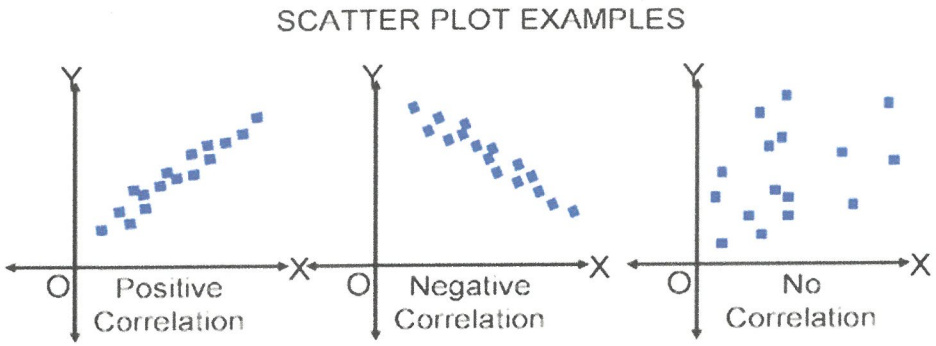
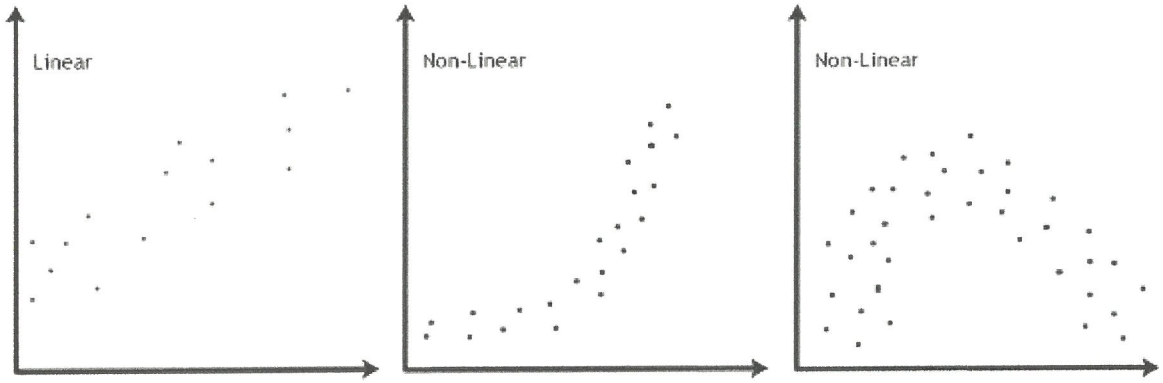
Correlation – Measuring the relationship between bivariate data

At GCSE we became familiar with the use of the term correlation when describing points plotted on a scatter-graph. Such points are based on the values of two variables, such as the monetary value compared to the mileage of a second hand car, or the marks awarded in Paper 1 and Paper 2 of a Maths exam. We would then describe these relationships as having 'positive', 'negative', or 'no' correlation.



Where there was evidence of correlation we learned how to draw and use a line of best fit on the plot.

In statistics there are different ways to **quantify** the strength of correlation exhibited between a pair of variables. These are known as correlation coefficients. When it comes to linear correlation (other forms of correlation exist, see below), the two measures used are **Pearson's Product Moment Correlation Coefficient (PMCC)** and **Spearman's Coefficient of Rank Correlation**.



Pearson's coefficient is calculated on raw bivariate data and measures how close the data lie to a straight line. A value of +1 would indicate that all the data lie on a straight line with a positive gradient (perfect positive correlation) and a correlation coefficient of -1 would indicate perfect negative correlation, with all the data laying on a straight line with a negative gradient.

