Dr Oliver Mathematics
Mathematics
Correlation and Regression
Past Examination Questions

This booklet consists of 38 questions across a variety of examination topics.
The total number of marks available is 420.

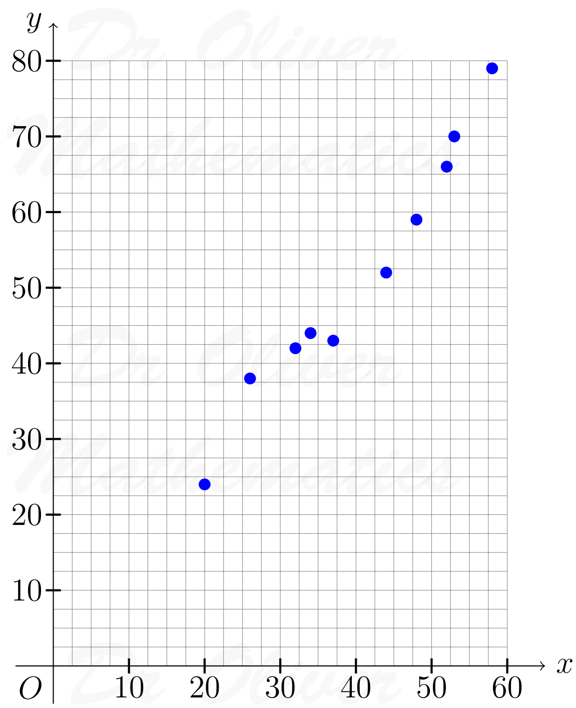
1. The chief executive of Rex cars wants to investigate the relationship between the number of new car sales and the amount of money spent on advertising. She collects data from company records on the number of new car sales, c, and the cost of advertising each year, p (£000). The data are shown in the table below.

Year	Number of new car sale, c	Cost of advertising (£000), p
1990	4240	120
1991	4380	126
1992	4420	132
1993	4440	134
1994	4430	137
1995	4520	144
1996	4590	148
1997	4660	150
1998	4700	153
1999	4790	158

(a) Using the coding $x = (p - 100)$ and $y = \frac{1}{10}(c - 4000)$, draw a scatter diagram to represent these data. Explain why x is the explanatory variable. (5)

Solution

Year	x	y
1990	20	24
1991	26	38
1992	32	42
1993	34	44
1994	37	43
1995	44	52
1996	48	59
1997	50	66
1998	53	70
1999	58	79



x is the change in the cost of advertising that influences the number of new car sales.

(b) Find the equation of the least squares regression line of y on x. You may use (7)

$$\Sigma\, x = 402,\ \Sigma\, y = 517,\ \Sigma\, x^2 = 17\,538,\ \text{ and } \Sigma\, xy = 22\,611.$$

Solution

$$\begin{aligned}
\overline{x} &= \frac{\Sigma x}{n} = \frac{402}{10} = 40.2 \\
\overline{y} &= \frac{\Sigma y}{n} = \frac{517}{10} = 51.7 \\
S_{xx} &= \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 17\,538 - \frac{402^2}{10} = 1377.6 \\
S_{xy} &= \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 22\,611 - \frac{402 \times 517}{10} = 1827.6 \\
b &= \frac{S_{xy}}{S_{xx}} = \frac{1827.6}{1377.6} = \tfrac{1523}{1148} \\
a &= \overline{y} - b\overline{x} = 51.7 - \tfrac{1523}{1148} \times 40.2 = -\tfrac{1873}{1148}.
\end{aligned}$$

Hence,

$$\underline{\underline{y = -\tfrac{1873}{1148} + \tfrac{1523}{1148}x}} \text{ or } \underline{\underline{y = -1.632 + 1.327x \text{ (4 sf)}}}.$$

(c) Deduce the equation of the least squares regression line of c on p in the form $c = a + bp$. (3)

Solution

$$\begin{aligned}
y = -\tfrac{1873}{1148} + \tfrac{1523}{1148}x &\Rightarrow \tfrac{1}{10}(c - 4000) = -\tfrac{1873}{1148} + \tfrac{1523}{1148}(p - 100) \\
&\Rightarrow c - 4000 = -\tfrac{18730}{1148} + \tfrac{15230}{1148}p - \tfrac{1523000}{1148} \\
&\Rightarrow \underline{\underline{c = 2657 + 13.27p \text{ (4 sf)}}}.
\end{aligned}$$

(d) Interpret the value of a. (2)

Solution

a is the <u>the number sold if no money spent is spent on advertising</u>. In that case, $p = 0$, and that is <u>well outside the valid range range</u>.

(e) Predict the number of extra new cars sales for an increase of £2000 in advertising budget. Comment on the validity of your answer. (2)

Solution

$$2 \times 13.27 = 26.54;$$

so either <u>26 or 27 cars sold</u>.

2. A researcher thinks there is a link between a person's height and level of confidence. She measured the height h, to the nearest cm, of a random sample of 9 people. She also devised a test to measure the level of confidence c of each person. The data are shown in the table below.

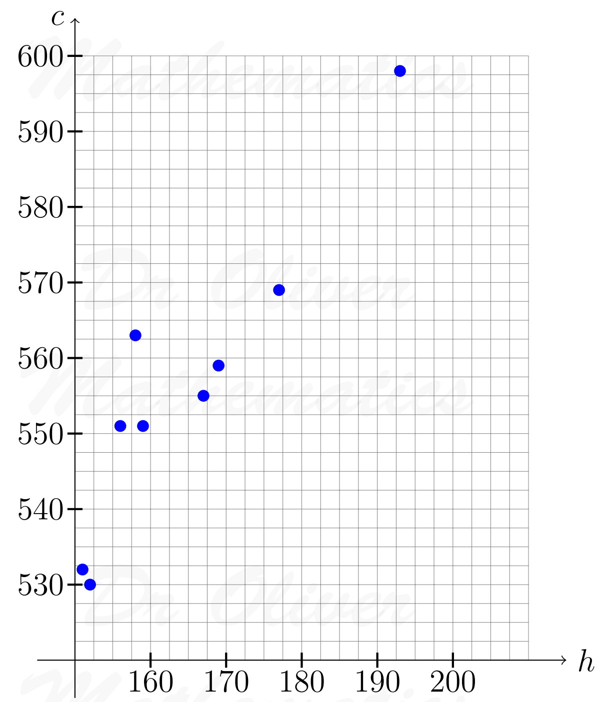
h	179	169	187	166	162	193	161	177	168
c	569	561	579	561	540	598	542	565	573

You may use

$$\Sigma h^2 = 272\,094,\ \Sigma c^2 = 2\,878\,966,\ \text{and}\ \Sigma hc = 884\,484.$$

(a) Draw a scatter diagram to illustrate these data. (4)

Solution



(b) Find exact values of S_{hc}, S_{hh}, and S_{cc}. (4)

Solution

$$
\begin{aligned}
\overline{h} &= \frac{\Sigma\, h}{n} = \frac{1562}{9} = 173\tfrac{5}{9} \\
\overline{c} &= \frac{\Sigma\, c}{n} = \frac{5088}{9} = 565\tfrac{1}{3} \\
S_{hh} &= \Sigma\, h^2 - \frac{(\Sigma\, h)^2}{n} = 272\,094 - \frac{(1562)^2}{9} = \underline{\underline{1000\tfrac{2}{9}}} \\
S_{cc} &= \Sigma\, c^2 - \frac{(\Sigma\, c)^2}{n} = 2\,878\,966 - \frac{(5088)^2}{9} = \underline{\underline{2550}} \\
S_{hc} &= \Sigma\, hc - \frac{(\Sigma\, h)(\Sigma\, c)}{n} = 884\,484 - \frac{(1562)(5088)}{9} = \underline{\underline{1433\tfrac{1}{3}}}.
\end{aligned}
$$

(c) Calculate the value of the product moment correlation coefficient for these data. (3)

Solution

$$
\begin{aligned}
r &= \frac{S_{hc}}{\sqrt{S_{hh} S_{cc}}} \\
&= \frac{1433\frac{1}{3}}{\sqrt{1000\frac{2}{9} \times 2550}} \\
&= 0.897\,488\,435\,2 \text{ (FCD)} \\
&= \underline{\underline{0.897 \text{ (3 dp)}}}.
\end{aligned}
$$

(d) Give an interpretation of your correlation coefficient. (1)

Solution

People who are taller tend to be more confident.

(e) Calculate the equation of the regression line of c on h in the form $c = a + bh$. (3)

Solution

$$
\begin{aligned}
b &= \frac{S_{hc}}{S_{cc}} = \frac{1433\frac{1}{3}}{1000\frac{2}{9}} = 1.433\,014\,886 \\
a &= \overline{c} - b\overline{h} = 565\tfrac{1}{3} - 1.433\ldots \times 173\tfrac{5}{9} = 316.625\,638\,7.
\end{aligned}
$$

Hence,

$$\underline{\underline{c = 316.6 + 1.433h \text{ (4 sf)}}}.$$

(f) Estimate the level of confidence of a person of height 180 cm. (2)

Solution

$h = 180 \Rightarrow c = 574.568\,318\,2$ (FCD) $= \underline{\underline{575 \text{ (nearest whole number)}}}$.

(g) State the range of values of h for which estimates of c are reliable. (1)

Solution

$$\underline{\underline{161 \leqslant h \leqslant 193}}.$$

3. Figure 1 shows the scatter diagrams that were drawn by a student.

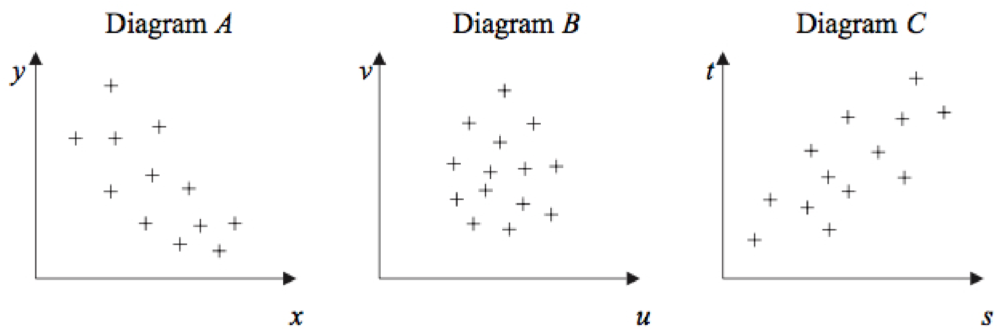


Figure 1: three scatter diagrams

The student calculated the value of the product moment correlation coefficient for each of the sets of data. The values were

$$0.68, \ -0.79, \ \text{and } 0.08.$$

Write down, with a reason, which value corresponds to which scatter diagram.

