KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12
Applied Mathematics
S2 Regression Correlation and hypothesis testing

HGS Maths

Dr Frost Course

Name: ______________________________

Class: ______________________________

Contents

Prior knowledge check

Prior knowledge check

1 Given that $y = 3 \times 2^x$

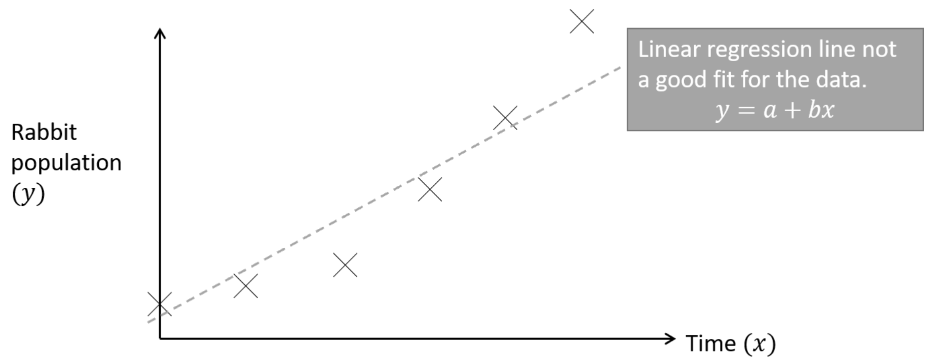
a show that $\log y = A + Bx$, where A and B are constants to be found.

b The straight-line graph of x against $\log y$ is plotted. Write down the gradient of the line and the intercept on the vertical axis. ← **Pure Year 1, Chapter 12**

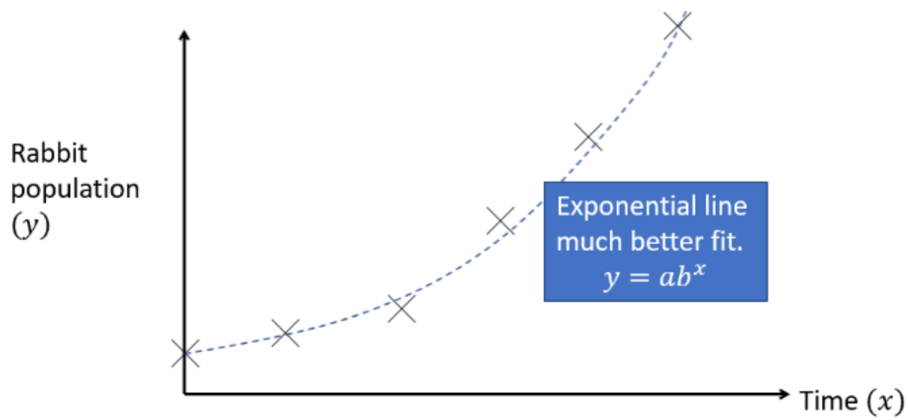
2 The height, h cm, and handspan, s cm, of 20 students are recorded. The regression line of h on s is found to be $h = 22 + 11.3s$. Give an interpretation of the value 11.3 in this model. ← **Year 1, Chapter 4**

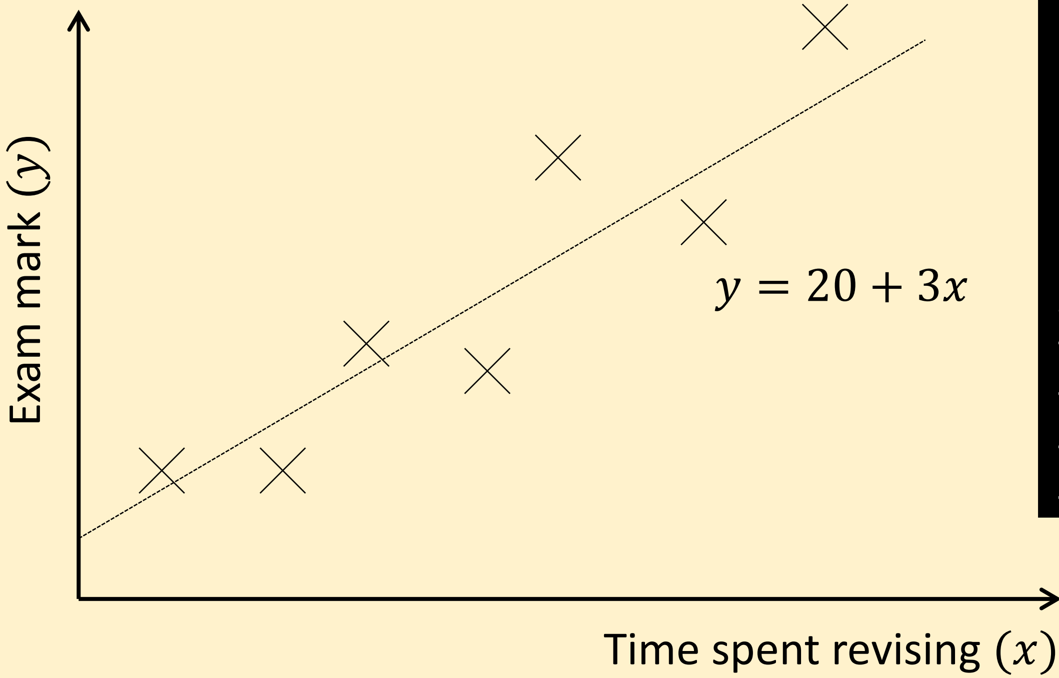
1.1) Exponential models

Linear regression sometimes doesn't quite fit the data neatly, as shown for the example of rabbit population vs time:



Instead, an exponential model is a better fit:

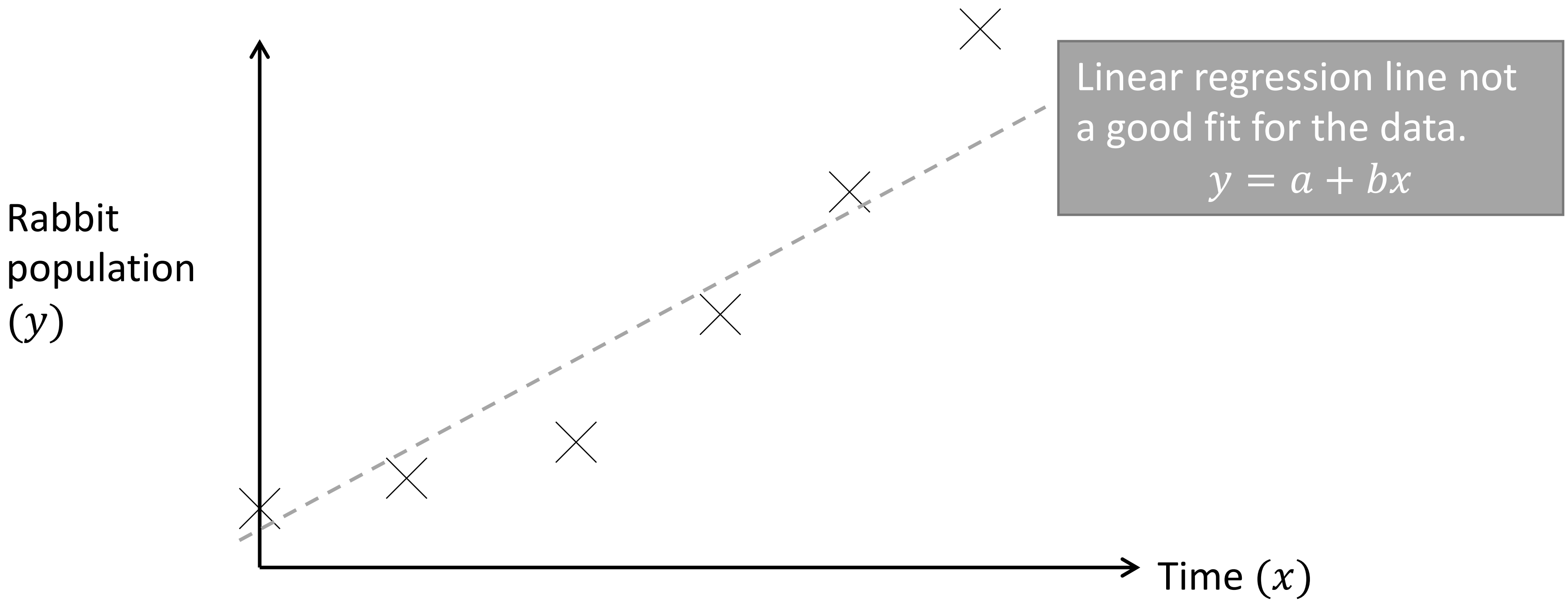




What we've done here is come up with a **model** to explain the data, in this case, a line $y = a + bx$. We've then tried to set a and b such that the resulting y value matches the actual exam marks as closely as possible.
The 'regression' bit is the act of setting the parameters of our model (here the gradient and y-intercept of the line of best fit) to best explain the data.

I record people's exam marks as well as the time they spent revising. I want to predict how well someone will do based on the time they spent revising. How would I do this?

Exponential Regression



For some variables, e.g. population with time, it may be more appropriate to use an **exponential** equation, i.e. $y = ab^x$, where a and b are constants we need to fix to best match the data.

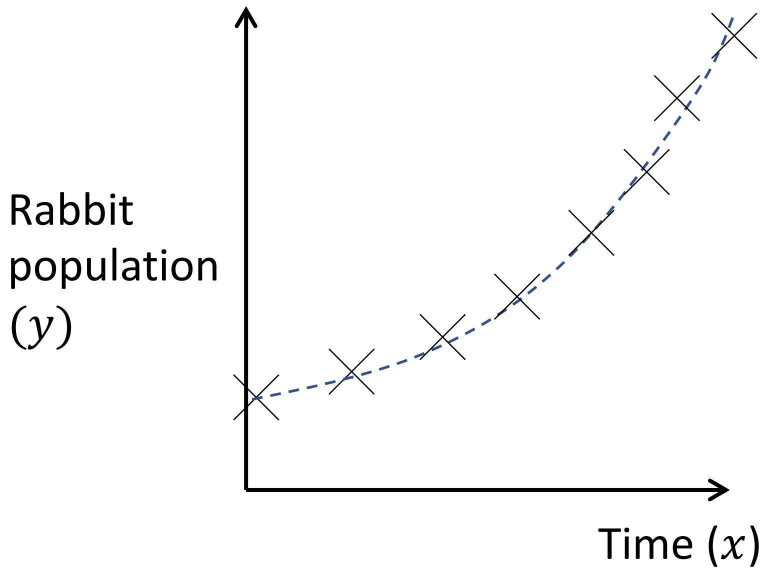
$$y = ab^x$$
$$\log y = \log(ab^x)$$
$$\log y = \log a + x \log b$$

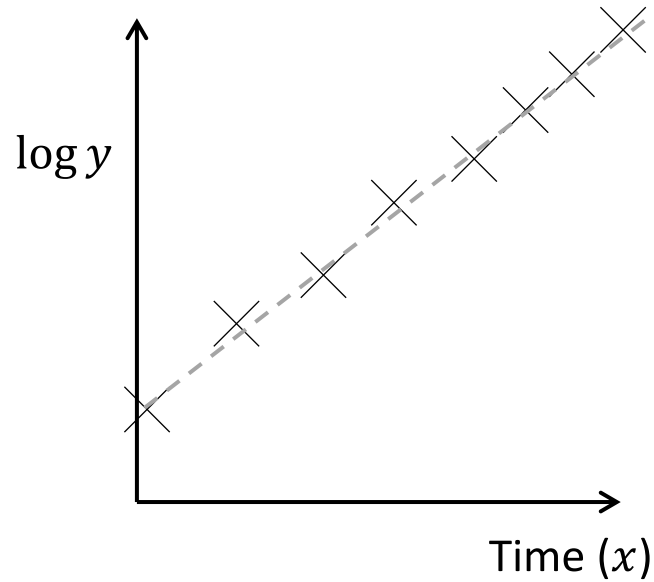
In Year 1, what did we do to both sides to end up with a straight line equation?

If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$

Exponential Regression

If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$





Comparing the equations, we can see that if we log the y values (although leave the x values), the data then forms a straight line, with y-intercept $\log k$ and gradient $\log b$.

Notes

Worked Example

The table shows some data collected on the temperature, in °C, of a colony of bacteria (t) and its growth rate (g).

Temperature, t (°C)	3	5	6	8	9	11
Growth rate, g	1.40	1.94	1.97	2.85	3.2	4.64

The data are coded using the changes of variable $x = t$ and $y = \log g$. The regression line of y on x is found to be $y = -0.0536 + 0.0637x$.

a) Find the initial growth rate
b) Given that the data can be modelled by an equation of the form $g = kb^t$ where k and b are constants, find the values of k and b.

Your Turn

The table shows some data collected on the temperature, in °C, of a colony of bacteria (t) and its growth rate (g).

Temperature, t (°C)	3	5	6	8	9	11
Growth rate, g	1.04	1.49	1.79	2.58	3.1	4.46

The data are coded using the changes of variable $x = t$ and $y = \log g$. The regression line of y on x is found to be $y = -0.2215 + 0.0792x$.

a) Find the initial growth rate

b) Given that the data can be modelled by an equation of the form $g = kb^t$ where k and b are constants, find the values of k and b.

a) 0.6

b) $k = 0.6,\ b = 1.20$ (3 sf)

Worked Example

A rabbit population, P, is modelled with respect to time in years, t. An exponential model is proposed:

$$P = kb^t$$

The data is coded using $x = t$ and $y = \log P$.

The regression line of y on x is found to be $y = 3 + 0.2x$. Determine the values of k and b.

T.3: Ex 1A all Qs, P.1 1.1 all Qs

Your Turn

A rabbit population, P, is modelled with respect to time in years, t. An exponential model is proposed:

$$P = kb^t$$

The data is coded using $x = t$ and $y = \log P$.

The regression line of y on x is found to be $y = 2 + 0.3x$. Determine the values of k and b.