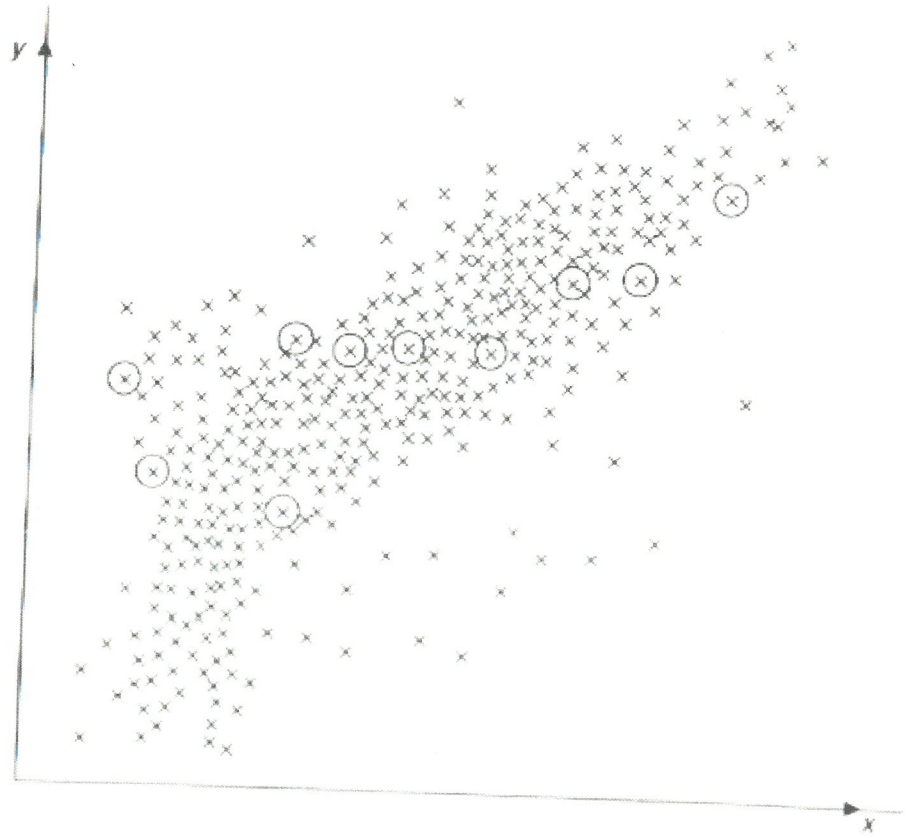
Spearman's coefficient is used when the data has been ranked in some way. For example if two judges in a bake off rank cakes baked by 10 different people. If both judges agreed exactly on the ranking of the cakes this would produce a coefficient of +1, a value of -1 would occur if the two judges had the ranking in exactly reverse order, 1st place for Judge 1 = 10th place for Judge 2, etc.

In Unit 4, we are not required to actually calculate the coefficients, just interpret them.

The meaning of a correlation coefficient

As described on the previous page, if the value of a correlation coefficient for a sample (denoted by the letter r), is close to +1 or -1, we can be satisfied that there is linear correlation. The question we are looking to address in this work is what happens in a case such as r = 0.6?

When calculating a value of r, we are actually using a sample of bivariate data from a much larger parent population. For example, in the scatter diagram below we may have sampled the ten circled points to calculate r.



There will be a level of correlation within the parent population and this is denoted by the Greek letter ρ (pronounced rho).

The calculated value of r, based on the sample can be used as an estimate for ρ. It can also be used to carry out an hypothesis test on the value of ρ, the parent population correlation coefficient. Used in this way it is described as a *test statistic*.

The simplest hypothesis test which you can carry out is that there is no correlation within the parent population. This gives rise to the null hypothesis:

H_0: $\rho = 0$ 'There is no correlation between the two variables'

There are three possible alternative hypotheses, according to the sense of the situation being investigated. These are:

H_1: $\rho \neq 0$ 'There is correlation between the variables' (2-tail test)

or H_1: $\rho > 0$ 'There is positive correlation between the variables' (1-tail test)

or H_1: $\rho < 0$ 'There is negative correlation between the variables' (1-tail test)

The test is carried out by comparing the calculated value of r with the appropriate entry in a table of critical values (actually easier to use than our calculator for this!). This will depend on the size of the sample, the significance level at which we are testing and whether the test is a 1-tail or a 2-tail test.

Eg7 A language teacher wished to test whether there is any correlation between students' ability in their own and a foreign language. Accordingly she collects the marks of 8 students, all native English speakers, in their end of year examinations in English and French.