Solution

Diagram A: $\underline{\underline{-0.79}}$ because as $\underline{\underline{x \text{ increases, } y \text{ decreases}}}$

Diagram B: $\underline{\underline{0.08}}$ because there is $\underline{\underline{\text{no real pattern}}}$

Diagram C: $\underline{\underline{0.68}}$ because as $\underline{\underline{s \text{ increases, } t \text{ decreases}}}$

4. A long distance lorry driver recorded the distance travelled, m miles, and the amount of fuel used, f litres, each day. Summarised below are data from the driver's records for a random sample of 8 days. The data are coded such that $x = m - 250$ and $y = f - 100$.

$$\Sigma x = 130, \ \Sigma y = 48, \ \Sigma xy = 8\,880, \ \text{ and } S_{xx} = 20\,487.5.$$

(a) Find the equation of the regression line of y on x in the form $y = a + bx$. (6)

Solution

$$\overline{x} = \frac{\Sigma x}{n} = \frac{130}{8} = 16.25$$
$$\overline{y} = \frac{\Sigma y}{n} = \frac{48}{8} = 6$$
$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 8\,880 - \frac{130 \times 48}{8} = 8100$$
$$b = \frac{S_{xy}}{S_{xx}} = \frac{8100}{20\,487.5} = \tfrac{648}{1639}$$
$$a = \overline{y} - b\overline{x} = 6 - \tfrac{648}{1639} \times 16.25 = -\tfrac{696}{1639}.$$

Hence,

$$\underline{\underline{y = -\tfrac{696}{1639} + \tfrac{648}{1639}x}} \text{ or } \underline{\underline{y = -0.4246 + 0.3954x \text{ (4 sf)}}}.$$

(b) Hence find the equation of the regression line of f on m. (3)

Solution

$$y = -\tfrac{696}{1639} + \tfrac{648}{1639}x \Rightarrow f - 100 = -\tfrac{696}{1639} + \tfrac{648}{1639}(m - 250)$$
$$\Rightarrow \underline{\underline{f = 0.7346 + 0.3954m}}.$$

(c) Predict the amount of fuel used on a journey of 235 miles. (1)

Solution

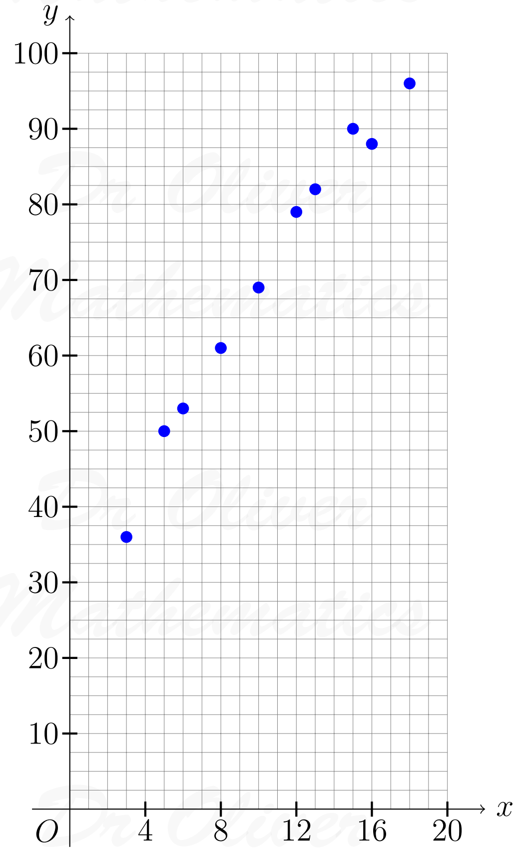
$$m = 235 \Rightarrow f = 93.6536 = \underline{\underline{93.7 \text{ (1 dp)}}}.$$

5. A manufacturer stores drums of chemicals. During storage, evaporation takes place. A random sample of 10 drums was taken and the time in storage, x weeks, and the evaporation loss, y ml, are shown in the table below.

x	3	5	6	8	10	12	13	15	16	18
y	36	50	53	61	69	79	82	90	88	96

(a) Draw a scatter diagram to represent these data. (3)

Solution



(b) Give a reason to support fitting a regression model of the form $y = a + bx$ to these data. (1)

Solution

The points lie close to a straight line.

(c) Find, to 2 decimal places, the value of a and the value of b. You may use (7)

$$\Sigma x^2 = 1\,352,\ \Sigma y^2 = 53\,112,\ \text{ and } \Sigma xy = 8\,354.$$

Solution

$$\overline{x} = \frac{\Sigma x}{n} = \frac{106}{10} = 10.6$$
$$\overline{y} = \frac{\Sigma y}{n} = \frac{704}{10} = 70.4$$
$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1\,352 - \frac{106^2}{10} = 228.4$$
$$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 8\,354 - \frac{106 \times 704}{10} = 891.6$$
$$b = \frac{S_{xy}}{S_{xx}} = \frac{891.6}{228.4} = \tfrac{2229}{571}$$
$$a = \overline{y} - b\overline{x} = 70.4 - \tfrac{2229}{571} \times 10.6 = \tfrac{16\,571}{571}.$$

Hence,

$$y = 29.02 + 3.90x \text{ (2 dp)}.$$

(d) Give an interpretation of the value of b. (1)

Solution

For every extra week in storage, another 3.90 ml of chemical evaporates.

(e) Using your model, predict the amount of evaporation that would take place after (2)

(i) 19 weeks,

Solution

$$x = 19 \Rightarrow y = 103.15.$$

(ii) 35 weeks.

Solution

$$x = 35 \Rightarrow y = 165.55.$$

(f) Comment, with a reason, on the reliability of each of your predictions. (4)

Solution

(i) The value is close to range of x, so this seems reasonably reliable.
(ii) Well outside range of x; it could be unreliable since no evidence that model will continue to hold.

6. A metallurgist measured the length, l mm, of a copper rod at various temperatures, t°C, and recorded the following results.

t	l
20.4	2 461.12
27.3	2 461.41
32.1	2 461.73
39.0	2 461.88
42.9	2 462.03
49.7	2 462.37
58.3	2 462.69
67.4	2 463.05

The results were then coded such that $x = t$ and $y = l - 2\,460.00$.

(a) Calculate S_{xy} and S_{xx}. You may use (5)

$$\Sigma\, x^2 = 15\,965.01 \text{ and } \Sigma\, xy = 757.467.$$

Solution

$$\overline{x} = \frac{\Sigma\, x}{n} = \frac{337.1}{8} = 42.1375$$
$$\overline{y} = \frac{\Sigma\, y}{n} = \frac{16.28}{8} = 2.035$$
$$S_{xx} = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 15\,965.01 - \frac{337.1^2}{8} = \underline{\underline{1\,760.458\,75}}$$
$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 757.467 - \frac{337.1 \times 16.28}{8} = \underline{\underline{71.4685}}$$

(b) Find the equation of the regression line of y on x in the form $y = a + bx$. (5)

Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{71.4685}{1\,760.458\,75} = 0.040\,596\,520\,65$$
$$a = \overline{y} - b\overline{x} = 2.035 - 0.040\,596\,520\,65 \times 42.1375 = 0.324\,364\,111\,1.$$

Hence,

$$\underline{\underline{y = 0.3244 + 0.0406x \text{ (4 sf)}}}.$$

(c) Estimate the length of the rod at 40°C. (3)

Solution

$$x = 40 \Rightarrow y = 1.9484 \Rightarrow \underline{\underline{l = 2461.9484}}.$$

(d) Find the equation of the regression line of l on t. (2)

Solution

$$\begin{aligned} y = 0.3244 + 0.0406x \Rightarrow l - 2460 &= 0.3244 + 0.0406t \\ \Rightarrow \underline{\underline{l}} &\underline{\underline{= 2460.3244 + 0.0406t}}. \end{aligned}$$

(e) Estimate the length of the rod at 90°C. (1)

Solution

$$t = 90 \Rightarrow \underline{\underline{l = 2463.9784}}.$$

(f) Comment on the reliability of your estimate in part (e). (2)

Solution

It is outside the range of data and so it may not be reliable.

7. As part of a statistics project, Gill collected data relating to the length of time, to the nearest minute, spent by shoppers in a supermarket and the amount of money they spent. Her data for a random sample of 10 shoppers are summarised in the table below, where t represents time and £m the amount spent over £20.

t (minutes)	£m
15	−3
23	17
5	−19
16	4
30	12
6	−9
32	27
23	6
35	20
27	6

(a) Write down the actual amount spent by the shopper who was in the supermarket for 15 minutes. (1)

> **Solution**
> £17.

(b) Calculate S_{tt}, S_{mm}, and S_{tm}. You may use (6)

$$\Sigma\, t^2 = 5478,\ \Sigma\, m^2 = 2101,\ \text{and } \Sigma\, tm = 2485.$$

> **Solution**

$$\begin{aligned}
\bar{t} &= \frac{\Sigma\, t}{n} = \frac{212}{10} = 21.2 \\
\overline{m} &= \frac{\Sigma\, m}{n} = \frac{61}{10} = 6.1 \\
S_{tt} &= \Sigma\, t^2 - \frac{(\Sigma\, t)^2}{n} = 5478 - \frac{212^2}{10} = \underline{\underline{983.6}} \\
S_{mm} &= \Sigma\, m^2 - \frac{(\Sigma\, m)^2}{n} = 2101 - \frac{61^2}{10} = \underline{\underline{1728.9}} \\
S_{tm} &= \Sigma\, tm - \frac{(\Sigma\, t)(\Sigma\, m)}{n} = 2485 - \frac{(212)(61)}{10} = \underline{\underline{1191.8}}.
\end{aligned}$$

(c) Calculate the value of the product moment correlation coefficient between t and m. (3)

> **Solution**

$$\begin{aligned}
r &= \frac{S_{tm}}{\sqrt{S_{tt} S_{mm}}} \\
&= \frac{1191.8}{\sqrt{983.6 \times 1728.9}} \\
&= 0.913\,922\,109\,6 \text{ (FCD)} \\
&= \underline{\underline{0.914 \text{ (3 dp)}}}.
\end{aligned}$$

(d) Write down the value of the product moment correlation coefficient between t and the actual amount spent. Give a reason to justify your value. (2)

Solution

It is $\underline{\underline{r = 0.914 \text{ (3 dp)}}}$ as coding does not change the coefficient.

On another day Gill collected similar data. For these data the product moment correlation coefficient was 0.178.

(e) Give an interpretation to both of these coefficients. (2)

Solution

$r = 0.914$ – The longer you are in the supermarket, the more you will spend.
$r = 0.178$ – Different amounts spent for same time shopping.

(f) Suggest a practical reason why these two values are so different. (1)

Solution

e.g., might spend short time buying one expensive item, might spend a long time checking for bargains, talking, buying lots of cheap items, etc.

8. A young family were looking for a new 3 bedroom semi-detached house. A local survey recorded the price x, in £1000, and the distance y, in miles, from the station of such houses. The following summary statistics were provided

$$S_{xx} = 113\,573,\ S_{yy} = 8.657,\ \text{and } S_{xy} = -808.917.$$

(a) Use these values to calculate the product moment correlation coefficient. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{-808.917}{\sqrt{113\,573 \times 8.657}} \\ &= -0.815\,798\,129 \quad \text{(FCD)} \\ &= \underline{\underline{-0.816 \text{ (3 dp)}}}. \end{aligned}$$

(b) Give an interpretation of your answer to part (a). (1)

Solution

e.g., houses are cheaper further away from the station.

Another family asked for the distances to be measured in km rather than miles.

(c) State the value of the product moment correlation coefficient in this case. (1)

Solution

-0.816 (3 dp).

9. A student is investigating the relationship between the price (y pence) of 100g of chocolate and the percentage (x%) of cocoa solids in the chocolate. The following data is obtained.

