Hypothesis Testing for Correlation

	B	C	D	E	G	H
1		English Exam Mark			Maths Exam Mark	
2		Mean	60		Mean	70
3	Student	S.D.	5		S.D.	10
4	1		63.90			70.13
5	2		55.24			65.99
6	3		58.80			80.18
7	4		59.65			57.16
8	5		66.44			72.76
9	6		59.53			79.82
10	7		57.43			71.48
11	8		58.33			60.56
12	9		67.43			69.56
13	10		63.11			87.13
16		r=	0.219			

	B	C	D	E	G	H
1		English Exam Mark			Maths Exam Mark	
2		Mean	60		Mean	70
3	Student	S.D.	5		S.D.	10
4	1		60.22			74.64
5	2		62.25			79.15
6	3		61.30			75.29
7	4		60.61			71.35
8	5		55.31			74.05
9	6		57.13			89.73
10	7		57.16			70.41
11	8		58.96			60.31
12	9		56.30			71.95
13	10		63.23			69.95
16		r=	-0.094			

Suppose we use a spreadsheet to **randomly** generate maths marks for students, and separately generate **random** English marks.

The **observed** PMCC between Maths and English marks in this first set of data is **0.219**

But the true PMCC between Maths and English is 0.

This is because they were generated independently of each other and so have no correlation. The observed PMCC may vary from the true PMCC because the data is randomly sampled, just as if we threw a fair die, we wouldn't necessarily see equal counts of each outcome.

✎ r denotes the PMCC of a **sample**.

✎ ρ (Greek letter *rho*) is the PMCC for the **whole population**.

✎ $\therefore r$ is the test statistic, ρ is the population parameter.

Let's carry out a hypothesis test on whether there is **positive** correlation between English and Maths marks, at 10% significance level:

H_0: $\rho = 0$

H_1: $\rho > 0$

$\rho < 0$

Sample size = 10

If our r is greater than 0.4428, there is evidence to reject H_0. i.e. there is a positive correlation.

$r = -0.219$ $r = 0.219$

-0.4428 $0.219 < 0.4428$, so not enough evidence to reject H_0, this suggests there is no correlation between Maths and English marks.

$-0.4428 < -0.219$

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						Spearman's Coefficient		
Level					Sample	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

Note: you take the negative value from the table if looking at significance for negative correlation

Two-tailed test

In the previous example we hypothesised that English/Maths marks were positively correlated. But we could also test whether there was **any** correlation, i.e. positive **or** negative.

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.

Product Moment Coefficient					
Level					Sample
0.10	0.05	0.025	0.01	0.005	size, n
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
0.2992	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

$H_0: \rho = 0$

$H_1: \rho \neq 0$ 2-tailed

Sample size 30, sig. level 5%

$-0.45 < -0.3061$

There is evidence to reject H_0. This suggests the masses are correlated, and the scientist is incorrect.

The table from the large data set shows the daily maximum gust, x kn, 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

a. Find the product moment correlation coefficient for this data.

b. 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.

a) $r = 0.1149$

b) $H_0: \rho = 0$

$H_1: \rho > 0$

$0.5067 > 0.1149$

So no evidence to reject H_0, this suggests there is no correlation between daily max. gust and rel. humidity.

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)

Using the same 9 days a location from the large data set gave $\bar{t} = 27.2$ and $\bar{w} = 3.5$

(d) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics.

(1)

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: $\rho = 0$ 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)	
(d)	Higher $\bar{t}$ suggests overseas and not Perth…lower wind speed so perhaps not close to the sea so suggest **Beijing**	B1	2.4
		(1)	
			(6 marks)

Notes:

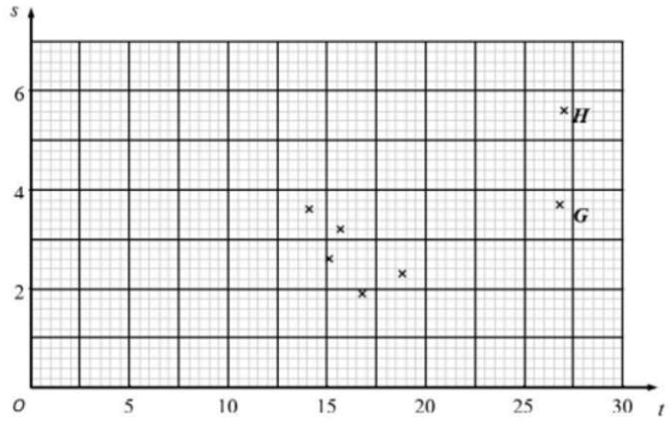
(a)
B1: for a correct statement (unreliable) with a suitable reason

(b)
B1: for a correct statement

(c)
B1: for both hypotheses in terms of ρ
M1: for selecting a suitable 5% critical value compatible with their H_1
A1: for a correct conclusion stated

(d)
B1: for suggesting Beijing with some supporting reason based on t or w
Allow Jacksonville with a reason based just on higher $\bar{t}$

2. A researcher believes that there is a linear relationship between daily mean temperature and daily total rainfall. The 7 places in the northern hemisphere from the large data set are used. The mean of the daily mean temperatures, t °C, and the mean of the daily total rainfall, s mm, for the month of July in 2015 are shown on the scatter diagram below.