Candidate	A	B	C	D	E	F	G	H
English	65	34	48	72	58	63	26	80
French	74	49	45	80	63	72	12	75

(i) Use your calculator to determine the PMCC

(ii) State the null and alternative hypotheses

(iii) Using the value of r as the test statistic, carry out the test at the 5% significance level.

(i) $r = 0.913$

(ii) $H_0: \rho = 0$ there is no correlation between performance in English & French

$H_1: \rho > 0$ " " a " " " " " "

testing @ 5% significance 1-tail test

From tables, critical value $r = 0.6215$

Acceptance region $r < 0.6215$

Our test correlation coefficient $r = 0.913 > 0.6215$ critical value

$\therefore$ reject H_0

Hence there is sufficient evidence to suggest a positive relationship between English & French test scores.

Eg8 "You can't win without scoring goals". So says the coach of a netball team. Jamila, who believes in solid defensive play, disagrees and sets out to prove that there is no correlation between scoring goals and winning matches. She collects the following data for the goals scored and the points scored by 12 teams in a netball league.

Goals Scored	**41**	**50**	**54**	**47**	**47**	**49**	**52**	**61**	**50**	**29**	**47**	**35**
Points Scored	21	20	19	18	16	14	12	11	11	7	5	2

(i) Use your calculator to determine the PMCC

(ii) State suitable null and alternative hypotheses, indicating whose position each represents

(iii) Carry out the hypothesis test and comment on the result

(i) Calculated sample $r = 0.38$

(ii) $H_0 : \rho = 0$ Jamila, there is no correlation between goals scored and winning

$H_1 : \rho > 0$ Coach, you score goals, you win.

(iii) 1 tail test @ 5% significance, $n = 12$

From tables, critical value $r = 0.4973$

Acceptance region of H_0 $r < 0.4973$

Test correlation coefficient $r = 0.38 < 0.4973$

$\therefore$ accept H_0

Hence there is evidence to support Jamila's view.

Eg9 Charlotte is a campaigner for temperance, believing that drinking alcohol is an evil habit. Michael, a representative of a wine company, presents her with figures which he claims show that wine drinking is good for marriages.

Country	Wine Consumption (kg/person/year)	Divorce Rate (/1000 inhabitants)
Belgium	20	2.0
Denmark	20	2.7
Germany	26	2.2
Greece	33	0.6
Italy	63	0.4
Portugal	54	0.9
Spain	41	0.6
UK	13	2.9

(i) Write Michael's claim in the form of an hypothesis test and carry it out at the 5% significance level.

(ii) Charlotte claims that Michael is 'indulging in pseudo-statistics'. What arguments could she use to support this point of view?

(i) $H_0: \rho = 0$ There is no correlation between wine consumption & divorce rate

$H_1: \rho < 0$ The greater the wine consumption, the lower the divorce rate

1-tail test @ 5% significance.

From sample, test correlation coefficient $r = -0.854$

From tables, critical value $r = 0.6215$

Acceptance region for H_0, $r < 0.6215$

Compare to $|r|$ test statistic $r = 0.854 >$ critical value

$\therefore$ reject H_0

So there is evidence to suggest that there is a negative relationship between wine consumption & divorce rate.

(ii) "Correlation does not imply causation"