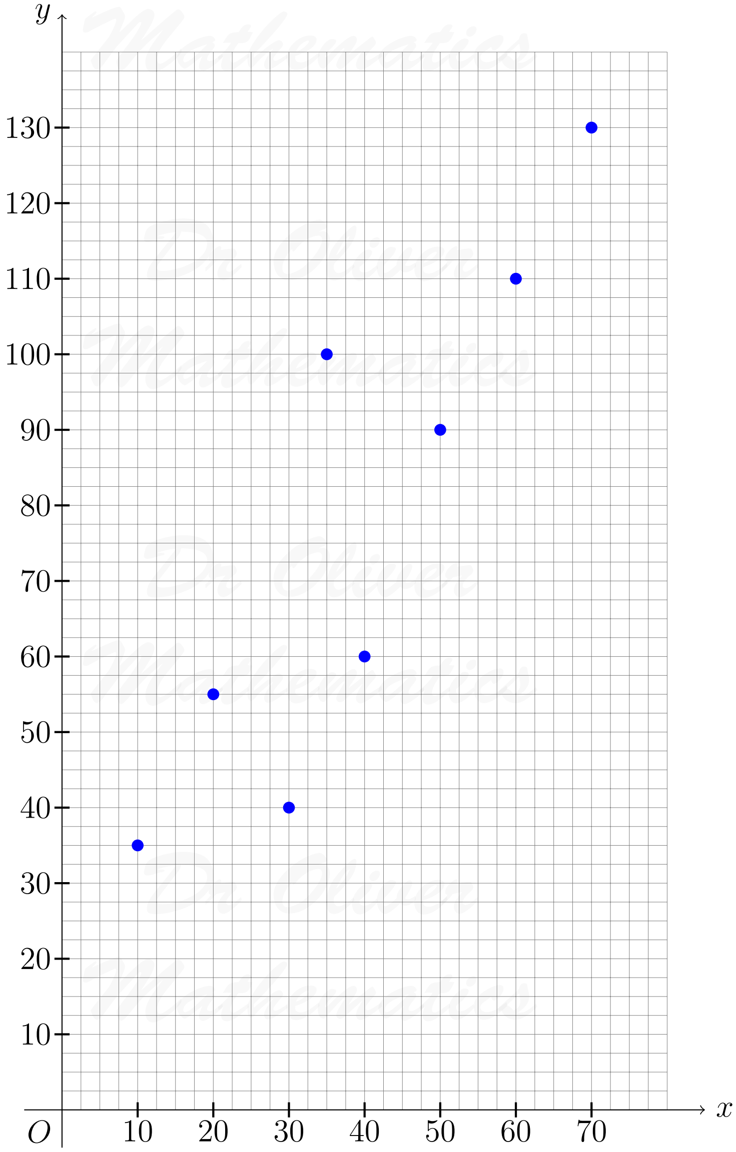
Chocolate brand	A	B	C	D	E	F	G	H
x	10	20	30	35	40	50	60	70
y	35	55	40	100	60	90	110	130

You may use

$$\Sigma x = 315,\ \Sigma x^2 = 15\,225,\ \Sigma y = 620,\ \Sigma y^2 = 56\,550,\ \text{and } \Sigma xy = 28\,750.$$

(a) Draw a scatter diagram to represent these data. (2)

Solution



(b) Show that $S_{xy} = 4\,337.5$ and find S_{xx} . (3)

Solution

$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 28\,750 - \frac{(315)(620)}{8} = \underline{\underline{4\,337.5}}$$

$$S_{xx} = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 15\,225 - \frac{315^2}{8} = \underline{\underline{2\,821.875}}.$$

The student believes that a linear relationship of the form $y = a + bx$ could be used to describe these data.

(c) Use linear regression to find the value of a and the value of b, giving your answers to 1 decimal place. (4)

Solution

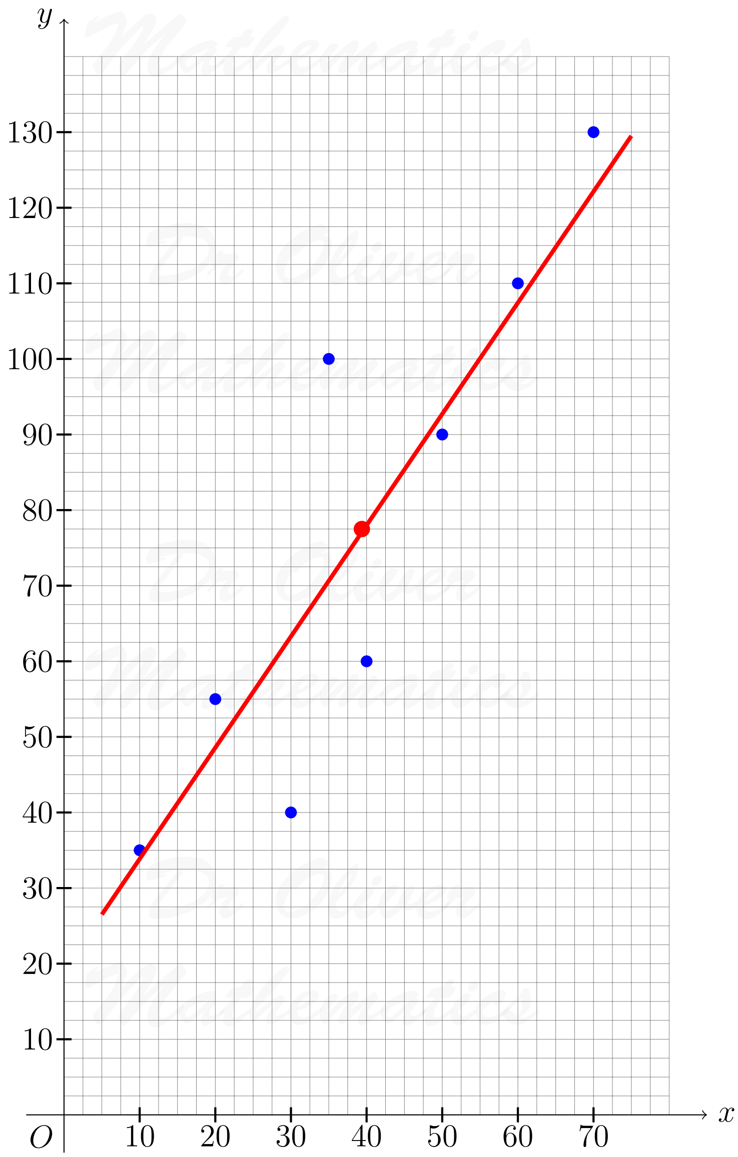
$$\begin{aligned}
\overline{x} &= \frac{\Sigma x}{n} = \frac{315}{8} = 39.375 \\
\overline{y} &= \frac{\Sigma y}{n} = \frac{620}{8} = 77.5 \\
b &= \frac{S_{xy}}{S_{xx}} = \frac{4\,337.5}{2\,821.875} = 1\tfrac{485}{903} \\
a &= \overline{y} - b\overline{x} = 77.5 - 1\tfrac{485}{903} \times 39.375 = 16\tfrac{42}{43}.
\end{aligned}$$

Hence,

$$\underline{\underline{y = 17.0 + 1.5x \text{ (1 dp)}}}.$$

(d) Draw the regression line on your scatter diagram. (2)

Solution



The student believes that one brand of chocolate is overpriced.

(e) Use the scatter diagram to (4)

(i) state which brand is overpriced,

Solution
35% cocoa (Brand D) because it is a long way above the line of regression.

(ii) suggest a fair price for this brand.

Solution

About $17.0 + 1.5 \times 35 = \underline{\underline{69.5 \text{ p}}}$.

Give reasons for both your answers.

10. A personnel manager wants to find out if a test carried out during an employee's interview and a skills assessment at the end of basic training is a guide to performance after working for the company for one year. The table below shows the results of the interview test of 10 employees and their performance after one year.

Employee	A	B	C	D	E	F	G	H	I	J
Interview test, x%	65	71	79	77	85	78	85	90	81	62
Performance after one year, y%	65	74	82	64	87	78	61	65	79	69

You may use

$$\Sigma\, x^2 = 60\,475, \Sigma\, y^2 = 53\,122, \text{ and } \Sigma\, xy = 56\,076.$$

(a) Showing your working clearly, calculate the product moment correlation coefficient between the interview test and the performance after one year. (5)

Solution

$$\begin{aligned}
\Sigma\, x &= 773 \\
\Sigma\, y &= 724 \\
S_{xy} &= \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 56\,076 - \frac{(773)(724)}{10} = 110.8 \\
S_{xx} &= \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 60\,475 - \frac{773^2}{10} = 722.1 \\
S_{yy} &= \Sigma\, y^2 - \frac{(\Sigma\, y)^2}{n} = 53\,122 - \frac{724^2}{10} = 704.4 \\
r &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \\
&= \frac{110.8}{\sqrt{722.1 \times 704.4}} \\
&= 0.155\,357\,210\,3 \quad \text{(FCD)} \\
&= \underline{\underline{0.155 \text{ (3 dp)}}}.
\end{aligned}$$

The product moment correlation coefficient between the skills assessment and the performance after one year is -0.156 to 3 significant figures.

(b) Use your answer to part (a) to comment on whether or not the interview test and skills assessment are a guide to the performance after one year. Give clear reasons for your answers. (2)

Solution

Both are weak correlation. Neither score is a good indication of future performance although interview test is slightly better since its correlation is positive.

11. A second hand car dealer has 10 cars for sale. She decides to investigate the link between the age of the cars, x years, and the mileage, y thousand miles. The data collected from the cars are shown in the table below.

Age, x (years)	2	2.5	3	4	4.5	4.5	5	3	6	6.5
Milage y (thousands)	22	34	33	37	40	45	49	30	58	58

You may assume that

$$\Sigma\, x = 41,\ \Sigma\, y = 406,\ \Sigma\, x^2 = 188,\ \text{and}\ \Sigma\, xy = 1818.5.$$

(a) Find S_{xx} and S_{xy}. (2)

Solution

$$S_{xx} = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 188 - \frac{41^2}{10} = \underline{\underline{19.9}}$$
$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 1818.5 - \frac{(41)(406)}{10} = \underline{\underline{153.9}}.$$

(b) Find the equation of the least squares regression line in the form $y = a + bx$. Give the values of a and b to 2 decimal places. (4)

Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{153.9}{19.9} = 7\tfrac{146}{199}$$
$$a = \overline{y} - b\overline{x} = 40.6 - 7\tfrac{146}{199} \times 4.1 = 8\tfrac{355}{398}.$$

Hence,

$$\underline{\underline{y = 8.89 + 7.73x \text{ (2 dp)}}}.$$

(c) Give a practical interpretation of the slope b. (1)

> **Solution**
>
> Every time a car is a year older, the mileage increases 7730 miles.

(d) Using your answer to part (b), find the mileage predicted by the regression line for a 5 year old car. (2)

> **Solution**
>
> $y = 8.89 + 7.73 \times 5 = 47.54$; the mileage is 47 540 miles.