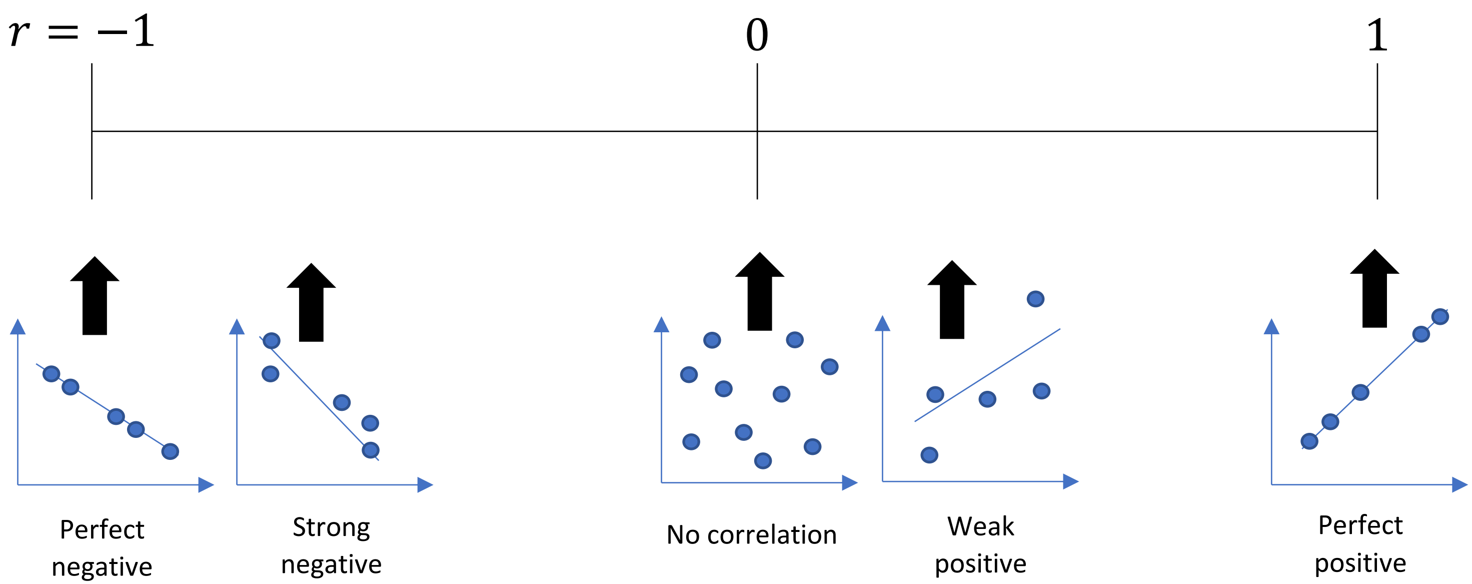
$k = 100, b = 2.00$ (3 sf)

1.1) Exponential models

You're used to use qualitative terms such as "positive correlation" and "negative correlation" and "no correlation" to describe the **type** of correlation, and terms such as "perfect", "strong" and "weak" to describe the **strength**.

The **Product Moment Correlation Coefficient** is one way to quantify this:

✎ The product moment correlation coefficient (PMCC), denoted by r, describes the linear correlation between two variables. It can take values between -1 and 1.

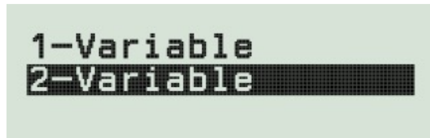


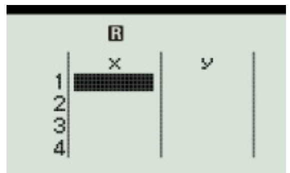
Rule of thumb: $r < -0.7$ or $r > 0.7$ is considered to be 'strong' correlation.

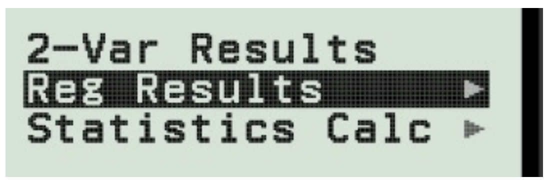
Note that PMCC is only applicable for a <u>linear</u> correlation, i.e. closeness of fit to a linear regression line (i.e. a <u>straight</u> 'line of best fit'). It may be the data exhibits strong correlation with respect to a different model (e.g. exponential) even when the PMCC is low.

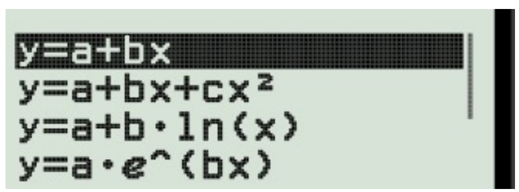
Calculating r on your calculator

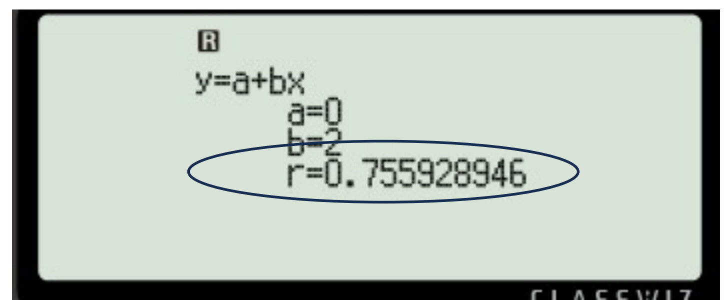
1. Press HOME
2. Select Statistics
3. Select 2 variable
4. Type in xy data
5. Press EXE
6. Select Reg Results
7. Choose y=a+bx (linear)
8. Read r value











Notes

Worked Example

Calculate the product moment correlation coefficient for the following data:

x	y
1	3
2	4
3	5
4	8

T.6: Ex 1B Qs 1-3, P.3 1.2 Qs 1-4

Your Turn

Calculate the product moment correlation coefficient for the following data:

x	y
1	3
2	6
3	5
4	8

$r = 0.868$ (3 sf)

Worked Example

From the large data set, the daily mean temperature, t °C, and the daily total rainfall, r mm, were recorded from 27th May to 5th June inclusive 1987 in Leuchars.

Day	1	2	3	4	5	6	7	8	9	10
t	8.5	9.0	10.3	12.8	13.5	12.8	9.8	8.8	10.0	10.4
r	0	2.4	8.1	0.2	0.4	tr	6.1	3.6	tr	31.8

a) State the meaning of tr in the table above.

b) Calculate the product moment correlation coefficient for the ten days, stating clearly how you deal with the 'tr' readings.

c) With reference to your answer to part b, comment on the suitability of a linear regression model for these data.

T.6: Ex 1B Qs 4+, P.3 1.2 Qs 5+

Your Turn

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 10 days in September in Hurn in 1987.

Day	1	2	3	4	5	6	7	8	9	10
w	4	4	8	7	12	12	3	4	7	10
g	13	12	19	23	33	37	10	n/a	n/a	23

a) State the meaning of n/a in the table above.
b) Calculate the product moment correlation coefficient for the remaining 8 days.
c) With reference to your answer to part b, comment on the suitability of a linear regression model for these data.

a) Data on daily maximum gust is not available for these days
b) $r = 0.9533$ (4 sf)
c) r is close to 1 so there is strong positive correlation between daily mean windspeed and daily maximum gust. This means that the data points lie close to a straight line, so a linear regression model is suitable.

1.3) Hypothesis testing for zero correlation

- Hypothesis testing can be best thought of as a court case.
- The person that is standing on trial is either found guilty or not guilty.
- The "default" position is that the person is not guilty as if there isn't enough evidence against the person, then the jury have to side on the side of caution and not convict.
- The stronger the evidence is against the person, the more the jury will start to believe that the person is in fact guilty.

H_0 is called the null hypothesis. (Think of this as "no change" or the default position.)

H_0 is the hypothesis that you assume to be correct (the jury assume that the person is not guilty until substantial evidence is provided to suggest otherwise).

H_1 is the alternative hypothesis (i.e. that the person is guilty)

Notes

We can use a hypothesis test to determine whether the product moment correlation coefficient for a particular sample indicates that there is likely to be a **linear relationship** within the whole population or not.