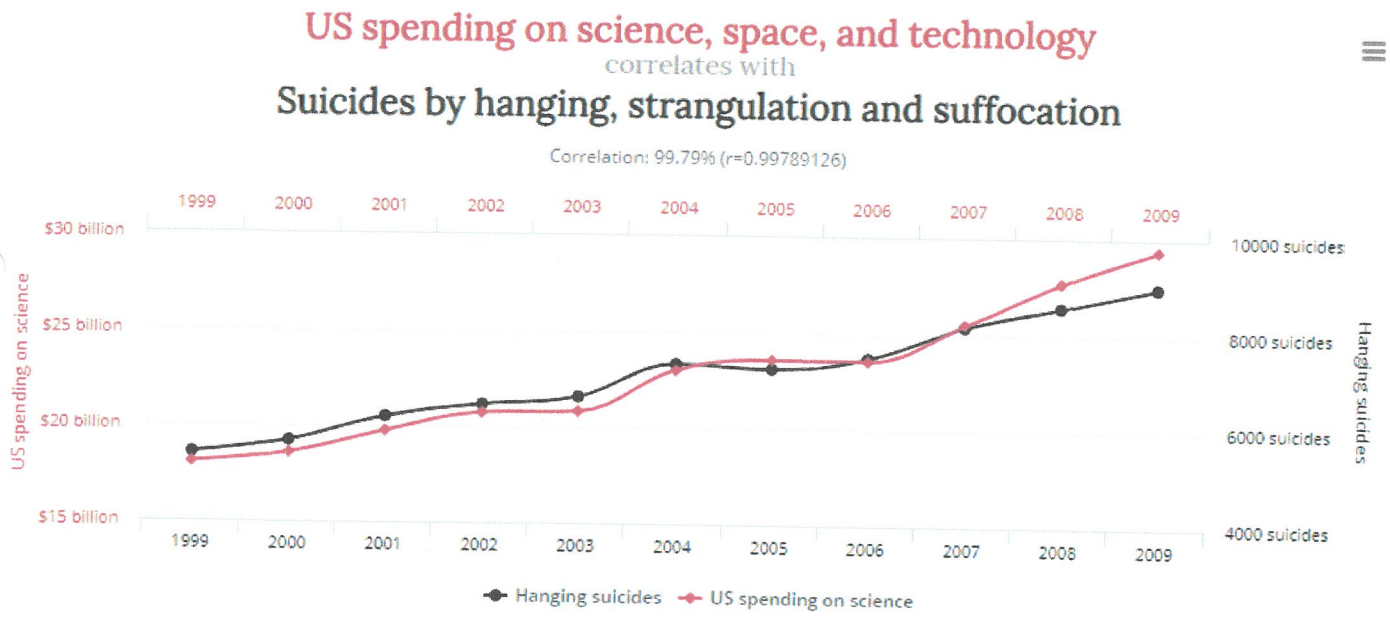
Interpreting Correlation

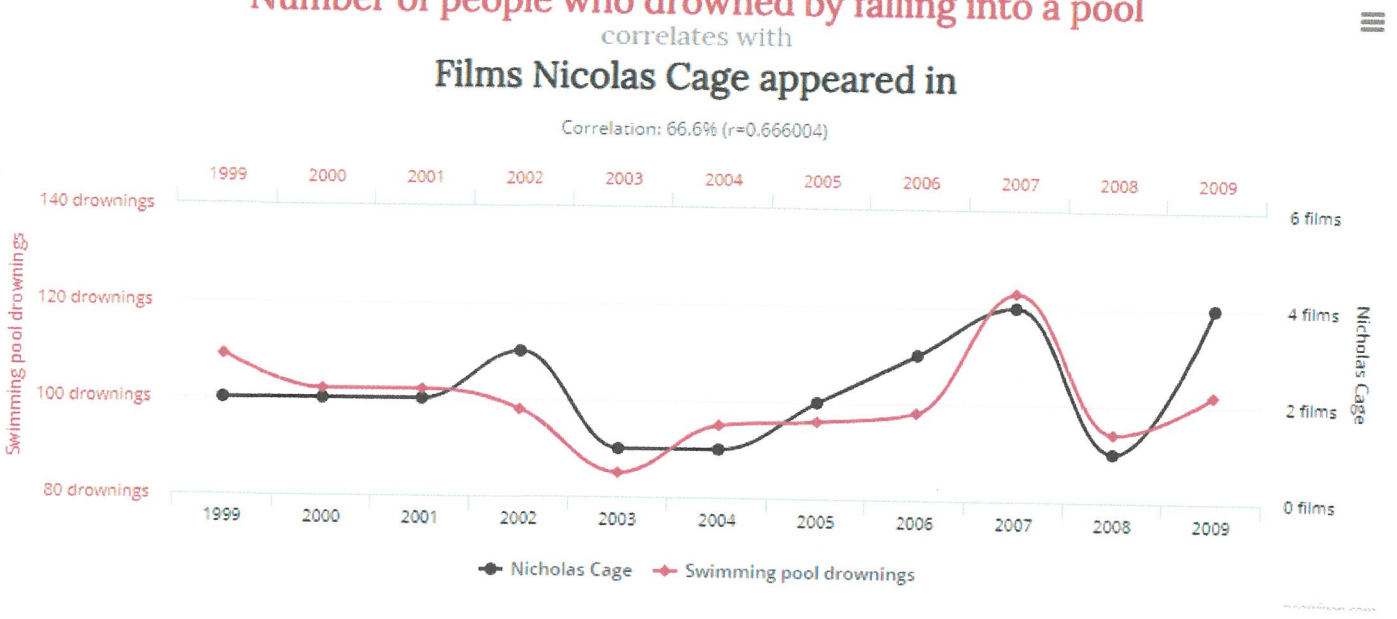
Correlation does not imply causation

The above example demonstrates a situation where a high level of correlation does not necessarily mean that there is a direct connection between the two variables.

Although there may be a high level of correlation between variables A and B it does not mean that A → B or that A ← B. It may well be that a third variable C causes both A and B or it may be a more complicated set of relationships.

Here are some examples taken from 'Spurious Correlations', www.tylervigen.com





Using a p-value

Sometimes you may be given a p-value instead of a value for the correlation coefficient. In this context, the p-value is the probability that, if there is no correlation, a random sample gives the given value for the correlation coefficient – in other words, the probability that this correlation coefficient could have been obtained by chance.

As with the hypothesis testing for sample means, we compare the p-value with the significance level. If the p-value is less than the significance level, this means that it is very unlikely that this value for the correlation coefficient could have risen by chance, so the null hypothesis is rejected.

Eg10 A random sample of 50 pairs of bivariate data (x, y) produce a product moment correlation coefficient of -0.3. This gives a 1-tail p-value of 0.0171.

What is the 2-tail p-value?

Stating your hypotheses clearly, test at the 5% significance level whether there is negative correlation between x and y within the parent population from which the values are drawn.

1-tail p-value of 0.0171 means a 1.71% chance of a rep sample of any 50 data pairs from the parent population having a negative correlation of -0.3 or stronger if the parent population had no correlation.

2-tail p-value would be 2x0.0171 = 0.0342

$H_0: \rho = 0$ parent population has no correlation

$H_1: \rho < 0$ " " " negative "