12. Crickets make a noise. The pitch, v kHz, of the noise made by a cricket was recorded at 15 different temperatures, t°C. These data are summarised below.

$$\Sigma\, t^2 = 10\,922.81,\ \Sigma\, v^2 = 42.3356,\ \Sigma\, tv = 677.971,\ \Sigma\, t = 401.3,\ \text{and } \Sigma\, v = 25.08.$$

(a) Find S_{tt}, S_{vv}, and S_{tv} for these data. (4)

> **Solution**
>
> $$\begin{aligned} S_{tt} &= \Sigma\, t^2 - \frac{(\Sigma\, t)^2}{n} = 10\,922.81 - \frac{401.3^2}{15} = 186.697\,\dot{3} \\ S_{vv} &= \Sigma\, v^2 - \frac{(\Sigma\, v)^2}{n} = 42.3356 - \frac{25.08^2}{15} = 0.401\,84 \\ S_{tv} &= \Sigma\, tv - \frac{(\Sigma\, t)(\Sigma\, v)}{n} = 677.971 - \frac{(401.3)(25.08)}{15} = 6.997\,4. \end{aligned}$$

(b) Find the product moment correlation coefficient between t and v. (3)

> **Solution**
>
> $$\begin{aligned} r &= \frac{S_{tv}}{\sqrt{S_{tt}S_{vv}}} \\ &= \frac{16.997\,4}{\sqrt{186.697\,\dot{3} \times 0.401\,84}} \\ &= 0.807\,869\,218\,1 \quad \text{(FCD)} \\ &= 0.808 \text{ (3 dp)}. \end{aligned}$$

(c) State, with a reason, which variable is the explanatory variable. (2)

Solution

t is the explanatory variable as we can control temperature but not frequency of noise.

(d) Give a reason to support fitting a regression model of the form $v = a + bt$ to these data. (1)

Solution

There is a strong correlation between the values of t and v.

(e) Find the value of a and the value of b. Give your answers to 3 significant figures. (4)

Solution

$$\begin{aligned} b &= \frac{S_{tv}}{S_{tt}} = \frac{6.997\,4}{186.697\dot{3}} \\ &= 0.037\,479\,914\,01 \text{ (FCD)} \\ &= \underline{\underline{0.037\,5 \text{ (3 sf)}}} \\ a &= \overline{y} - b\overline{x} = \frac{25.08}{15} - 0.037\,479\ldots \times \frac{401.3}{15} \\ &= 0.669\,287\,367\,1 \text{ (FCD)} \\ &= \underline{\underline{0.669 \text{ (3 sf)}}}. \end{aligned}$$

(f) Using this model, predict the pitch of the noise at 19°C. (1)

Solution

$v = 0.669 + 0.037\,5 \times 19 = \underline{\underline{1.381\,5}}$.

13. A teacher is monitoring the progress of students using a computer based revision course. The improvement in performance, y marks, is recorded for each student along with the time, x hours, that the student spent using the revision course. The results for a random sample of 10 students are recorded below.

x hours	1.0	3.5	4.0	1.5	1.3	0.5	1.8	2.5	2.3	3.0
y marks	5	30	27	10	-3	-5	7	15	-10	20

You may use

$$\Sigma\, x = 21.4,\ \Sigma\, y = 96,\ \Sigma\, x^2 = 57.22, \text{ and } \Sigma\, xy = 313.7.$$

(a) Calculate S_{xx} and S_{xy}. (3)

Solution

$$S_{xx} = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 57.22 - \frac{21.4^2}{10} = \underline{\underline{11.424}}$$
$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 313.7 - \frac{(21.4)(96)}{10} = \underline{\underline{108.26}}.$$

(b) Find the equation of the least squares regression line of y on x in the form $y = a + bx$. (4)

Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{108.26}{11.424} = 9.476\,540\,616 \text{ (FCD)}$$
$$a = \overline{y} - b\overline{x} = 9.6 - 9.476\ldots \times 2.14 = -10.679\,796\,92 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{y = -10.7 + 9.48x \text{ (3 sf)}}}.$$

(c) Give an interpretation of the gradient of your regression line. (1)

Solution

For every hour they spend on revision, their marks go up by 9.48.

Rosemary spends 3.3 hours using the revision course.

(d) Predict her improvement in marks. (2)

Solution

$y = -10.7 + 9.48 \times 3.3 =$ 20 to 21 marks.

Lee spends 8 hours using the revision course claiming that this should give him an improvement in performance of over 60 marks.

(e) Comment on Lee's claim. (1)

Solution

The model may not be valid since 8 hours outside the range $0.5 \leqslant x \leqslant 4$ hours.

14. In a study of how students use their mobile telephones, the phone usage of a random sample of 10 students was examined for a particular week. The total length of calls, y minutes, for the 11 students were 17, 23, 35, 36, 51, 53, 54, 55, 60, and 77.

(a) Show that S_{yy} for the 10 students is 2966.9. (3)

Solution

$$\begin{aligned}\Sigma y &= 17 + 23 + \ldots + 77 = 461\\ \Sigma y^2 &= 17^2 + 23^2 + \ldots + 77^2 = 24\,219\\ S_{yy} &= \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 24\,219 - \frac{461^2}{10} = 2\,966.9,\end{aligned}$$

as required.

These 10 students were each asked how many text messages, x, they sent in the same week. The values of S_{xx} and S_{xy} for these 10 students are $S_{xx} = 3\,463.6$ and $S_{xy} = -18.3$.

(b) Calculate the product moment correlation coefficient between the number of text messages sent and the total length of calls for these 10 students. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\\ &= \frac{-18.3}{\sqrt{3\,463.6 \times 2\,966.9}}\\ &= -0.005\,550\,672\,581 \text{ (FCD)}\\ &= -0.006 \text{ (3 dp)}.\end{aligned}$$

A parent believes that a student who sends a large number of text messages will spend fewer minutes on calls.

(c) Comment on this belief in the light of your calculation in part (b). (1)

Solution

r suggests correlation is close to zero so the parent's claim is not justified.

15. The volume of a sample of gas is kept constant. The gas is heated and the pressure, p, is measured at 10 different temperatures, t. The results are summarised below.

$$\Sigma\, p = 445,\ \Sigma\, p^2 = 38\,125,\ \Sigma\, t = 240,\ \Sigma\, t^2 = 27\,520,\ \text{and}\ \Sigma\, pt = 26\,830.$$

(a) Find S_{pp} and S_{pt}. (3)

Solution

$$S_{pp} = \Sigma\, p^2 - \frac{(\Sigma\, p)^2}{n} = 38\,125 - \frac{445^2}{10} = \underline{\underline{18\,322.5}}$$
$$S_{pt} = \Sigma\, pt - \frac{(\Sigma\, p)(\Sigma\, t)}{n} = 26\,830 - \frac{(445)(240)}{10} = \underline{\underline{16\,150}}.$$

Given that $S_{tt} = 21\,760$,

(b) calculate the product moment correlation coefficient. (2)

Solution

$$\begin{aligned} r &= \frac{S_{pt}}{\sqrt{S_{pp}S_{tt}}} \\ &= \frac{16\,150}{\sqrt{18\,322.5 \times 21\,760}} \\ &= 0.808\,817\,829 \text{ (FCD)} \\ &= \underline{\underline{0.809 \text{ (3 dp)}}}. \end{aligned}$$

(c) Give an interpretation of your answer to part (b). (1)

Solution

e.g., as the temperature increases, the pressure increases.

16. The weight, w grams, and the length, l mm, of 10 randomly selected newborn turtles are given in the table below.

l	49.0	52.0	53.0	54.5	54.1	53.4	50.0	51.6	49.5	51.2
w	29	32	34	39	38	35	30	31	29	30

You may use

$$S_{ll} = 33.381,\ S_{wl} = 59.99,\ \text{and}\ S_{ww} = 120.1.$$

(a) Find the equation of the regression line of w on l in the form $w = a + bl$. (5)

Solution

$$\bar{l} = \frac{\Sigma l}{n} = \frac{518.3}{10} = 51.83$$
$$\overline{w} = \frac{\Sigma w}{n} = \frac{327}{10} = 32.7$$
$$b = \frac{S_{wl}}{S_{ll}} = \frac{59.99}{33.381} = 1.797\,130\,010\,4 \text{ (FCD)}$$
$$a = \overline{y} - b\overline{x} = 32.7 - 1.797\ldots \times 51.83 = -60.445\,253\,29 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{w = -60.4 + 1.80l \text{ (3 sf)}}}.$$

(b) Use your regression line to estimate the weight of a newborn turtle of length 60 mm. (2)

Solution

$w = -60.4 + 1.80 \times 60 = \underline{\underline{47.6}}$.

(c) Comment on the reliability of your estimate giving a reason for your answer. (2)

Solution

The model <u>may not be valid</u> since 60 mm outside the range $49 \leqslant l \leqslant 54.5$ hours.

17. The blood pressures, p mmHg, and the ages, $t$years, of 7 hospital patients are shown in the table below.

Patient	A	B	C	D	E	F	G
t	42	74	48	35	56	26	60
p	98	130	120	88	182	80	135

You may use

$$\Sigma t = 341,\ \Sigma p = 833,\ \Sigma t^2 = 18\,181,\ \Sigma p^2 = 106\,397,\ \text{and } \Sigma tp = 42\,948.$$

(a) Find S_{pp}, S_{tp}, and S_{tt} for these data. (4)

Solution

$$S_{tt} = \Sigma\, t^2 - \frac{(\Sigma\, t)^2}{n} = 18\,181 - \frac{341^2}{7} = \underline{\underline{1569\tfrac{3}{7}}}$$

$$S_{pp} = \Sigma\, p^2 - \frac{(\Sigma\, p)^2}{n} = 106\,397 - \frac{833^2}{7} = \underline{\underline{7270}}$$

$$S_{tp} = \Sigma\, tv - \frac{(\Sigma\, t)(\Sigma\, p)}{n} = 42\,948 - \frac{(341)(833)}{7} = \underline{\underline{2369}}.$$

(b) Calculate the product moment correlation coefficient for these data. (3)

Solution

$$\begin{aligned} r &= \frac{S_{tp}}{\sqrt{S_{pp} S_{tt}}} \\ &= \frac{2369}{\sqrt{7270 \times 1569\frac{3}{7}}} \\ &= 0.701\,337\,527\,5 \text{ (FCD)} \\ &= \underline{\underline{0.701 \text{ (3 dp)}}}. \end{aligned}$$

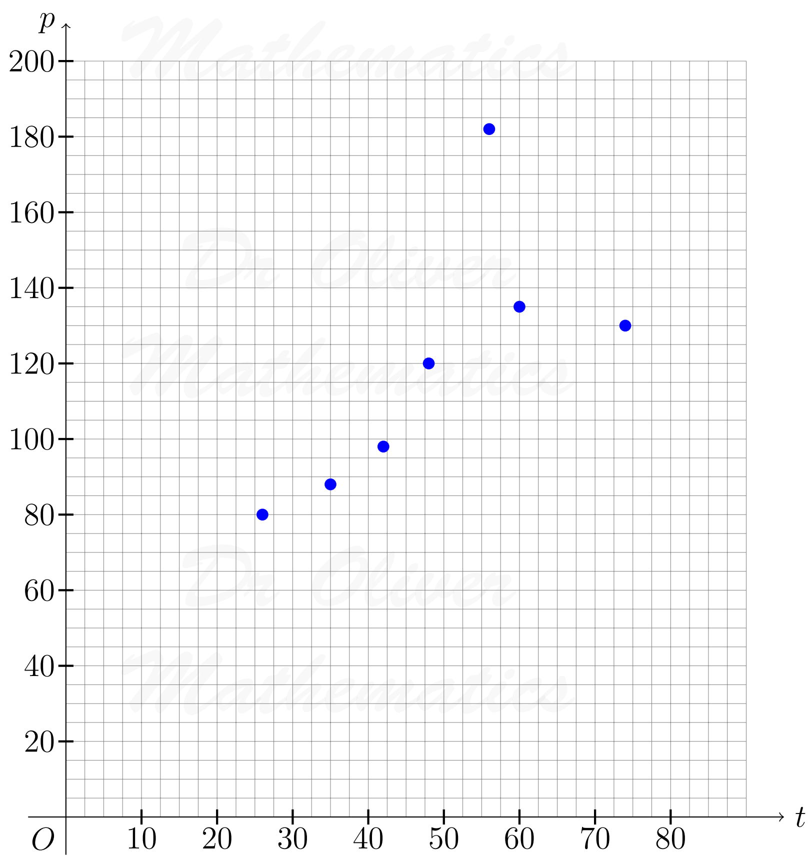
(c) Interpret the correlation coefficient. (1)

Solution

PMCC shows positive correlation. Hence, older patients have higher blood pressure.