Here is how we do the test:

1. Assume innocence: there is no correlation for the whole data set (population)
2. Measure *r* (correlation) of a sample of size *n*.
3. Check if this *r value* is **extreme enough** to conclude there is a correlation of some sort (enough evidence of guilt), otherwise assume innocence (the default of no correlation).

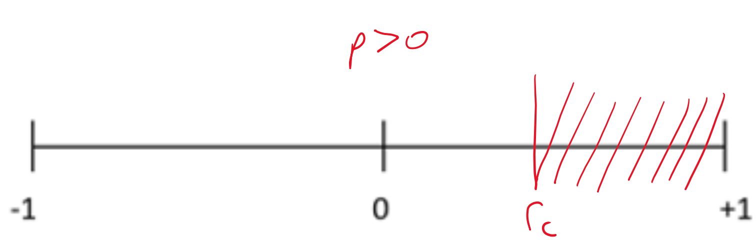
$$\rho\ (rho) is\ the\ PMCC\ for\ the\ population, r\ is\ the\ PMCC\ for\ a\ sample$$

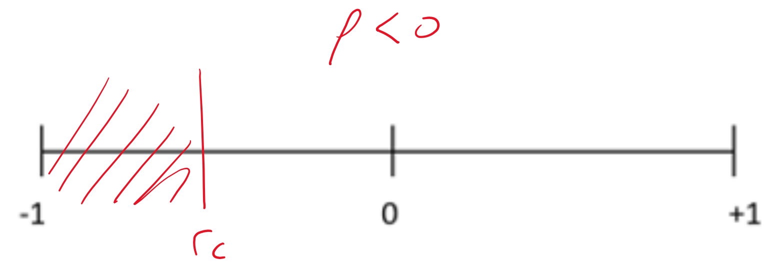
The 'evidence' we refer to is the *r* from the SAMPLE, and it used to conclude something about the ρ of the population.

Notes

One-tailed hypothesis test

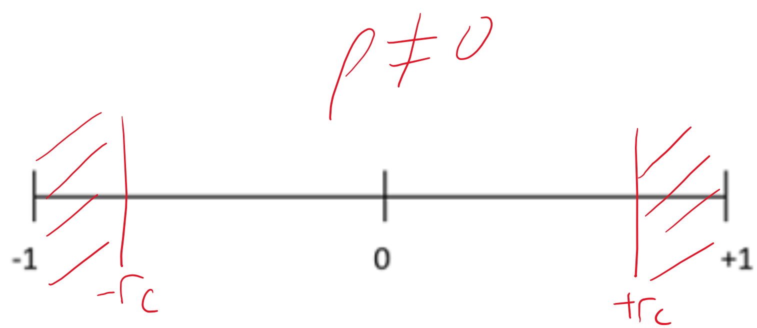
- This type of test involves an alternative hypothesis that solely relates to either ρ indicating a positive or a negative correlation
- H_1: $\rho > 0$ or $\rho < 0$ (one way or the other).





Two-tailed hypothesis test

- This type of test involves an alternative hypothesis that relates to ρ indicating any possible correlation at all
- H_1: $\rho \neq 0$ (either way).



Notes

Let us go back to the court case example…

- There have been cases where a jury have convicted wrong and the person is actually innocent!
- This "risk" of convicting wrong is known as a significance level which is denoted by α.
- In industrial environments, it is normally 5% or 1% as industry wants a good chance that they have concluded correctly.

Critical value tables

Earlier it was stated that we only start to believe '$guilt$' when substantial evidence is given to us, but what is substantial evidence? We have critical values that depend on sample size and how sure we want to be that we aren't making a mistake (i.e. the chance that the jury makes the mistake of convicting when the person is innocent).

CRITICAL VALUES FOR CORRELATION COEFFICIENTS

These tables concern tests of the hypothesis that a population correlation coefficient ρ is 0. The values in the tables are the minimum values which need to be reached by a sample correlation coefficient in order to be significant at the level shown, on a one-tailed test.

Product Moment Coefficient					Sample	Spearman's Coefficient		
		Level					Level	
0.10	0.05	0.025	0.01	0.005	Level	0.05	0.025	0.01
0.8000	0.9000	0.9500	0.9800	0.9900	4	1.0000	-	-
0.6870	0.8054	0.8783	0.9343	0.9587	5	0.9000	1.0000	1.0000
0.6084	0.7293	0.8114	0.8822	0.9172	6	0.8286	0.8857	0.9429
0.5509	0.6694	0.7545	0.8329	0.8745	7	0.7143	0.7857	0.8929
0.5067	0.6215	0.7067	0.7887	0.8343	8	0.6429	0.7381	0.8333
0.4716	0.5822	0.6664	0.7498	0.7977	9	0.6000	0.7000	0.7833
0.4428	0.5494	0.6319	0.7155	0.7646	10	0.5636	0.6485	0.7455
0.4187	0.5214	0.6021	0.6851	0.7348	11	0.5364	0.6182	0.7091
0.3981	0.4973	0.5760	0.6581	0.7079	12	0.5035	0.5874	0.6783
0.3802	0.4762	0.5529	0.6339	0.6835	13	0.4835	0.5604	0.6484
0.3646	0.4575	0.5324	0.6120	0.6614	14	0.4637	0.5385	0.6264

The value of 0.6215 (sample size of 8 and a 5% significance level) means this value is the minimum that our r value can be in order for us to conclude that we believe that there is sufficient evidence of positive correlation.

The table works the other way for testing negative correlation by placing a negative in front of the numbers.

The value of -0.6215 (sample size of 8 and a 5% significance level) means this value is the maximum that our r value can be in order for us to conclude that we believe that there is sufficient evidence of negative correlation.

Definitions

Definitions – Examples for testing PMCC

Null hypothesis: H_0, is the hypothesis that the population product moment correlation coefficient is zero, i.e. there is no correlation between daily mean windspeed and daily maximum temperature.

It is important that candidates understand they are using a sample to make inferences about the population. The null and alternative hypotheses always refer to the population.

Alternative hypothesis: H_1, is the hypothesis that the population product moment correlation coefficient is either different to zero, greater than zero or less than zero. In other words, the sample observations are influenced by some non-random cause.

	H_1 One- or two-tailed	H_1
Test for a difference or a change	Two-tailed	H_1: $\rho \neq 0$
Test for an increase or greater than	One-tailed	H_1: $\rho > 0$
Test for a decrease, Reduction or less than	One-tailed	H_1: $\rho < 0$

The significance level of a hypothesis test is the probability of rejecting the null hypothesis given that it is true.

Test statistics: In hypothesis testing, the test statistic is a value computed from sample data. The test statistic is used to assess the strength of evidence in support of a null hypothesis.

Critical value: In hypothesis testing, a critical value is a value that is compared to the test statistic to determine whether to reject the null hypothesis. If the absolute value of your test statistic is greater than (or less than when considering a negative test statistic) the critical value, you can declare statistical significance and reject the null hypothesis.

Critical region: The rejection region for the null hypothesis in the testing of a hypothesis.

Acceptance region: The rejection region for the alternative hypothesis in the testing of a hypothesis.