1-tail test @ 5% significance.

because p-value, 1.71% < 5% we can reject H_0

∴ there is evidence to suggest a negative correlation exists within the parent population

Exercise 3.2 (Edexcel S3 Ex 5B)

1 A product-moment correlation coefficient of 0.3275 was obtained from a sample of 40 pairs of values. Test whether or not this value shows evidence of correlation:

a at the 0.05 level (use a two-tailed test), **b** at the 0.02 level (use a two-tailed test).

2 **a** Calculate the product-moment correlation coefficient for the following data, giving values for S_{xx}, S_{yy} and S_{xy}.

x	2	3	4	4	5	5	6
y	7	6	5	4	3	2	1

b Test, for these data, the null hypothesis that there is no correlation between x and y. Use a 1% significance level.
State any assumptions you have made.

3 The ages X (years) and heights Y (cm) of 11 members of a football team were recorded and the following statistics were used to summarise the results.

$$\sum X = 168, \quad \sum Y = 1275, \quad \sum XY = 20\,704, \quad \sum X^2 = 2585 \quad \sum Y^2 = 320\,019$$

a Calculate the product-moment correlation coefficient for these data. r = 0.677

b Test the assertion that height and weight are positively correlated by using a suitable test. State your conclusion in words and any assumptions you have made. (Use a 5% level of significance.)

4 **a** Explain briefly your understanding of the term 'correlation'. Describe how you used, or could have used, correlation in a project or in class work.

b Twelve students sat two Biology tests, one theoretical the other practical. Their marks are shown below.

Marks in theoretical test (t)	5	9	7	11	20	4	6	17	12	10	15	16
Marks in practical test (p)	6	8	9	13	20	9	8	17	14	8	17	18

Find to 3 significant figures,
i the value of S_{tp} **ii** the product-moment correlation coefficient.

c Use a 0.05 significance level and a suitable test to check the statement that 'students who do well in theoretical Biology also do well in practical Biology tests'.