(d) Draw the scatter diagram of blood pressure against age for these 7 patients. (2)

Solution



(e) Find the equation of the regression line of p on t. (4)

Solution

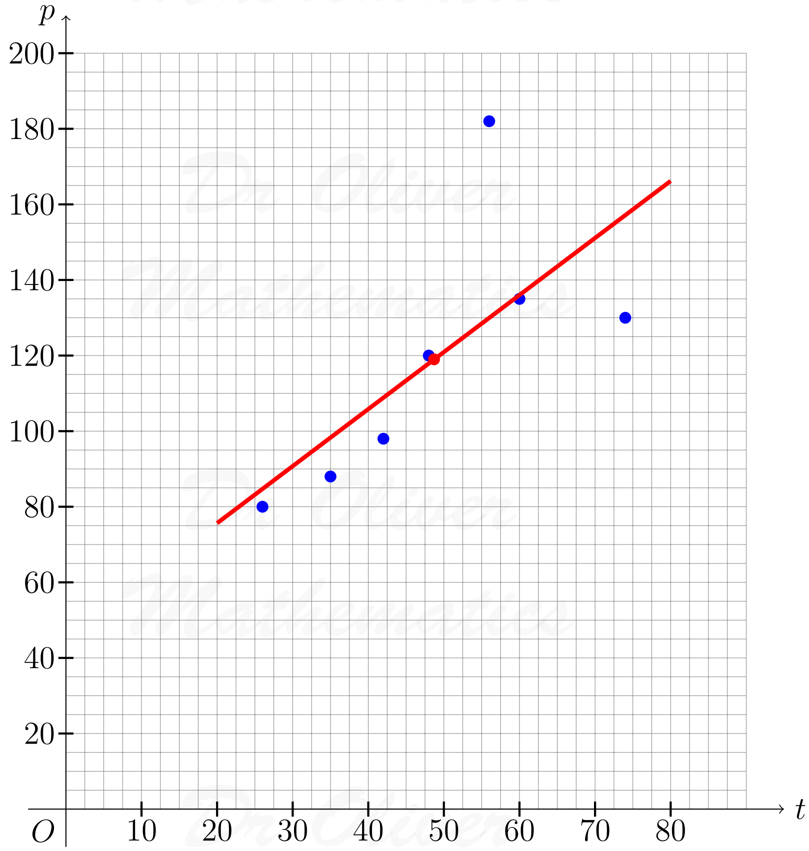
$$\bar{t} = \frac{\Sigma\, t}{n} = \frac{341}{7} = 48\tfrac{5}{7}$$
$$\bar{p} = \frac{\Sigma\, p}{n} = \frac{833}{7} = 119$$
$$b = \frac{S_{tp}}{S_{tt}} = \frac{2369}{1569\frac{3}{7}} = 1.509\,466\,594 \text{ (FCD)}$$
$$a = \bar{y} - b\bar{x} = 119 - 1.509\ldots \times 48\tfrac{5}{7} = 45.467\,413\,07 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{p = 45.47 + 1.509t \text{ (4 sf)}}}.$$

(f) Plot your regression line on your scatter diagram. (2)

Solution



(g) Use your regression line to estimate the blood pressure of a 40 year old patient. (2)

Solution

$p = 45.47 + 1.509 \times 40 = \underline{\underline{105.83}}$.

18. Gary compared the total attendance, x, at home matches and the total number of goals, y, scored at home during a season for each of 12 football teams playing in a league. He correctly calculated:

$$S_{xx} = 1\,022\,500,\ S_{yy} = 130.9,\ \text{and}\ S_{xy} = 8\,825.$$

(a) Calculate the product moment correlation coefficient for these data. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{8\,825}{\sqrt{1\,022\,500 \times 130.9}} \\ &= 0.762\,804\,475\,5 \text{ (FCD)} \\ &= \underline{\underline{0.763 \text{ (3 dp)}}}. \end{aligned}$$

(b) Interpret the value of the correlation coefficient. (1)

Solution

PMCC shows positive correlation. Hence, teams with high attendance scored more goals.

Helen was given the same data to analyse. In view of the large numbers involved she decided to divide the attendance figures by 100. She then calculated the product moment correlation coefficient between $\frac{x}{100}$ and y.

(c) Write down the value Helen should have obtained. (1)

Solution

0.763 (3 dp).

19. A travel agent sells flights to different destinations from *Beerow* airport. The distance d, measured in 100 km, of the destination from the airport and the fare £f are recorded for a random sample of 6 destinations.

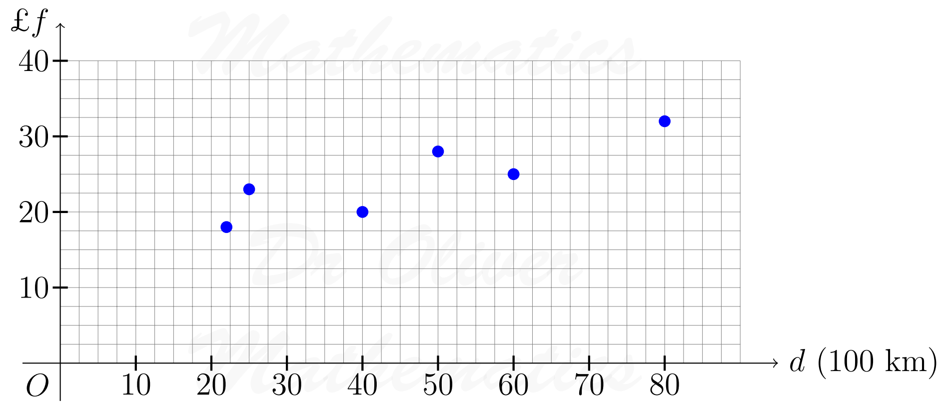
Destination	A	B	C	D	E	F
d	2.2	4.0	6.0	2.5	8.0	5.0
f	18	20	25	23	32	28

You may use

$$\Sigma\, d^2 = 152.09,\ \Sigma\, f^2 = 3\,686,\ \text{and}\ \Sigma\, fd = 723.1.$$

(a) Complete a scatter diagram to illustrate this information. (2)

Solution



(b) Explain why a linear regression model may be appropriate to describe the relationship between f and d. (1)

Solution

e.g., the points lie reasonably close to a straight line.

(c) Calculate S_{dd} and S_{fd}. (4)

Solution

$$\overline{d} = \frac{\Sigma d}{n} = \frac{27.7}{6} = \tfrac{277}{60}$$
$$\overline{f} = \frac{\Sigma f}{n} = \frac{146}{6} = \tfrac{73}{3}$$
$$S_{dd} = \Sigma d^2 - \frac{(\Sigma d)^2}{n} = 152.09 - \frac{27.7^2}{6} = \tfrac{581}{24}$$
$$S_{fd} = \Sigma fd - \frac{(\Sigma f)(\Sigma d)}{n} = 723.1 - \frac{(146)(27.7)}{6} = \tfrac{736}{15}$$

(d) Calculate the equation of the regression line of f on d giving your answer in the form $f = a + bd$. (4)

Solution

$$b = \frac{S_{fd}}{S_{dd}} = \frac{\frac{736}{15}}{\frac{581}{24}} = 2\tfrac{78}{2905}$$
$$a = \overline{f} - b\overline{d} = \tfrac{73}{3} - 2\tfrac{78}{2905} \times \tfrac{277}{60} = 14.976\,041\,31 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{f = 15.0 + 2.03d \text{ (3 sf)}}}.$$

(e) Give an interpretation of the value of b. (1)

Solution

It costs $\underline{\underline{\text{2.03p to fly a kilometre}}}$.

Jane is planning her holiday and wishes to fly from *Beerow* airport to a destination t km away. A rival travel agent charges 5p per km.

(f) Find the range of values of t for which the first travel agent is cheaper than the rival. (2)

Solution

$$\begin{aligned} 15.0 + 2.03d < 5d &\Rightarrow 2.97d > 15.0 \\ &\Rightarrow d > 5.\dot{0}\dot{5}; \end{aligned}$$

so, $\underline{\underline{t > 505 \text{ km}}}$.

20. A random sample of 50 salmon was caught by a scientist. He recorded the length l cm and weight w kg of each salmon. The following summary statistics were calculated from these data:

$$\Sigma\, l = 4\,027,\ \Sigma\, l^2 = 327\,754.5,\ \Sigma\, w = 357.1,\ \Sigma\, lw = 29\,330.5,\ \text{and } S_{ww} = 289.6.$$

(a) Find S_{ll} and S_{lw}. (3)

Solution

$$S_{ll} = \Sigma\, l^2 - \frac{(\Sigma\, l)^2}{n} = 327\,754.5 - \frac{4\,027^2}{50} = \underline{\underline{3\,419.92}}$$
$$S_{lw} = \Sigma\, lw - \frac{(\Sigma\, l)(\Sigma\, w)}{n} = 29\,330.5 - \frac{(4\,027)(357.1)}{50} = \underline{\underline{569.666}}$$

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between l and w. (2)

Solution

$$\begin{aligned} r &= \frac{S_{lw}}{\sqrt{S_{ll}S_{ww}}} \\ &= \frac{569.666}{\sqrt{3\,419.92 \times 289.6}} \\ &= 0.572\,417\,691 \text{ (FCD)} \\ &= \underline{\underline{0.572 \text{ (3 sf)}}}. \end{aligned}$$

(c) Give an interpretation of your coefficient. (1)

Solution

e.g., as the length of the salmon increases, the weight increases.