***p*-value**: Defined informally as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.

Conducting the test

1. State hypotheses

H_0: $\rho = 0$

H_1:

- $\rho > 0$ → Simply read value from table corresponding to correct sig. level and sample size *n*
- $\rho < 0$ → Do as above but make negative
- $\rho \neq 0$ → Half sig. level and put a $\pm$

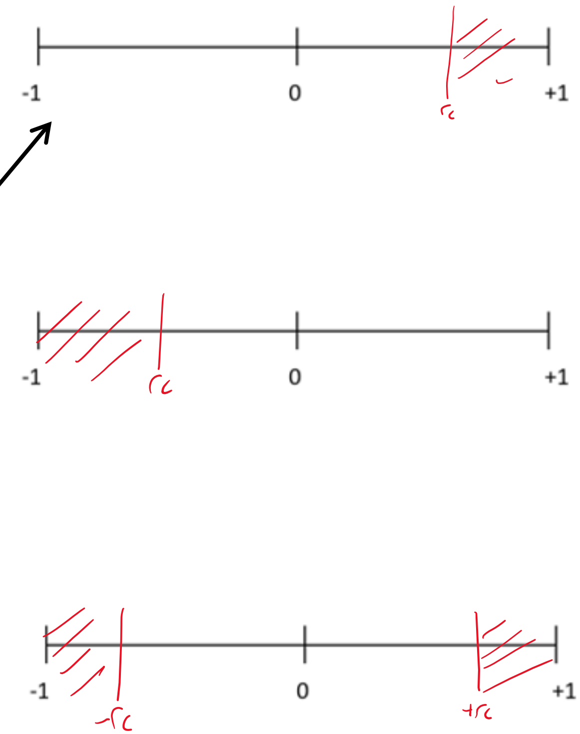
2. Find critical value from table:

Simply read value from table corresponding to correct sig. level and sample size *n*

Do as above but make negative

Half sig. level and put a $\pm$

3. Diagram and comparison:



Concluding the test

1. State hypotheses

H_0: $\rho = 0$

$\rho > 0$

H_1: $\rho < 0$

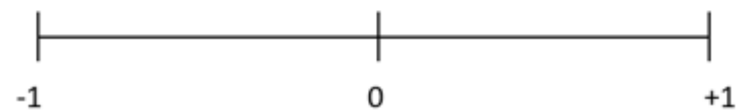
$\rho \neq 0$

4. Conclusion in context:

Accept H_0 , there is **insufficient** evidence to suggest [insert wording from question] at the …% significance level

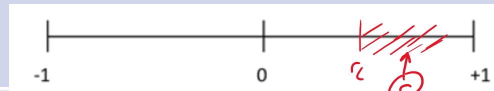
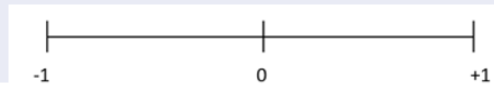
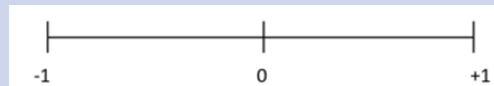
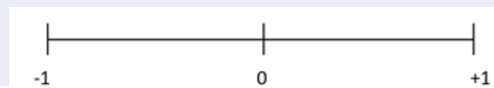
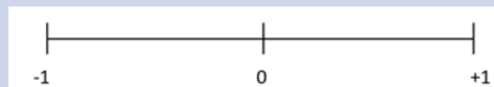
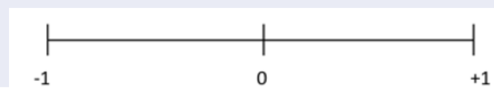
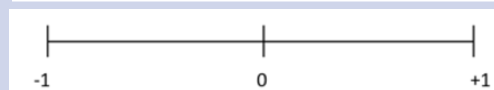
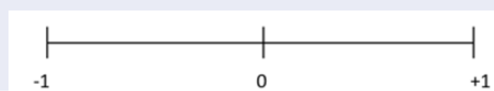
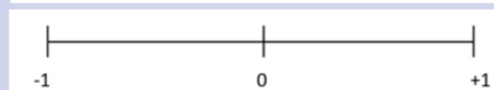
Reject H_0 , there is **sufficient** evidence to suggest [insert wording from question] at the …% significance level

Writing Frame to use

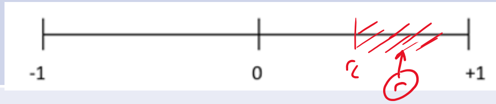
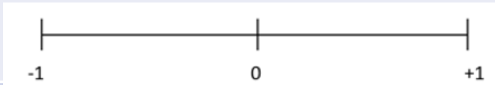
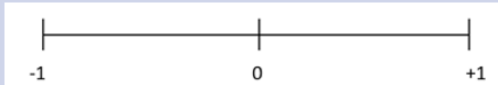
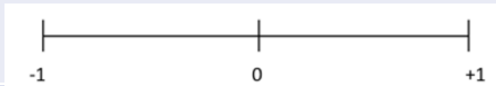
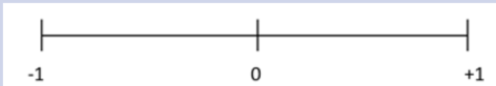
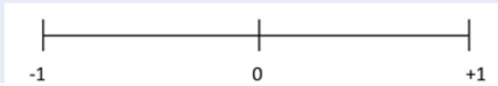
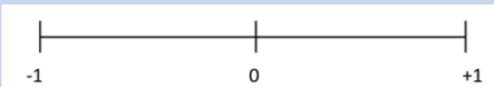
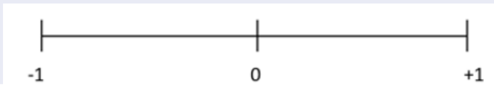
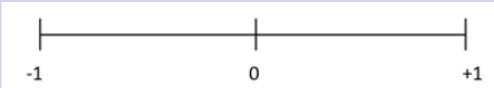
State hypotheses	$H_0\!: \rho = 0$ $\qquad$ $H_1\!: \rho \quad 0$
Find critical value (r_c) from table	$r_c =$
Draw diagram	
Inequality test stat r against r_c	
Accept/reject H_0	
Conclusion in context – *using wording from Q*	there is __________ evidence to suggest ____ ____ at the___ % significance level

Writing frame word doc in folder in booklets section

Fill in the blank

H_0	H_1	Test stat r	Sample size n	Sig. level %	Critical value r_c	Diagram	inequality	Accept/reject H_0
$\rho = 0$	$\rho > 0$	$r = 0.6$	10	5	0.5494		$r > r_c$ $0.6 > 0.5494$	reject H_0
$\rho = 0$	$\rho > 0$	$r = 0.6$	10	1				
$\rho = 0$	$\rho > 0$	$r = 0.6$	22	2.5				
$\rho = 0$	$\rho < 0$	$r = -0.5$	10	10				
$\rho = 0$	$\rho < 0$	$r = -0.5$	10	1				
$\rho = 0$	$\rho < 0$	$r = -0.5$	25	0.5				
$\rho = 0$	$\rho \neq 0$	$r = -0.5$	12	10				
$\rho = 0$	$\rho \neq 0$	$r = 0.5$	12	5				
$\rho = 0$	$\rho \neq 0$	$r = -0.5$	12	1				

Fill in the blank

H_0	H_1	Test stat r	Sample size n	Sig. level %	Critical value r_c	Diagram	inequality	Accept/reject H_0
$\rho = 0$	$\rho > 0$	$r = 0.6$	10	5	0.5494		$r > r_c$ $0.6 > 0.5494$	reject H_0
$\rho = 0$	$\rho > 0$	$r = 0.6$	10	1	0.7155			
$\rho = 0$	$\rho > 0$	$r = 0.6$	22	2.5	0.4227			
$\rho = 0$	$\rho < 0$	$r = -0.5$	10	10	-0.4428			
$\rho = 0$	$\rho < 0$	$r = -0.5$	10	1	-0.7155			
$\rho = 0$	$\rho < 0$	$r = -0.5$	25	0.5	-0.5052			
$\rho = 0$	$\rho \neq 0$	$r = -0.5$	12	10	± 0.4973			
$\rho = 0$	$\rho \neq 0$	$r = 0.5$	12	5	± 0.5760			
$\rho = 0$	$\rho \neq 0$	$r = -0.5$	12	1	± 0.7079			

Worked Example

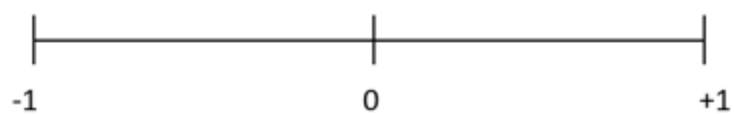
A scientist takes 19 observations of the masses of two reactants in an experiment.
She calculates a product moment correlation coefficient of $r = 0.54$.

