The Product Moment Correlation Coefficient

Correlation coefficients measure how strongly two variables are linked. For this course you only need to know about the product moment correlation coefficient — and that's only so you can do a hypothesis test on it.

The **Product Moment Correlation Coefficient** measures correlation

The **Product Moment Correlation Coefficient** (**PMCC**, or r, for short) measures the **strength** of linear correlation between two variables — i.e. how close to a **straight line** the points on a scatter diagram lie.

The PMCC is always between +1 and –1:

- If all your points lie **exactly** on a **straight line** with a **positive gradient** (perfect positive correlation), $r = +1$.
- If all your points lie **exactly** on a **straight line** with a **negative gradient** (perfect negative correlation), $r = -1$.
- If the variables **aren't correlated**, $r = 0$ (or more likely, pretty close to 0).

Example: Marcus records the amount of exercise (in hours) that 30 people do in a month, and the length of time (in mins) it takes them to complete a lap of a cycle route. He calculates the product moment correlation coefficient of his data to be $r = -0.866$. Interpret this result in context.

r is close to –1, so this means the data has a **strong negative correlation**.

This shows that as one variable increases, the other decreases.

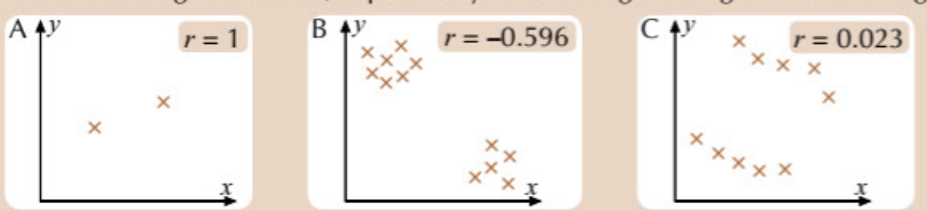
This suggests that the **more** hours of exercise a person does, the **less** time it takes them to cycle the route (i.e. they can cycle quicker).

Use the context of the question.

Always say that the data 'suggests' a conclusion like this. It's based on a sample, so you can't say for sure that it proves anything about the population.

You need to be careful with your conclusion. There isn't enough information to conclude that one variable necessarily **causes** the change — there might be **another** factor that has an effect on **both** variables, such as age.

Example: For the following sets of data, explain why the PMCC given might be misleading.



In A, there are only **two data points**, so r would always be +1 or –1 (you can always draw a straight line between two points).
This **doesn't** tell you anything about the relationship between the variables.

The scatter graph of B shows two **clusters** of data — there appears to be negative correlation overall, but none within the clusters. Similarly, C contains two data **clusters**, each with negative correlation, despite r suggesting virtually no correlation. Establishing what has caused the two clusters in each case and separating the data would lead to better conclusions.

PMCC Hypothesis Testing tests the value of r

To test whether your value of r is likely to mean that the two variables are **actually correlated**, you need to do a **hypothesis test**. The method used is like the hypothesis testing you did in Section 14, but with a few differences:

1) The **test statistic** is r (which is calculated from sample data for you).
2) The **population parameter** is the PMCC of the population, ρ (the Greek letter rho).
3) The **null hypothesis** is always that there is no correlation between the two variables — $H_0: \rho = 0$.
4) There are two kinds of **alternative hypothesis** — for a **one-tailed** test, $H_1: \rho > 0$ or $\rho < 0$
   for a **two-tailed** test, $H_1: \rho \neq 0$.

In a PMCC hypothesis testing question, you could be given a **table of critical values** and be expected to use the **sample size** and **significance level** to choose the right value. Or if the examiners are feeling kind they'll just give you the critical value instead.

The Product Moment Correlation Coefficient

Example: a) A teacher claims that test scores and hours spent revising are positively correlated. State the null and alternate hypotheses for a test of the teacher's claim.

- The null hypothesis is that there is no correlation, so H_0: $\rho = 0$
- This is a **one-tailed** test, so the alternative hypothesis is H_1: $\rho > 0$

b) The teacher samples 10 students and finds that the PMCC is 0.76. Using the table of critical values provided, carry out a hypothesis test at the 5% significance level to investigate whether this result is significant.

Sample Size	Significance Level		
	0.10	0.05	0.025
4	0.8000	0.9000	0.9500
⋮	⋮	⋮	⋮
9	0.4716	0.5822	0.6664
10	0.4428	0.5494	0.6319
11	0.4187	0.5214	0.6021

A table or critical value will be given in the exam if needed.

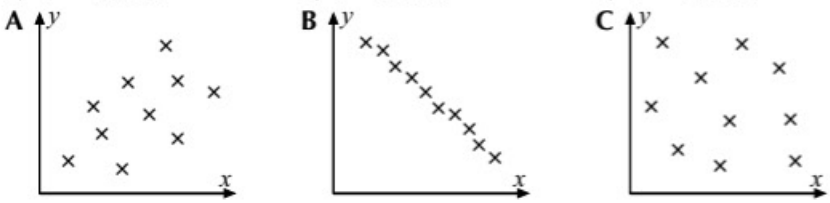
- Test for significance using the significance level column $\alpha = 0.05$ and the sample size row 10:
  Using the table provided, the critical value is **0.5494**, so you would **reject H_0 if $r \geq 0.5494$**.
  Since $0.76 > 0.5494$ the result is **significant**.
- Write your conclusion — you either reject the null hypothesis H_0 or have insufficient evidence to do so:
  There is evidence at the 5% level of significance to reject H_0 and to support the alternative hypothesis that test scores and hours spent revising are positively correlated.

For more on hypothesis testing, look at pages 176-177.

Practice Questions

1 Interpret the following values of r and match each one to the most suitable set of data:

a) $r = -0.912$ b) $r = 0.431$ c) $r = -0.041$



2 A shop owner records the number of pairs of sunglasses and ice creams that are sold each day in his shop, over one year. The product moment correlation coefficient is calculated to be $r = 0.786$.

a) Interpret the value of r in context.

b) The shop owner says 'if there's positive correlation, then that means high sunglasses sales cause high ice cream sales'. Comment on the shop owner's claim.

Exam Questions

Q1 The age of 50 adults and their time taken to do a times table test are measured. The product moment correlation coefficient is calculated to be $r = -0.24$. Stating your hypotheses clearly, carry out a hypothesis test, at the 2.5% significance level, to investigate whether the evidence suggests that age and time taken to do the test are negatively correlated, given that the critical value for such a test is -0.2787. [2 marks]

Q2 The diameter and the weight of 8 randomly selected biscuits are measured and the product moment correlation coefficient is calculated to be $r = 0.958$.

a) Using the table of critical values provided, find the critical region for a test at the 0.5% significance level of whether the diameter and weight of the biscuits are positively correlated. State your hypotheses clearly. [2 marks]

b) Comment on the significance of r at the 0.5% significance level. [1 mark]

Sample Size	Significance Level		
	0.025	0.01	0.005
4	0.9500	0.9800	0.9900
⋮	⋮	⋮	⋮
7	0.7545	0.8329	0.8745
8	0.7067	0.7887	0.8343
9	0.6664	0.7498	0.7977

It's fun to study the P.M.C.C...

You only need to be able to interpret the value of r (not find it) — remember numbers close to +1 or −1 suggest strong positive or negative correlation (respectively) and numbers close to 0 mean no correlation. This is also a good time to check that you're comfortable with the theory behind hypothesis testing — if not, flick back to Section 14 and rejoice.