21. A farmer collected data on the annual rainfall, x cm, and the annual yield of peas, p tonnes per acre. The data for annual rainfall was coded using $v = \frac{x-5}{10}$ and the following statistics were found:

$$S_{vv} = 5.753,\ S_{pv} = 1.688,\ S_{pp} = 1.168,\ \overline{p} = 3.22, \text{ and } \overline{v} = 4.42.$$

(a) Find the equation of the regression line of p on v in the form $p = a + bv$. (4)

Solution

$$b = \frac{S_{pv}}{S_{vv}} = \frac{1.688}{5.753} = 0.293\,412\,132\,8 \text{ (FCD)}$$
$$a = \overline{f} - b\overline{d} = 3.22 - 0.293\ldots \times 4.42 = 1.923\,118\,373 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{p = 1.92 + 0.293v \text{ (3 sf)}}}.$$

(b) Using your regression line estimate the annual yield of peas per acre when the annual rainfall is 85 cm. (2)

Solution

Now,

$$p = 1.92 + 0.0293(x - 5)$$

and

$$x = 85 \Rightarrow p = 1.92 + 0.0293(85 - 5) = \underline{\underline{4.264}}.$$

22. On a particular day the height above sea level, x metres, and the mid-day temperature, y°C, were recorded in 8 north European towns. These data are summarised below:

$$S_{xx} = 3\,535\,237.5,\ \Sigma\, y = 181\ \Sigma\, y^2 = 4\,305,\ \text{and}\ S_{xy} = -23\,726.25.$$

(a) Find S_{yy}. (2)

Solution

$$S_{yy} = \Sigma\, y^2 - \frac{(\Sigma\, y)^2}{n} = 4\,305 - \frac{181^2}{8} = \underline{\underline{209.875}}.$$

(b) Calculate, to 3 significant figures, the product moment correlation coefficient for these data. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{-23\,726.25}{\sqrt{3\,535\,237.5 \times 209.875}} \\ &= -0.871\,042\,709\,1 \text{ (FCD)} \\ &= \underline{\underline{-0.871 \text{ (3 sf)}}}. \end{aligned}$$

(c) Give an interpretation of your coefficient. (1)

Solution

PMCC shows negative correlation. Hence, towns closer to sea level are warmer.

A student thought that the calculations would be simpler if the height above sea level, h, was measured in kilometres and used the variable $h = \frac{x}{1000}$ instead of x.

(d) Write down the value of S_{hh}. (1)

Solution

$\underline{\underline{S_{hh} = 3\,535\,237.5}}$.

(e) Write down the value of the correlation coefficient between h and y. (1)

Solution

$\underline{\underline{-0.871 \text{ (3 sf)}}}$.

23. A teacher took a random sample of 8 children from a class. For each child the teacher recorded the length of their left foot, f cm, and their height, h cm. The results are given in the table below.

f	23	26	23	22	27	24	20	21
h	135	144	134	136	140	134	130	132

You may use

$$\Sigma f = 186,\ \Sigma h = 1\,085,\ S_{ff} = 39.5,\ S_{hh} = 139.875,\ \text{and } \Sigma fh = 25\,291.$$

(a) Calculate S_{fh}. (2)

Solution

$$S_{fh} = \Sigma fh - \frac{(\Sigma f)(\Sigma h)}{n} = 25\,291 - \frac{(186)(1\,085)}{8} = \underline{\underline{64.75}}.$$

(b) Find the equation of the regression line of h on f in the form $h = a + bf$. Give the value of a and the value of b correct to 3 significant figures. (5)

Solution

$$b = \frac{S_{fh}}{S_{ff}} = \frac{64.75}{39.5} = 1\tfrac{101}{158}$$
$$a = \overline{h} - b\overline{f} = \frac{1\,085}{8} - 1\tfrac{101}{158} \times \frac{186}{8} = 97\tfrac{81}{158}.$$

Hence,

$$\underline{\underline{h = 97.5 + 1.64f\ \text{(3 sf)}}}.$$

(c) Use your equation to estimate the height of a child with a left foot length of 25 cm. (2)

Solution

$$h = 97.5 + 1.64 \times 25 = \underline{\underline{138.5}}.$$

(d) Comment on the reliability of your estimate in (c), giving a reason for your answer. (2)

Solution

It should be reliable, since 25 cm is inside the range $20 \leqslant f \leqslant 27$.

The left foot length of the teacher is 25 cm.

(e) Give a reason why the equation in (b) should not be used to estimate the teacher's height. (1)

Solution

e.g., line is for children: a different equation would apply to adults; children are still growing and height will increase more than foot length.

24. The age, t years, and weight, w grams, of each of 10 coins were recorded. These data are summarised below.

$$\Sigma t^2 = 2\,688,\ \Sigma tw = 1\,760.62,\ \Sigma t = 158,\ \Sigma w = 111.75,\ \text{and } S_{ww} = 0.16.$$

(a) Find S_{tt} and S_{tw} for these data. (3)

Solution

$$S_{tt} = \Sigma t^2 - \frac{(\Sigma t)^2}{n} = 2\,688 - \frac{158^2}{10} = \underline{\underline{191.6}}$$
$$S_{tw} = \Sigma tw - \frac{(\Sigma w)(\Sigma w)}{n} = 1\,760.62 - \frac{(158)(111.75)}{10} = \underline{\underline{-5.03}}$$

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between t and w. (2)

Solution

$$\begin{aligned} r &= \frac{S_{tw}}{\sqrt{S_{tt}S_{ww}}} \\ &= \frac{-5.03}{\sqrt{191.6 \times 0.16}} \\ &= -0.908\,469\,269\,9 \text{ (FCD)} \\ &= \underline{\underline{-0.908 \text{ (3 sf)}}}. \end{aligned}$$

(c) Find the equation of the regression line of w on t in the form $w = a + bt$. (4)

Solution

$$b = \frac{S_{tw}}{S_{tt}} = \frac{-5.03}{191.6} = -\tfrac{503}{19\,160}$$
$$a = \overline{h} - b\overline{f} = 11.175 + \tfrac{503}{19\,160} \times 15.8 = 11.589\,791\,23 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{w = 11.6 - 0.0263t \text{ (3 sf)}}}.$$

(d) State, with a reason, which variable is the explanatory variable. (2)

Solution

The explanatory variable is the age of each coin. This is because the age is set and the mass varies.

(e) Using this model, estimate (2)

(i) the weight of a coin which is 5 years old,

Solution

$$w = 11.6 - 0.0263 \times 5 = 11.468\,5 = \underline{\underline{11.5 \text{ g (1 dp)}}}.$$

(ii) the effect of an increase of 4 years in age on the weight of a coin.

Solution

A decrease of

$$0.0263 \times 4 = \underline{\underline{0.1052 \text{ g}}}.$$

It was discovered that a coin in the original sample, which was 5 years old and weighed 20 grams, was a fake.

(f) State, without any further calculations, whether the exclusion of this coin would increase or decrease the value of the product moment correlation coefficient. Give a reason for your answer. (2)

Solution

It will decrease; by removing it, the other nine coins will result in a better linear fit so r will be closer to -1.

25. A bank reviews its customer records at the end of each month to find out how many customers have become unemployed, u, and how many have had their house repossessed,

h, during that month. The bank codes the data using variables $x = \frac{u-100}{3}$ and $y = \frac{h-20}{7}$. The results for the 12 months of 2009 are summarised below.

$$\Sigma x = 477,\ S_{xx} = 5\,606.25,\ \Sigma y = 480,\ S_{yy} = 4\,244,\ \text{and } \Sigma xy = 23\,070.$$

(a) Calculate the value of the product moment correlation coefficient for x and y. (3)

> **Solution**
>
> $$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 23\,070 - \frac{(477)(480)}{12} = 3\,990$$
>
> and
>
> $$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{3\,990}{\sqrt{5\,606.25 \times 4\,244}} \\ &= 0.817\,991\,852\,3 \text{ (FCD)} \\ &= \underline{\underline{0.818 \text{ (3 dp)}}}. \end{aligned}$$

(b) Write down the product moment correlation coefficient for u and h. (1)

> **Solution**
>
> 0.818 (3 dp).

The bank claims that an increase in unemployment among its customers is associated with an increase in house repossessions.

(c) State, with a reason, whether or not the bank's claim is supported by these data. (2)

> **Solution**
>
> PMCC shows positive correlation. Hence, increase in unemployment is matched by increase in house repossessions.

26. A scientist is researching whether or not birds of prey exposed to pollutants lay eggs with thinner shells. He collects a random sample of egg shells from each of 6 different nests and tests for pollutant level, p, and measures the thinning of the shell, t. The results are shown in the table below.

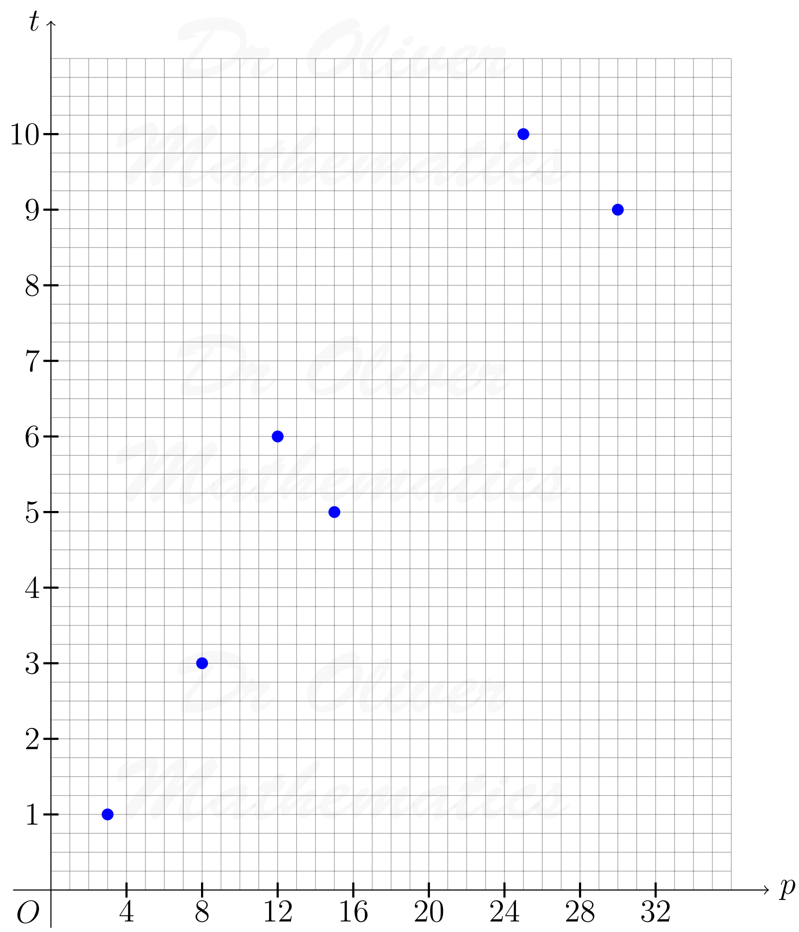
p	3	8	30	25	15	12
t	1	3	9	10	5	6

You may use

$$\Sigma p^2 = 1\,967 \text{ and } \Sigma pt = 694.$$

(a) Draw a scatter diagram to represent these data. (2)

Solution



(b) Explain why a linear regression model may be appropriate to describe the relationship between p and t. (1)

Solution

e.g., the points lie reasonably close to a straight line.