The scientist believes there is a positive correlation between the masses of the two reactants.
Test at the 1% level of significance, the scientist's claim, stating your hypotheses clearly.

Your Turn

A scientist takes 14 observations of the masses of two reactants in an experiment.
She calculates a product moment correlation coefficient of $r = -0.45$.

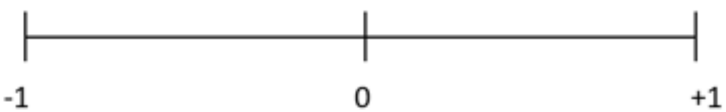
The scientist believes there is a negative correlation between the masses of the two reactants.
Test at the 5% level of significance, the scientist's claim, stating your hypotheses clearly.

State hypotheses	$H_0\colon \rho = 0 \qquad H_1\colon \rho \quad 0$
Find critical value (r_c) from table	$r_c =$
Draw diagram	
Inequality test stat r against r_c	
Accept/reject H_0	
Conclusion in context – *using wording from Q*	there is _________ evidence to suggest ____ ____ at the___ % significance level

Worked Example

A scientist takes 20 observations of the masses of two reactants in an experiment.
She calculates a product moment correlation coefficient of $r = 0.54$.

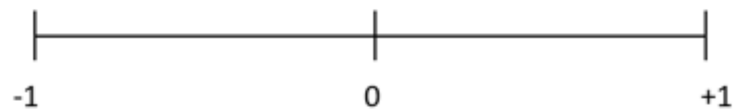
The scientist believes there is no correlation between the masses of the two reactants.
Test at the 1% level of significance, the scientist's claim, stating your hypotheses clearly.

State hypotheses	H_0: $\rho = 0$ $\qquad$ H_1: $\rho \quad 0$
Find critical value (r_c) from table	$r_c =$
Draw diagram	
Inequality test stat r against r_c	
Accept/reject H_0	
Conclusion in context – *using wording from Q*	there is ________ evidence to suggest ______________________________________ ______________________________________ at the___ % significance level

Your Turn

A scientist takes 30 observations of the masses of two reactants in an experiment.
She calculates a product moment correlation coefficient of $r = -0.45$.

The scientist believes there is no correlation between the masses of the two reactants.
Test at the 10% level of significance, the scientist's claim, stating your hypotheses clearly.

State hypotheses	H_0: $\rho = 0$ $\qquad$ H_1: $\rho \quad 0$
Find critical value (r_c) from table	$r_c =$
Draw diagram	
Inequality test stat r against r_c	
Accept/reject H_0	
Conclusion in context – *using wording from Q*	there is __________ evidence to suggest ____ ____ at the___ % significance level

Worked Example

The table from the large data set shows the daily mean temperature, t °C, and the daily total rainfall, r mm, in Leuchars for a sample of nine days in October 1987.

$\boldsymbol{t}$	11.4	10.5	6.5	8.3	8.2	5.7	7.6	12.1	11.2
$\boldsymbol{r}$	0	1	3.9	16.3	7.9	4.1	15.2	0	tr

Test, at the 10% level of significance, whether there is evidence of a negative correlation between daily mean temperature and daily total rainfall.
State your hypotheses clearly.

T.10: Ex 1C all Qs 1-3, P.5 1.3 all Qs

Your Turn

The table from the large data set shows the daily maximum gust, x knots, and the daily maximum relative humidity, $y\%$, in Leeming for a sample of eight days in May 2015.

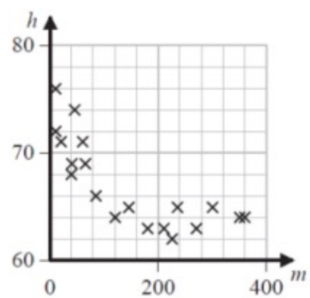
x	31	28	38	37	18	17	21	29
y	99	94	87	80	80	89	84	86

Test, at the 10% level of significance, whether there is evidence of a positive correlation between daily maximum gust and daily maximum relative humidity. State your hypotheses clearly.

6. Anna is investigating the relationship between exercise and resting heart rate. She takes a random sample of 19 people in her year at school and records for each person

- their resting heart rate, h beats per minute
- the number of minutes, m, spent exercising each week

Her results are shown on the scatter diagram.



Anna codes the data using the formulae

$$x = \log_{10} m$$

$$y = \log_{10} h$$

The product moment correlation coefficient between x and y is – 0.897

(*b*) Test whether or not there is significant evidence of a negative correlation between x and y
You should

- state your hypotheses clearly
- use a 5% level of significance
- state the critical value used

(3)

(b)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	Critical value – 0.3887 (Allow $\pm$)	M1	1.1b
	There is evidence that the product moment **correlation** is **less than 0/ there is a negative correlation**	A1	2.2b
		(3)	

(b)	**B1**	Both hypotheses correct in terms of ρ (allow p)
	M1	For the cv of -0.3887 or any cv such that $0.3 < \lvert cv \rvert < 0.5$
	A1	Independent of hypotheses. Correct conclusion that implies reject H_0 on basis of seeing –0.3887 or if they give 0.3887 we must see the comparison $0.3887 < 0.897$ **and** which mentions "pmcc/correlation/relationship" and less than 0/ negative or $\rho < 0$ A contradictory statement scores A0 eg Accept H_0 therefore negative correlation

Learning from exams

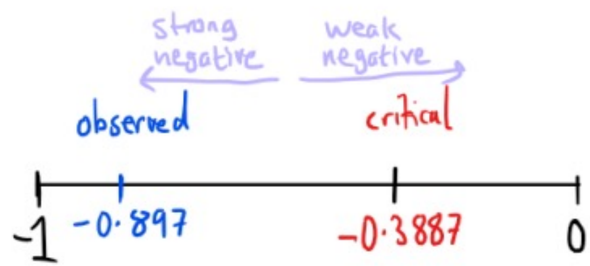
Product Moment Coefficient					
Level					Sample size, n
0.10	0.05	0.025	0.01	0.005	
0.8000	0.9000	0.9500	0.9800	0.9900	4
0.6870	0.8054	0.8783	0.9343	0.9587	5
0.6084	0.7293	0.8114	0.8822	0.9172	6
0.5509	0.6694	0.7545	0.8329	0.8745	7
0.5067	0.6215	0.7067	0.7887	0.8343	8
0.4716	0.5822	0.6664	0.7498	0.7977	9
0.4428	0.5494	0.6319	0.7155	0.7646	10
0.4187	0.5214	0.6021	0.6851	0.7348	11
0.3981	0.4973	0.5760	0.6581	0.7079	12
0.3802	0.4762	0.5529	0.6339	0.6835	13
0.3646	0.4575	0.5324	0.6120	0.6614	14
0.3507	0.4409	0.5140	0.5923	0.6411	15
0.3383	0.4259	0.4973	0.5742	0.6226	16
0.3271	0.4124	0.4821	0.5577	0.6055	17
0.3170	0.4000	0.4683	0.5425	0.5897	18
0.3077	0.3887	0.4555	0.5285	0.5751	19
0.2992	0.3783	0.4438	0.5155	0.5614	20
0.2914	0.3687	0.4329	0.5034	0.5487	21
0.2841	0.3598	0.4227	0.4921	0.5368	22