5 The following table shows the marks attained by 8 students in English and Mathematics tests.

Student	A	B	C	D	E	F	G	H
English	25	18	32	27	21	35	28	30
Mathematics	16	11	20	17	15	26	32	20

a Calculate the product-moment correlation coefficient.

A teacher thinks that the population correlation coefficient between the marks is likely to be zero.

b Test the teacher's idea at the 5% level of significance.

6 A small company decided to import fine Chinese porcelain. They believed that in the long term this would prove to be an increasingly profitable arrangement with profits increasing proportionally to sales. Over the next 6 years their sales and profits were as shown in the table below.

Year	1994	1995	1996	1997	1998	1999
Sales in thousands	165	165	170	178	178	175
Profits in £1000	65	72	75	76	80	83

Using a 1% significance level test to see if there is any evidence that the company's beliefs were correct, and that profit and sales were positively correlated.

Answers

1 $H_0: \rho = 0$; $H_1: \rho \neq 0$

a Critical values $\pm$ 0.3120. Reject H_0.

b Critical values $\pm$ 0.3665. Do not reject H_0.

2 **a** $S_{xx} = 10.857\,14$, $S_{yy} = 28$, $S_{xy} = -17$, $r = -0.975\ldots$

b Assume data are jointly normally distributed.

Critical values $\pm$0.8745.

Reject H_0: there is a correlation between x and y.

3 **a** $r = 0.677\ldots$

b Assume data are jointly normally distributed.

H_0: $\rho = 0$;

H_1: $\rho > 0$, 5% critical value is 0.5214. Reject H_0.

There is evidence to suggest that the taller you are the older you are.

4 **a** The product-moment coefficient of correlation is the measure of the strength of the linear link between two variables. You could use it to investigate whether there is correlation between the age of a lichen and its diameter, for example.

b **i** $S_{dp} = 255$ **ii** $r = 0.935$

c H_0: $\rho = 0$; H_1: $\rho > 0$;

Critical value = 0.4973

Reject H_0: there is reason to believe that students who do well in theoretical biology are likely to do well in practical biology.

5 **a** 0.686

b critical value = $\pm$0.7067,

Reject H_0. There is some evidence to show that the theory is correct.

6 $r = 0.793$, Critical value = 0.8822 Accept H_0. There is evidence to suggest the that company is incorrect to believe that profits increase with sales

TABLE 9 CRITICAL VALUES OF THE PRODUCT MOMENT CORRELATION COEFFICIENT

The table gives the critical values, for different significance levels, of the sample product moment correlation coefficient r based on n independent pairs of observations from a bivariate normal distribution with correlation coefficient p = 0.

