigma c)^2}{n}$ $S_{cc} = 53031 - \frac{(711)^2}{10}$ $= 2478.9$
Find S_{pc}.	$S_{pc} = \Sigma pc - \frac{\Sigma p \Sigma c}{n}$ $S_{pc} = 55499 - \frac{747 \times 711}{10}$ $= 2387.3$
Find r using the formula.	$r = \frac{S_{pc}}{\sqrt{S_{pp} \times S_{cc}}}$ $= \frac{2387.3}{\sqrt{3010.1 \times 2478.9}}$ $= 0.874\ (3s.f.)$

ii) Giving a reason, state whether there is a correlation between the marks for Physics and Chemistry.

State the type of correlation (correlated or not, positive or negative, strong or weak) and explain using the value of r calculated earlier.	There is a strong positive correlation because the value of r is close to +1.

Spearman's Rank Correlation Coefficient

Spearman's Rank Correlation Coefficient can be used as an alternative when PMCC is not a suitable measure. This might be due to :

- Variables do not seem to have a linear correlation
- For data which are not continuous
- For data which do not have a normal distribution

The two sets of data are given a rank starting from 1. Rank 1 can be assigned to either the highest or lowest value, but you must rank both sets in a consistent way (i.e., highest as rank 1 in both or lowest as rank 1 in both). If more than 1 data has the same value, the mean of the ranks is assigned. For example, the 5th and 6th rank are tied, so the rank assigned to both will be $\frac{5+6}{2} = 5.5$.

The Spearman's rank correlation coefficient, r_s, can be calculated using the following formula:

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

where d = difference in ranks for each observation
n = number of pairs of observation

The interpretation of the Spearman's rank correlation is similar as the PMCC:

r_s	Interpretation
-1	Rankings are exactly opposite
-0.8	Rankings are mostly in disagreement
0	No correlation
0.4	Rankings agree to some extent
1	Rankings agree with each other completely

Remember to look back at the question and look at the context.

Example 2: In a competition, participants were ranked by 2 different judges, which are recorded in the table below.

	A	B	C	D	E	F	G	H	I	J
Judge 1	6	2	10	1	7	5	9	3	4	8
Judge 2	4	6	10	2	8	5	7	1	3	9

i) Using the Spearman's rank correlation, explain whether the judges agree with each other.

Calculate the difference in ranks (d) and their squares (d^2).	A: d = 2, d^2 = 4; B: -4, 16; C: 0, 0; D: -1, 1; E: -1, 1; F: 0, 0; G: 2, 4; H: 2, 4; I: 1, 1; J: -1, 1
Find the sum of d^2.	$\Sigma d^2 = 4 + 16 + 0 + 1 + 1 + 0 + 4 + 4 + 1 + 1 = 32$
Substitute d^2 into the formula to find r_s.	$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$ $= 1 - \frac{6 \times 32}{10(10^2-1)}$ $= 1 - \frac{192}{10(99)}$ $= 1 - 0.193939$ $= 0.806\ (3s.f.)$
Interpret your value of r_s.	The 2 judges are mostly in agreement with each other as the Spearman's rank correlation coefficient is close to +1.

ii) Judge 1 decided to swap the ranks of participant E and participant G. Without calculating the Spearman's rank correlation coefficient again, state how the correlation has changed.

Compare how similar the ranks given by the two judges to participants E and G are before and after the change.	Judge 1 (before): E = 7, G = 9 Judge 1 after: E = 9, G = 7 Judge 2: E = 8, G = 7
Since the ranks are more similar, d will decrease as well as Σd^2.	The r_s value will increase therefore the correlation will be stronger. The judges agree to a greater extent.

Hypothesis Testing

Hypothesis testing for zero correlation can be used to infer whether a correlation is likely in the population, based on whether a correlation is evident in the sample. Both PMCC and Spearman's rank correlation coefficient can be used, depending on which is more suitable for the data given.

The null hypothesis is always $H_0: \rho = 0$. It assumes that there is no correlation.

The alternative hypothesis for a one-tailed test is either $H_0: \rho < 0$ or $H_0: \rho > 0$. For a two-tailed test, it is $H_0: \rho \neq 0$.

You can find the critical values for the given significance level and sample size from a table in the formula booklet. There is a table for PMCC and another for Spearman's rank correlation coefficient.

Example 3: A teacher believes that the hours of sleep students have are associated with their exam performance. Using the data from the table below, perform a hypothesis test to show whether the teacher's belief is true at a significance level of 0.05.

Hours of sleep (x)	7	8.5	9	8	6	8	7.5
Exam score (y)	67	95	82	78	72	89	90

You can use $\Sigma x = 54$, $\Sigma y = 573$, $\Sigma x^2 = 422.5$, $\Sigma y^2 = 47527$ and $\Sigma xy = 4457.5$.

Write down your H_0 and H_1. Since the question does not specify what kind of association we are testing for, we use a two-tailed test.	$H_0: \rho = 0$ $H_1: \rho \neq 0$
Since this is a two-tailed test, we need to divide the significance level by 2 to find the significance level at each tail.	$0.05 \div 2 = 0.025$ $\therefore$ significance level at each tail: 0.025
Find the critical region for significance level: 0.025 and sample size: 7 using the table in the formula booklet.	$CV = \pm 0.7857$ Critical region: > 0.7857 and < -0.7857
Find S_{xy}.	$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$ $S_{xy} = 4457.5 - \frac{54 \times 573}{7}$ $= 37.214$
Find S_{xx}.	$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$ $S_{xx} = 422.5 - \frac{54^2}{7}$ $= 5.929$
Find S_{yy}.	$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$ $S_{yy} = 47527 - \frac{(573)^2}{7}$ $= 622.857$
Calculate your summary statistics.	$r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$ $= \frac{37.214}{\sqrt{5.929 \times 622.857}}$ $= 0.612\ (3s.f.)$
Check whether your summary statistic falls in the critical region.	$0.612 < 0.785$ Not in critical region
Interpret your results.	Accept H_0. 0.612 is not in the critical region. The teacher's belief is not supported at a significance level of 0.025.