The product moment correlation coefficient between x and y is – 0.897

(*b*) Test whether or not there is significant evidence of a negative correlation between x and y
You should

- state your hypotheses clearly
- use a 5% level of significance
- state the critical value used

$H_0 : \rho = 0 \quad H_1 : \rho < 0$
Critical value – 0.3887 (Allow $\pm$)



$-0{\cdot}897$ is stronger than $-0{\cdot}3887$

Exemplar – Student B and Student C

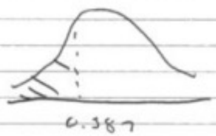
b) $x = \log_{10} m$

$y = \log_{10} h$

PMTT = $-0.897 = r$

$H_0: \rho = 0$

$H_1: \rho < 0$ critical value at 0.05 = 0.3887



evidence to reject H_0 and except H_1 as it falls within

b) $H_0: \rho = 0$ critical value = -0.3887

$H_1: \rho < 0$

5% => 0.05

n=19

$-0.897 < -0.3887$, this value is significant Reject H_0 and accept H_1, as there is sufficient evidence to suggest there is a negative correlation between x and y.

Part (b)
B1 Correct hypotheses
M1 The CV is correct and has the minus sign.
A1 A correct conclusion that mentions correlation and negative and does not contradict their "reject H_0" statement.

3/3

Whilst many students realised that one- tailed hypotheses were needed in part (b) few used ρ for the parameter. Most students had $\pm$ the correct CV but the final A1 was lost by many students, as they compared the positive value 0.3887 to the negative test statistic. This showed these students were unfamiliar with the meaning of the CV/CR and how it related to their test statistic.

2/3

Part (b)
B1 Correct hypotheses. We do not get too fussy how close their □ looks to a p
M1 0.3887 lies within the range given in the MS 9it is in fact the correct value but note it is positive and the alternative hypothesis is < 0
A0 Because they have given the positive CV we need to see the comparison 0.3887 < 0.897 We would also need to see the correct conclusion.

9MA0 – Oct 21 Q2

2. Marc took a random sample of 16 students from a school and for each student recorded

- the number of letters, x, in their last name
- the number of letters, y, in their first name

His results are shown in the scatter diagram on the next page.

(a) Describe the correlation between x and y.

(1)

Marc suggests that parents with long last names tend to give their children shorter first names.

(b) Using the scatter diagram comment on Marc's suggestion, giving a reason for your answer.

(1)

The results from Marc's random sample of 16 observations are given in the table below.

x	3	6	8	7	5	3	11	3	4	5	4	9	7	10	6	6
y	7	7	4	4	6	8	5	5	8	4	7	4	5	5	6	3

(c) Use your calculator to find the product moment correlation coefficient between x and y for these data.

(1)

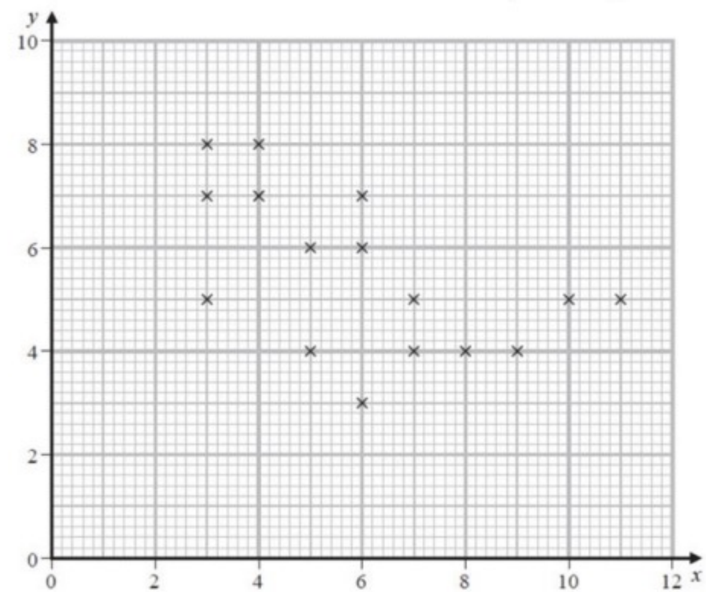
(d) Test whether or not there is evidence of a negative correlation between the number of letters in the last name and the number of letters in the first name.

You should

- state your hypotheses clearly
- use a 5% level of significance

(3)

(Total for Question 2 is 6 marks)



 3'

Qu 2	Scheme	Marks	AO
(a)	Negative	B1 (1)	1.2
(b)	Marc's suggestion is compatible because it's negative correlation	B1 (1)	2.4
(c)	$(r=)\ -0.54458266\ldots$ awrt $-$ **0.545**	B1 (1)	1.1b
(d)	$H_0: \rho = 0 \quad H_1: \rho < 0$	B1	2.5
	[5% 1-tail cv =] $(\pm)\ 0.4259$	M1	1.1a
	(significant result / reject H_0) There is evidence of negative correlation between the number of letters in (or length of) a student's last name and their first name	A1 (3)	2.2b
		(6 marks)	
	Notes		
(a)	B1 for "negative" Allow "slight" or "weak" etc Allow a description e.g. "as x increases y decreases" or in context e.g. "people with longer last names tend to have shorter first names" A comment of "negative skew" is B0 **Need to see distinct or separate responses for (a) and (b)**		
(b)	B1 for a comment that suggests data is compatible with the suggestion **and** a suitable reason such as "there is negative correlation" or a description in x and y or in context or the points lie close to a line with negative gradient or draw line $y = x$ and state that more points below the line so supports (or is compatible with) his suggestion A reason based on just a **single point** is B0 e.g. " 11 letters in last name has only 5 in first name"		
(c)	B1 for awrt -0.545		
(d)	B1 for both hypotheses correct in terms of ρ M1 for a critical value compatible with their H_1: 1-tail: awrt ± 0.426 (condone ± 0.425) **or** 2-tail (B0 scored for H_1) : awrt ± 0.497 If hypotheses are in words and can deduce whether one or two-tail then use their words. If no hypotheses or their H_1 is not clearly one or two tail assume one-tail A1 for compatible signs between cv and r **and** a correct conclusion in context mentioning correlation and number of letters or length and name (ft their value from (c)) Do NOT award this A mark if contradictory comments or working seen e.g. "accept H_0" or comparison of 0.426 with significance level of 0.05 etc		
NB	The M1A1 can be scored independently of the hypotheses		

Most answered part (c) correctly though a number "lost" the minus sign between their calculator and writing down the answer on the page. Usually there were sufficient figures given for us to award the mark when the minus sign was included. In part (d) most attempted the hypotheses and the majority used ρ. Some didn't have "= 0" for their null hypothesis with alternative values or inequalities being used and of course some used r or simply wrote the hypotheses in words. We allowed $(\pm)$ for the critical value provided the actual value was compatible with their alternative hypothesis and many achieved this mark but few scored the final mark as the conclusion was not related to the context of the question.