(a) With reference to the scatter diagram, explain why a linear regression model may not be suitable for the relationship between t and s.

(1)

The researcher calculated the product moment correlation coefficient for the 7 places and obtained $r = 0.658$

(b) Stating your hypotheses clearly, test at the 10% level of significance, whether or not the product moment correlation coefficient for the population is greater than zero.

(3)

(c) Using your knowledge of the large data set, suggest the names of the 2 places labelled G and H.

(1)

(d) Using your knowledge from the large data set, and with reference to the locations of the 2 places labelled G and H, give a reason why these places have the highest temperatures in July.

(1)

(e) Suggest how you could make better use of the large data set to investigate the relationship between daily mean temperature and daily total rainfall.

(1)

Question	Scheme	Marks	AOs
2(a)	Not suitable with a correct reason eg the points do not lie close to a straight line. there appear to be two populations if *G* and *H* were removed it appears to be a negative correlation	B1	1.2
		(1)	
(b)	$H_0 : \rho = 0 \quad H_1 : \rho > 0$	B1	2.5
	Critical value 0.5509	M1	1.1a
	Reject H_0		
	There is evidence that pmcc is greater than zero	A1	2.2b
		(3)	
(c)	Beijing and Jacksonville	B1	2.2a
		(1)	
(d)	Beijing and Jacksonville are the closest to the equator	B1	2.4
		(1)	
(e)	Use data from one place.	B1	2.4
		(1)	
			(7 marks)

Notes:

(a) B1: for a correct statement using the data in the table

(b) B1: for both hypotheses in terms of ρ

M1: for selecting a suitable critical value compatible with their H_1

A1: for a correct conclusion stated

(c) B1: both Beijing and Jacksonville – they do not need to be attached to G and H correctly.

(d) B1: for the idea they are near the equator dependent only Beijing or Jacksonville being given in part(c)

2. An ornithologist believes that there is a relationship between the tail length, t mm, and the wing length, w mm, of female hook-billed kites. A random sample of size 10 is taken from a database of these kites and the relevant data is given in the table below.

t **(mm)**	191	197	208	180	188	210	196	191	179	208
w **(mm)**	284	285	288	273	280	283	288	271	257	289

The ornithologist plans to use a linear regression model based on these data and interpolate or extrapolate as necessary to estimate the wing length of other female hook-billed kites from their tail length.

(a) (i) Explain what is meant by extrapolation.

(1)

(ii) Explain the dangers of extrapolation.

(1)

The ornithologist attempts to calculate the product moment correlation coefficient, r, and obtains a value of 1.3

(b) Explain how she should be able to identify that this is incorrect without carrying out any further calculations.

(1)

(c) Use your calculator to find the correct value of the product moment correlation coefficient, r.

(1)

(d) Stating your hypotheses clearly test, at the 1% significance level, whether or not there is evidence that the product moment correlation coefficient for the population is positive.

(3)

(e) Explain what your test in part (d) suggests about female hook-billed kites.

(1)

Question	Scheme	Marks	AOs
2(a)(i)	Extrapolation is making predictions outside the original data range.	B1	1.2
(a)(ii)	This is unreliable as the trend may not continue.	B1	2.4
		(2)	
(b)	The product moment correlation coefficient cannot be greater than 1	B1	1.2
		(1)	
(c)	$r = 0.76279\ldots$ awrt 0.763	B1	1.1b
		(1)	
(d)	$H_0: \rho = 0 \qquad H_1: \rho > 0$	B1	2.5
	Critical value 0.7155	M1	1.1a
	Reject H_0		
	There is evidence that the product moment correlation coefficient is greater than 0	A1ft	2.2b
		(3)	
(e)	This suggests that on average (female hook-billed) kites with longer tails have longer wings.	B1	3.2a
		(1)	
			(8 marks)

Notes:
(a) B1: for a correct definition of extrapolation B1: for a correct statement of the dangers of extrapolation
(b) B1: for a correct statement
(c) B1: for awrt 0.763
(d) B1: for both hypotheses in terms of ρ M1: for selecting a suitable 1% critical value compatible with their H_1 A1: for correct conclusion stated ft their (c) provided $-1 \leq r \leq 1$
(e) B1: for correct interpretation in context ft their (d) provided $-1 \leq r \leq 1$

2. Tessa owns a small clothes shop in a seaside town. She records the weekly sales figures, £w, and the average weekly temperature, t°C, for 8 weeks during the summer.
The product moment correlation coefficient for these data is −0.915

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not the correlation between sales figures and average weekly temperature is negative.

(3)

(b) Suggest a possible reason for this correlation.

(1)

Tessa suggests that a linear regression model could be used to model these data.

(c) State, giving a reason, whether or not the correlation coefficient is consistent with Tessa's suggestion.