One tail Two tail n	10% 20%	5% 10%	2.5% 5%	1% 2%	0.5% 1%
4	0.8000	0.9000	0.9500	0.9800	0.9900
5	0.6870	0.8054	0.8783	0.9343	0.9587
6	0.6084	0.7293	0.8114	0.8822	0.9172
7	0.5509	0.6694	0.7545	0.8329	0.8745
8	0.5067	0.6215	0.7067	0.7887	0.8343
9	0.4716	0.5822	0.6664	0.7498	0.7977
10	0.4428	0.5494	0.6319	0.7155	0.7646
11	0.4187	0.5214	0.6021	0.6851	0.7348
12	0.3981	0.4973	0.5760	0.6581	0.7079
13	0.3802	0.4762	0.5529	0.6339	0.6835
14	0.3646	0.4575	0.5324	0.6120	0.6614
15	0.3507	0.4409	0.5140	0.5923	0.6411
16	0.3383	0.4259	0.4973	0.5742	0.6226
17	0.3271	0.4124	0.4821	0.5577	0.6055
18	0.3170	0.4000	0.4683	0.5425	0.5897
19	0.3077	0.3887	0.4555	0.5285	0.5751
20	0.2992	0.3783	0.4438	0.5155	0.5614
21	0.2914	0.3687	0.4329	0.5034	0.5487
22	0.2841	0.3598	0.4227	0.4921	0.5368
23	0.2774	0.3515	0.4132	0.4815	0.5256
24	0.2711	0.3438	0.4044	0.4716	0.5151
25	0.2653	0.3365	0.3961	0.4622	0.5052
26	0.2598	0.3297	0.3882	0.4534	0.4958
27	0.2546	0.3233	0.3809	0.4451	0.4869
28	0.2497	0.3172	0.3739	0.4372	0.4785
29	0.2451	0.3115	0.3673	0.4297	0.4705
30	0.2407	0.3061	0.3610	0.4226	0.4629
31	0.2366	0.3009	0.3550	0.4158	0.4556
32	0.2327	0.2960	0.3494	0.4093	0.4487
33	0.2289	0.2913	0.3440	0.4032	0.4421
34	0.2254	0.2869	0.3388	0.3972	0.4357
35	0.2220	0.2826	0.3338	0.3916	0.4296
36	0.2187	0.2785	0.3291	0.3862	0.4238
37	0.2156	0.2746	0.3246	0.3810	0.4182
38	0.2126	0.2709	0.3202	0.3760	0.4128
39	0.2097	0.2673	0.3160	0.3712	0.4076
40	0.2070	0.2638	0.3120	0.3665	0.4026
41	0.2043	0.2605	0.3081	0.3621	0.3978
42	0.2018	0.2573	0.3044	0.3578	0.3932
43	0.1993	0.2542	0.3008	0.3536	0.3887
44	0.1970	0.2512	0.2973	0.3496	0.3843
45	0.1947	0.2483	0.2940	0.3457	0.3801
46	0.1925	0.2455	0.2907	0.3420	0.3761
47	0.1903	0.2429	0.2876	0.3384	0.3721
48	0.1883	0.2403	0.2845	0.3348	0.3683
49	0.1863	0.2377	0.2816	0.3314	0.3646
50	0.1843	0.2353	0.2787	0.3281	0.3610
60	0.1678	0.2144	0.2542	0.2997	0.3301
70	0.1550	0.1982	0.2352	0.2776	0.3060
80	0.1448	0.1852	0.2199	0.2597	0.2864
90	0.1364	0.1745	0.2072	0.2449	0.2702
100	0.1292	0.1654	0.1966	0.2324	0.2565

Ex 3.2

(1) $r = 0.3275$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

(a) @ 5% Critical value 0.3120

So acceptance region for H_0, $r < 0.3120$

So reject H_0

(b) @ 2% critical value 0.3665

So acceptance region for H_0, $r < 0.3665$

So accept H_0

(2). (a) $r = -0.975$

(b) $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

2-tail test @ 1% sig, $n = 7$

From tables, critical value 0.8745

So acceptance region for H_0 $r < 0.8745$

test value $r = 0.975$

∴ reject H_0

Assuming data are jointly normally distributed.

(3) $r = 0.677$

$H_0: \rho = 0$ no correlation between height and weight.

$H_1: \rho > 0$ positive " " " "

1 tail test @ 5% sig, $n = 11$

from the tables, critical value 0.5214

So acceptance region for H_0, $r < 0.5214$

from sample test statistic $r = 0.677$

∴ reject H_0

there is evidence to suggest the taller the learner you are.

Assume variables are normally distributed

(4)(b) $r = 0.935$

(c) $H_0: \rho = 0$ no correlation between marks in theory & prac

$H_1: \rho > 0$ good at one, good at both.

1-tail test @ 5% sig, $n = 12$

from tables, critical value 0.4973

So acceptance region for H_0, $r < 0.4973$

test statistic $r = 0.935$

So reject H_0

there is evidence to suggest that students who do well in theory also do well at prac

(5) (a) $r = 0.686$

(b) $H_0 : \rho = 0$

$H_1 : \rho \neq 0$

2-tail test @ 5% sig $n = 8$

from tables, critical value $= 0.7067$

acceptance regn for H_0, $r < 0.7067$

test stat $r = 0.686$

$\therefore$ accept H_0

(6) test statistic $r = 0.793$

$H_0 : \rho = 0$

$H_1 : \rho > 0$

1-tail test @ 1% sig, $n = 6$

critical value $= 0.8822$

Acceptance regn for H_0, $r < 0.8822$

$\therefore$ accept H_0

there is not enough evidence to accept company's belief.