Students B and C

c) $r = -0.54458262$

d) $H_0: r < 0$ $\quad H_1: r > 0$

Assuming H_0:

from the table the critical region is -0.4259

$-0.54458... < -0.4259$

$\therefore$ -0.54452 is more extreme than the critical region

$\therefore$ there is enough evidence to reject H_0 in favour of H_1

$\therefore$ There is evidence that a negative correlation between the number of letters in the last name and number of letters in the first name exists.

3/6

(a) B0: They have said there is no correlation which does not score the mark. It is worth noting that this student did infer a negative correlation in part (d) but did not consider changing their answers to (a) and (b)

(b) B0: They have said that the graph does NOT support Marc's suggestion. Even ignoring the negative correlation it is clear that the points below the line $y = x$ offer support for Marc's suggestion.

(c) B1: This is correct.

(d) B0: The hypotheses are incorrect though suggest a one-tail test (in terms of r not ρ and with incorrect inequalities)

M1: The critical value is correct for a one-tail test.

A1: They have correctly rejected H_0 and given a correct contextual comment with the key words. We allowed students to score the M1 and A1 even if their hypotheses were incorrect this time.

c) $r = -0.544582$

$r = -0.545$

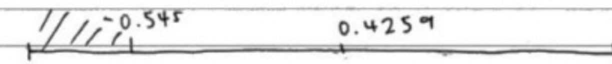
d) $H_0: \rho = 0$ $\quad$ 5% level of sig

$H_0: \rho < 0$ $\quad$ 1-tailed test

0.05

$n = 16$

$r = 0.4259$



dont reject H_0, there is no evidence that at a 5% significance level there is evidence of a negative correlation between the number of letters in the last name & the number of letter in the first name

5/6

(a) B1: They have described the correlation as "negative".

(b) B1: They have said that Marc's suggestion IS supported and given a description of the negative correlation in terms of x and y

(c) B1: This is correct. Notice how the student has first quoted several figures from their calculator and then given a rounded answer as we would recommend.

(d) B1: The hypotheses are correct.

M1: They have stated the correct critical value. We allow this mark for the correct value irrespective of the sign.

A0: Their conclusion though is incorrect (the incompatible signs have caused them to not reject H_0)

This is a good example to demonstrate the extra difficulty when testing using the left-hand tail. Some students simply compare $|r|$ with the critical value and this, of course, is fine but students need to be aware that the critical values in the tables are for the right-hand tail and minus signs will need to be introduced to use them correctly with negative value of r.

Critical Values for Correlation Coefficients

These tables concern tests of the hypothesis that a population correlation coefficient ρ is 0. The values in the tables are the minimum values which need to be reached by a sample correlation coefficient in order to be significant at the level shown, on a one-tailed test.

Product Moment Coefficient					
Level					Sample size, n
0.10	0.05	0.025	0.01	0.005	
0.8000	0.9000	0.9500	0.9800	0.9900	4
0.6870	0.8054	0.8783	0.9343	0.9587	5
0.6084	0.7293	0.8114	0.8822	0.9172	6
0.5509	0.6694	0.7545	0.8329	0.8745	7
0.5067	0.6215	0.7067	0.7887	0.8343	8
0.4716	0.5822	0.6664	0.7498	0.7977	9
0.4428	0.5494	0.6319	0.7155	0.7646	10
0.4187	0.5214	0.6021	0.6851	0.7348	11
0.3981	0.4973	0.5760	0.6581	0.7079	12
0.3802	0.4762	0.5529	0.6339	0.6835	13
0.3646	0.4575	0.5324	0.6120	0.6614	14
0.3507	0.4409	0.5140	0.5923	0.6411	15
0.3383	0.4259	0.4973	0.5742	0.6226	16
0.3271	0.4124	0.4821	0.5577	0.6055	17
0.3170	0.4000	0.4683	0.5425	0.5897	18
0.3077	0.3887	0.4555	0.5285	0.5751	19
0.2992	0.3783	0.4438	0.5155	0.5614	20
0.2914	0.3687	0.4329	0.5034	0.5487	21
0.2841	0.3598	0.4227	0.4921	0.5368	22
0.2774	0.3515	0.4133	0.4815	0.5256	23
0.2711	0.3438	0.4044	0.4716	0.5151	24
0.2653	0.3365	0.3961	0.4622	0.5052	25
0.2598	0.3297	0.3882	0.4534	0.4958	26
0.2546	0.3233	0.3809	0.4451	0.4869	27
0.2497	0.3172	0.3739	0.4372	0.4785	28
0.2451	0.3115	0.3673	0.4297	0.4705	29
0.2407	0.3061	0.3610	0.4226	0.4629	30
0.2070	0.2638	0.3120	0.3665	0.4026	40
0.1843	0.2353	0.2787	0.3281	0.3610	50
0.1678	0.2144	0.2542	0.2997	0.3301	60
0.1550	0.1982	0.2352	0.2776	0.3060	70
0.1448	0.1852	0.2199	0.2597	0.2864	80
0.1364	0.1745	0.2072	0.2449	0.2702	90
0.1292	0.1654	0.1966	0.2324	0.2565	100

2. A meteorologist believes that there is a relationship between the daily mean windspeed, w kn, and the daily mean temperature, t °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below.

t	13.3	16.2	15.7	16.6	16.3	16.4	19.3	17.1	13.2
w	7	11	8	11	13	8	15	10	11

The meteorologist calculated the product moment correlation coefficient for the 9 days and obtained $r = 0.609$

(a) Explain why a linear regression model based on these data is unreliable on a day when the mean temperature is 24 °C

(1)

(b) State what is measured by the product moment correlation coefficient.

(1)

(c) Stating your hypotheses clearly test, at the 5% significance level, whether or not the product moment correlation coefficient for the population is greater than zero.

(3)

Exams

- Formula Booklet
- Past Papers
- Practice Papers
- past paper Qs by topic

Past paper practice by topic. Both new and old specification can be found via this link on hgsmaths.com

Question	Scheme	Marks	AOs
2(a)	e.g. It requires extrapolation so will be unreliable (o.e.)	B1	1.2
		(1)	
(b)	e.g. Linear association between w and t	B1	1.2
		(1)	
(c)	$H_0: \rho = 0 \quad H_1: \rho > 0$	B1	2.5
	Critical value 0.5822	M1	1.1a
	Reject H_0		
	There is evidence that the product moment correlation coefficient is greater than 0	A1	2.2b
		(3)	

Summary of Key Points

Summary of key points

1. If $y = ax^n$ for constants a and n then $\log y = \log a + n \log x$

2. If $y = kb^x$ for constants k and b then $\log y = \log k + x \log b$

3. The **product moment correlation coefficient** describes the linear correlation between two variables. It can take values between -1 and 1.

4. For a one-tailed test use either:
   - $H_0 : \rho = 0$, $H_1 : \rho > 0$ or
   - $H_0 : \rho = 0$, $H_1 : \rho < 0$

   For a two-tailed test use:
   - $H_0 : \rho = 0$, $H_1 : \rho \neq 0$

T.12: mixed ex Qs 1-3, P.7: BSG