(c) Calculate the value of S_{pt} and the value of S_{pp}. (4)

Solution

$$\overline{p} = \Sigma\, p = 93$$
$$\overline{t} = \Sigma\, t = 34$$
$$S_{pt} = \Sigma\, pt - \frac{(\Sigma\, p)(\Sigma\, t)}{n} = 694 - \frac{(93)(34)}{6} = \underline{\underline{167}}$$
$$S_{pp} = \Sigma\, p^2 - \frac{(\Sigma\, p)^2}{n} = 1\,967 - \frac{93^2}{6} = \underline{\underline{525.5}}$$

(d) Find the equation of the regression line of t on p, giving your answer in the form $t = a + bp$. (4)

Solution

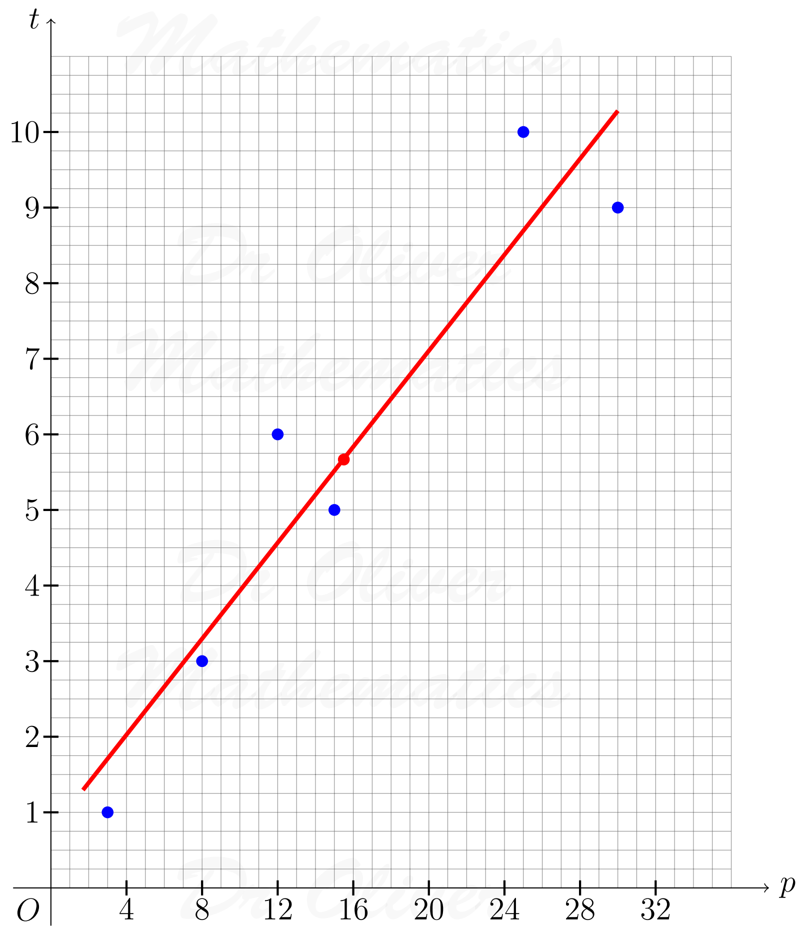
$$b = \frac{S_{pt}}{S_{pp}} = \frac{167}{525.5} = \tfrac{334}{1\,051}$$
$$a = \overline{h} - b\overline{f} = \frac{34}{6} - \tfrac{334}{1\,051} \times \frac{93}{6}\frac{34}{6} = \tfrac{2\,336}{3\,153}.$$

Hence,

$$\underline{\underline{t = 0.741 + 0.318p \text{ (3 sf)}}}.$$

(e) Plot the point $(\overline{p}, \overline{t})$ and draw the regression line on your scatter diagram. (2)

Solution



The scientist reviews similar studies and finds that pollutant levels above 16 are likely to result in the death of a chick soon after hatching.

(f) Estimate the minimum thinning of the shell that is likely to result in the death of a chick. (2)

Solution

$$t = 0.741 + 0.318 \times 16 = \underline{\underline{5.829}}.$$

27. A teacher asked a random sample of 10 students to record the number of hours of television, t, they watched in the week before their mock exam. She then calculated their grade, g, in their mock exam. The results are summarised as follows.

$$\Sigma\, t = 258,\ \Sigma\, t^2 = 8\,702,\ \Sigma\, g = 63.6,\ S_{gg} = 7.864,\ \text{and}\ \Sigma\, gt = 1\,550.2.$$

(a) Find S_{tt} and S_{gt}. (3)

Solution

$$S_{tt} = \Sigma\, t^2 - \frac{(\Sigma\, t)^2}{n} = 8\,702 - \frac{258^2}{10} = \underline{\underline{2\,045.6}}$$
$$S_{gt} = \Sigma\, gt - \frac{(\Sigma\, t)(\Sigma\, t)}{n} = 1\,550.2 - \frac{(258)(63.6)}{10} = \underline{\underline{-90.68}}$$

(b) Calculate, to 3 significant figures, the product moment correlation coefficient between t and g. (2)

Solution

$$\begin{aligned} r &= \frac{S_{gt}}{\sqrt{S_{tt}S_{gg}}} \\ &= \frac{-90.68}{\sqrt{2\,045.6 \times 7.864}} \\ &= -0.714\,956\,142\,2 \text{ (FCD)} \\ &= \underline{\underline{-0.715 \text{ (3 sf)}}}. \end{aligned}$$

The teacher also recorded the number of hours of revision, v, these 10 students completed during the week before their mock exam. The correlation coefficient between t and v was -0.753.

(c) Describe, giving a reason, the nature of the correlation you would expect to find between v and g. (2)

Solution

PMCC shows positive correlation. Hence, the more time they spent on revision, the higher the mark.

28. A biologist is comparing the intervals (m seconds) between the mating calls of a certain species of tree frog and the surrounding temperature (t°C). The following results were obtained.

t°C	8	13	14	15	15	20	25	30
m secs	6.5	4.5	6	5	4	3	2	1

You may use

$$\Sigma\, tm = 469.5,\ S_{tt} = 354,\ \text{and}\ S_{mm} = 25.5.$$

(a) Show that $S_{tm} = -90.5$. (4)

Solution

$$\Sigma\, t = 140$$
$$\Sigma\, m = 32$$
$$S_{tm} = \Sigma\, tm - \frac{(\Sigma\, t)(\Sigma\, m)}{n} = 469.5 - \frac{(140)(32)}{8} = \underline{\underline{-90.5}}$$

(b) Find the equation of the regression line of m on t giving your answer in the form $m = a + bt$. (4)

Solution

$$b = \frac{S_{mt}}{S_{tt}} = \frac{-90.5}{354} = -\tfrac{181}{708}$$
$$a = \overline{h} - b\overline{f} = \frac{32}{8} + \tfrac{181}{708} \times \frac{140}{8} = 8\tfrac{671}{1\,416}.$$

Hence,

$$\underline{\underline{m = 8.47 - 0.256t \text{ (3 sf)}}}.$$

(c) Use your regression line to estimate the time interval between mating calls when the surrounding temperature is 10°C. (1)

Solution

$$m = 8.47 - 0.256 \times 10 = \underline{\underline{5.91}}.$$

(d) Comment on the reliability of this estimate, giving a reason for your answer. (1)

Solution

It should be <u>reliable</u>, since 10°C is inside the range $8 \leqslant t \leqslant 30$.

29. A meteorologist believes that there is a relationship between the height above sea level, h m, and the air temperature, t°C. Data is collected at the same time from 9 different places on the same mountain. The data is summarised in the table below.

h	1400	1100	260	840	900	550	1230	100	770
t	3	10	20	9	10	13	5	24	16

You may assume that

$\Sigma h = 7\,150$, $\Sigma t = 110$, $\Sigma h^2 = 7\,171\,500$, $\Sigma t^2 = 1\,716$, $\Sigma th = 64\,980$, and $S_{tt} = 371.56$.

(a) Calculate S_{th} and S_{hh}. Give your answers to 3 significant figures. (3)

Solution

$$S_{th} = \Sigma\, th - \frac{(\Sigma\, t)(\Sigma\, h)}{n} = 64\,980 - \frac{(110)(7\,150)}{9} = \underline{\underline{-22\,400 \text{ (3 sf)}}}$$

$$S_{hh} = \Sigma\, h^2 - \frac{(\Sigma\, h)^2}{n} = 7\,171\,500 - \frac{7\,150^2}{9} = \underline{\underline{1\,490\,000 \text{ (3 sf)}}}$$

(b) Calculate the product moment correlation coefficient for this data. (2)

Solution

$$\begin{aligned} r &= \frac{S_{th}}{\sqrt{S_{tt} S_{hh}}} \\ &= \frac{-22\,400}{\sqrt{371.56 \times 1\,490\,000}} \\ &= -0.952\,007\,581\,6 \text{ (FCD)} \\ &= \underline{\underline{-0.952 \text{ (3 sf)}}}. \end{aligned}$$

(c) State whether or not your value supports the use of a regression equation to predict the air temperature at different heights on this mountain. Give a reason for your answer. (1)

Solution

PMCC shows negative correlation. Hence, <u>the higher above sea level, the cooler it is</u>.

(d) Find the equation of the regression line of t on h giving your answer in the form $t = a + bh$. (4)

Solution

$$b = \frac{S_{th}}{S_{hh}} = \frac{-22\,400}{1\,490\,000} = -\tfrac{56}{3\,725}$$

$$a = \overline{h} - b\overline{f} = \frac{110}{9} + \tfrac{56}{3\,725} \times \frac{7\,150}{9} = 24\tfrac{74}{447}.$$

Hence,

$$\underline{\underline{t = 24.2 - 0.0150h \text{ (3 sf)}}}.$$

(e) Interpret the value of b. (1)

Solution

e.g., for every 1 m you go up, the temperature decreases by 0.0150°C.

(f) Estimate the difference in air temperature between a height of 500 m and a height of 1000 m. (2)

Solution

$$500 \times 0.0150 = 7.5;$$

hence, a decrease of 7.5°C.

30. Sammy is studying the number of units of gas, g, and the number of units of electricity, e, used in her house each week. A random sample of 10 weeks use was recorded and the data for each week were coded so that $x = \frac{g-60}{4}$ and $y = \frac{e}{10}$. The results for the coded data are summarised below:

$$\Sigma\, x = 48.0,\ \Sigma\, y = 58.0,\ S_{xx} = 312.1,\ S_{yy} = 2.10,\ \text{and } S_{xy} = 18.35.$$

(a) Find the equation of the regression line of y on x in the form $y = a + bx$. Give the values of a and b correct to 3 significant figures. (4)

Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{18.35}{312.1} = \tfrac{367}{6\,242}$$
$$a = \overline{y} - b\overline{x} = \tfrac{58}{10} + \tfrac{367}{6\,242} \times \tfrac{48}{10} = 5.517\,782\,762 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{y = 5.52 + 0.0588 \text{ (3 sf)}}}.$$

(b) Hence find the equation of the regression line of e on g in the form $e = c + dg$. Give the values of c and d correct to 2 significant figures. (4)

Solution

$$\begin{aligned} y = 5.52\ldots + \tfrac{367}{6\,242}x \Rightarrow \frac{e}{10} &= 5.52\ldots + \tfrac{367}{6\,242}\left(\frac{g-60}{4}\right) \\ \Rightarrow \frac{e}{10} &= 5.52\ldots + \tfrac{367}{24\,968}(g-60) \\ \Rightarrow e &= 46.358\,538\,93 + 0.146\,988\,144\,8g \text{ (FCD)} \\ \Rightarrow e &= 46 + 0.15g \text{ (2 sf)}. \end{aligned}$$

(c) Use your regression equation to estimate the number of units of electricity used in a week when 100 units of gas were used. (2)

Solution

$$e = 46 + 0.15 \times 100 = 61.$$

31. A researcher believes that parents with a short family name tended to give their children a long first name. A random sample of 10 children was selected and the number of letters in their family name, x, and the number of letters in their first name, y, were recorded. The data are summarised as:

$$\Sigma\, x = 60,\ \Sigma\, y = 61, \Sigma\, y^2 = 393,\ \Sigma\, xy = 382, \text{ and } S_{xx} = 28.$$

(a) Find S_{yy} and S_{xy}. (3)

Solution

$$\begin{aligned} S_{yy} &= \Sigma\, y^2 - \frac{(\Sigma\, y)^2}{n} = 393 - \frac{61^2}{10} = 20.9 \\ S_{xy} &= \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 382 - \frac{(60)(61)}{10} = 16 \end{aligned}$$

(b) Calculate the product moment correlation coefficient, r, between x and y. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{16}{\sqrt{28 \times 20.9}} \\ &= 0.661\,405\,533\,4 \text{ (FCD)} \\ &= \underline{\underline{0.661 \text{ (3 dp)}}}. \end{aligned}$$

(c) State, giving a reason, whether or not these data support the researcher's belief. (2)

Solution

Researcher's belief suggests negative correlation whereas the data suggests positive correlation. So the data does not support researcher's belief.