(1)

(d) State, giving a reason, which variable would be the explanatory variable.

(1)

Tessa calculated the linear regression equation as $w = 10\,755 - 171t$

(e) Give an interpretation of the gradient of this regression equation.

(1)

Qu 2	Scheme	Marks	AO
(a)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	Critical value: -0.6215 (Allow any cv in range $0.5 < \lvert cv \rvert < 0.75$)	M1	1.1a
	$r < -0.6215$ so significant result and there is evidence of a negative correlation between w and t	A1	2.2b
		(3)	
(b)	e.g. As temperature increases people spend more time on the beach and less time shopping (o.e.)	B1	2.4
		(1)	
(c)	Since r is close to -1 , it **is** consistent with the suggestion	B1	2.4
		(1)	
(d)	t will be the explanatory variable since sales are likely to depend on the temperature	B1	2.4
		(1)	
(e)	Every degree rise in temperature leads to a drop in weekly earnings of £171	B1	3.4
		(1)	
		(7 marks)	

	Notes
(a)	B1 for both hypotheses in terms of ρ
	M1 for the critical value: sight of ± 0.6215 or any cv such that $0.5 < \lvert cv \rvert < 0.75$
	A1 must reject H_0 on basis of comparing -0.915 with -0.6215 (if $-0.915 < 0.6215$ is seen then A0 but may use $\lvert r \rvert$ o.e. which is fine) and mention "negative", "correlation/relationship" and at least "w" and "t"
(b)	B1 for a suitable reason to explain negative correlation using the context given. e.g. "As temperature drops people are more likely to go shopping (than to the beach)" e.g. "As temperature increases people will be outside rather than in shops" A mere description in context of negative correlation is B0
	SO e.g. "As temperature increases people don't want to go shopping/buy clothes" is B0 e.g. "Less clothes needed as temp increases" is B0
(c)	B1 for a suitable reason e.g. "strong"/"significant"/"near perfect" "correlation", $\lvert r \rvert$ close to 1 and saying it is consistent with the suggestion. Allow "yes" followed by the reason.
(d)	B1 For identifying t and giving a suitable reason. Need idea that "w depends on t" or "w responds to t" or "t affects w" (o.e.) Allow t (temperature) affects the other variable etc Just saying "t is the independent variable" or "t explains change in w" is B0 N. B. Suggesting causation is B0 e.g. "t causes w to decrease"
(e)	B1 for a description that conveys the idea of rate per degree Celsius. Must have 171, condone missing "£" sign.

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						Spearman's Coefficient		
Level					Sample size, n	Level		
0.10	0.05	0.025	0.01	0.005		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
0.3507	0.4409	0.5140	0.5923	0.6411	15	0.4464	0.5214	0.6036
0.3383	0.4259	0.4973	0.5742	0.6226	16	0.4294	0.5029	0.5824
0.3271	0.4124	0.4821	0.5577	0.6055	17	0.4142	0.4877	0.5662
0.3170	0.4000	0.4683	0.5425	0.5897	18	0.4014	0.4716	0.5501
0.3077	0.3887	0.4555	0.5285	0.5751	19	0.3912	0.4596	0.5351
0.2992	0.3783	0.4438	0.5155	0.5614	20	0.3805	0.4466	0.5218

0.2914	0.3687	0.4329	0.5034	0.5487	21	0.3701	0.4364	0.5091
0.2841	0.3598	0.4227	0.4921	0.5368	22	0.3608	0.4252	0.4975
0.2774	0.3515	0.4133	0.4815	0.5256	23	0.3528	0.4160	0.4862
0.2711	0.3438	0.4044	0.4716	0.5151	24	0.3443	0.4070	0.4757
0.2653	0.3365	0.3961	0.4622	0.5052	25	0.3369	0.3977	0.4662
0.2598	0.3297	0.3882	0.4534	0.4958	26	0.3306	0.3901	0.4571
0.2546	0.3233	0.3809	0.4451	0.4869	27	0.3242	0.3828	0.4487
0.2497	0.3172	0.3739	0.4372	0.4785	28	0.3180	0.3755	0.4401
0.2451	0.3115	0.3673	0.4297	0.4705	29	0.3118	0.3685	0.4325
0.2407	0.3061	0.3610	0.4226	0.4629	30	0.3063	0.3624	0.4251
0.2070	0.2638	0.3120	0.3665	0.4026	40	0.2640	0.3128	0.3681
0.1843	0.2353	0.2787	0.3281	0.3610	50	0.2353	0.2791	0.3293
0.1678	0.2144	0.2542	0.2997	0.3301	60	0.2144	0.2545	0.3005
0.1550	0.1982	0.2352	0.2776	0.3060	70	0.1982	0.2354	0.2782
0.1448	0.1852	0.2199	0.2597	0.2864	80	0.1852	0.2201	0.2602
0.1364	0.1745	0.2072	0.2449	0.2702	90	0.1745	0.2074	0.2453
0.1292	0.1654	0.1966	0.2324	0.2565	100	0.1654	0.1967	0.2327