The researcher decides to add a child with family name "Turner" to the sample.

(d) Using the definition (2)

$$S_{xx} = \Sigma\,(x - \overline{x})^2,$$

state the new value of S_{xx} giving a reason for your answer.

Solution

Since it is the same (six letters), S_{xx} is unchanged.

Given that the addition of the child with family name "Turner" to the sample leads to an increase in S_{yy},

(e) use the definition (2)

$$S_{xy} = \Sigma\,(x - \overline{x})(y - \overline{y}),$$

to determine whether or not the value of r will increase, decrease, or stay the same. Give a reason for your answer.

Solution

S_{xy} will remain the same (six letters). However,

$$S_{yy} = \Sigma\,y^2 - \frac{(\Sigma\,y)^2}{n} = 429 - \frac{67^2}{11} = 20.\dot{9}\dot{0}$$

will increase therefore r will decrease.

32. The table shows data on the number of visitors to the UK in a month, v (1000s), and the amount of money they spent, m (£ millions), for each of 8 months.

v	2450	2480	2540	2420	2350	2290	2400	2460
m	1370	1350	1400	1330	1270	1210	1330	1350

You may use:

$$S_{vv} = 42\,587.5,\ S_{vm} = 31\,512.5,\ S_{mm} = 25\,187.5,\ \Sigma\, v = 19\,390,\ \text{and}\ \Sigma\, m = 10\,610.$$

(a) Find the product moment correlation coefficient between m and v. (2)

Solution

$$\begin{aligned} r &= \frac{S_{vw}}{\sqrt{S_{vv}S_{ww}}} \\ &= \frac{31\,512.5}{\sqrt{42\,587.5 \times 25\,187.5}} \\ &= 0.962\,164\,332\,8 \text{ (FCD)} \\ &= \underline{\underline{0.962 \text{ (3 dp)}}}. \end{aligned}$$

(b) Give a reason to support fitting a regression model of the form $m = a + bv$ to these data. (1)

Solution

PMCC shows positive correlation. Hence, the higher the visitor numbers, the more they spend.

(c) Find the value of b correct to 3 decimal places. (2)

Solution

$$b = \frac{S_{vw}}{S_{vv}} = \tfrac{2\,521}{3\,407} = \underline{\underline{0.740 \text{ (3 dp)}}}.$$

(d) Find the equation of the regression line of m on v. (2)

Solution

$$a = \overline{m} - b\overline{v} = \tfrac{10\,610}{8} - \tfrac{2\,521}{3\,407} \times \tfrac{19\,390}{8} = -467.196\,947\,5 \text{ (FCD)}$$

Hence,

$$\underline{\underline{m = -467 + 0.740v \text{ (3 sf)}}}.$$

(e) Interpret your value of b. (2)

Solution

b is the money spent per visitor, which is £740.

(f) Use your answer to part (d) to estimate the amount of money spent when the number of visitors to the UK in a month is 2 500 000. (2)

Solution

$$m = -467 + 0.740 \times 2\,500 = 1\,383;$$

hence, it is £1 383 000 000.

(g) Comment on the reliability of your estimate in part (f). Give a reason for your answer. (2)

Solution

The value of money spent is reliable as 2 500 is within the range $2\,290 \leqslant v \leqslant 2\,540$.

33. A large company is analysing how much money it spends on paper in its offices every year. The number of employees, x, and the amount of money spent on paper, p (£ hundreds), in 8 randomly selected offices are given in the table below.

x	8	9	12	14	7	3	16	19
p	40.5	36.1	30.4	39.4	32.6	31.1	43.4	45.7

You may us:

$$\Sigma\, x^2 = 1\,160,\ \Sigma\, p = 299.2,\ \Sigma\, p^2 = 11\,422,\ \text{and}\ \Sigma\, xp = 3\,449.5.$$

(a) Show that $S_{pp} = 231.92$ and find the value of S_{xx} and the value of S_{xp}. (5)

Solution

$$\Sigma\, x = 88$$

$$S_{pp} = \Sigma\, p^2 - \frac{(\Sigma\, p)^2}{n} = 11\,422 - \frac{299.2^2}{8} = \underline{\underline{231.92}}$$

$$S_{xx} = \Sigma\, x^2 - \frac{(\Sigma\, x)^2}{n} = 1\,160 - \frac{88^2}{8} = \underline{\underline{192}}$$

$$S_{xp} = \Sigma\, xp - \frac{(\Sigma\, x)(\Sigma\, p)}{n} = 3\,449.5 - \frac{(88)(299.2)}{8} = \underline{\underline{158.3}}.$$

(b) Calculate the product moment correlation coefficient between x and p. (2)

Solution

$$\begin{aligned} r &= \frac{S_{xp}}{\sqrt{S_{xx}S_{pp}}} \\ &= \frac{158.3}{\sqrt{192 \times 231.92}} \\ &= 0.750\,172\,603\,1 \text{ (FCD)} \\ &= \underline{\underline{0.750 \text{ (3 dp)}}}. \end{aligned}$$

The equation of the regression line of p on x is given in the form $p = a + bx$.

(c) Show that, to 3 significant figures, $b = 0.824$ and find the value of a. (4)

Solution

$$b = \frac{S_{xp}}{S_{xx}} = \frac{158.3}{192} = 0.824\,479\,1\dot{6} = \underline{\underline{0.824 \text{ (3 sf)}}}$$

$$a = \overline{p} - b\overline{x} = \tfrac{299.2}{8} + \tfrac{1\,583}{1\,920} \times \tfrac{88}{8} = 28\tfrac{127}{384} = \underline{\underline{28.3 \text{ (3 sf)}}}.$$

(d) Estimate the amount of money spent on paper in an office with 10 employees. (2)

Solution

$$p = 28.3 + 0.824 \times 10 = 36.54;$$

hence, $\underline{\underline{\text{£}3700}}$.

(e) Explain the effect each additional employee has on the amount of money spent on paper. (1)

Solution

An extra employee means £82.40 of photocopying.

Later the company realised it had made a mistake in adding up its costs, p. The true costs were actually half of the values recorded. The product moment correlation coefficient and the equation of the linear regression line are recalculated using this information.

(f) Write down the new value of (2)

(i) the product moment correlation coefficient,

Solution

$r = 0.750$ (3 dp).

(ii) the gradient of the regression line.

Solution

$b = 0.412$ (3 sf).

34. An estate agent recorded the price per square metre, p £/m^2, for 7 two-bedroom houses. He then coded the data using the coding $q = \frac{p-a}{b}$, where a and b are positive constants. His results are shown in the table below.

p	1840	1848	1830	1824	1819	1834	1850
q	4.0	4.8	3.0	2.4	1.9	3.4	5.0

(a) Find the value of a and the value of b. (2)

Solution

$$4 = \frac{1840 - a}{b} \Rightarrow 4b = 1840 - a$$

and

$$3 = \frac{1830 - a}{b} \Rightarrow 3b = 1830 - a$$

subtract:

$$b = 10 \text{ and } a = 1800.$$

The estate agent also recorded the distance, d km, of each house from the nearest train station. The results are summarised below.

$$S_{dd} = 1.02, \ S_{qq} = 8.22, \ \text{and } S_{dq} = -2.17.$$

(b) Calculate the product moment correlation coefficient between d and q. (2)

Solution

$$\begin{aligned} r &= \frac{S_{dq}}{\sqrt{S_{dd}S_{qq}}} \\ &= \frac{-2.17}{\sqrt{1.02 \times 8.22}} \\ &= -0.749\,417\,342\,8 \text{ (FCD)} \\ &= \underline{\underline{-0.749 \text{ (3 dp)}}}. \end{aligned}$$

(c) Write down the value of the product moment correlation coefficient between d and p. (1)

Solution

$\underline{\underline{-0.749 \text{ (3 dp)}}}$.

The estate agent records the price and size of 2 additional two-bedroom houses, H and J.

House	Price (£)	Size (m^2)
H	156 400	85
J	172 900	95

(d) Suggest which house is most likely to be closer to a train station. Justify your answer. (3)

Solution

$$\begin{aligned} \text{House } H: &\quad \frac{156\,400}{85} = 1840 \\ \text{House } J: &\quad \frac{172\,900}{95} = 1820. \end{aligned}$$

PMCC shows negative correlation which means the higher the price (per square metre), the lower the distance from the train station. Hence, House H is likely to be closer.

35. Statistical models can provide a cheap and quick way to describe a real world situation.

(a) Give two other reasons why statistical models are used. (2)

Solution

e.g., to simplify or represent a real world problem
to improve understanding
to analyse a real world problem
we change variables easily
we change replicate easily
to make predictions or find estimates.

A scientist wants to develop a model to describe the relationship between the average daily temperature, t°C, and her household's daily energy consumption, y kWh, in winter. A random sample of the average daily temperature and her household's daily energy consumption are taken from 10 winter days and shown in the table.

x	−0.4	−0.2	0.3	0.8	1.1	1.4	1.8	2.1	2.5	2.6
y	28	30	26	25	26	27	26	24	22	21

You may use

$$\Sigma\, x^2 = 24.76,\ \Sigma\, y = 255,\ \Sigma\, xy = 283.8,\ \text{and}\ S_{xx} = 10.36.$$

(b) Find S_{xy} for these data. (3)

Solution

$$\Sigma\, x = 12$$
$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 283.8 - \frac{(12)(255)}{10} = \underline{\underline{-22.2}}.$$

(c) Find the equation of the regression line of y on x in the form $y = a + bx$. Give the value of a and the value of b to 3 significant figures. (4)

Solution

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-22.2}{10.36} = -2\tfrac{1}{7} = \underline{\underline{-2.14 \text{ (3 sf)}}}$$
$$a = \overline{y} - b\overline{x} = \tfrac{255}{10} + \tfrac{15}{7} \times \tfrac{12}{10} = 28\tfrac{1}{44} = \underline{\underline{28.1 \text{ (3 sf)}}}.$$

(d) Give an interpretation of the value of a. (1)

Solution

When the temperature is 0°C, then $\underline{\underline{\text{28.1 kWh}}}$ of energy are used.

(e) Estimate her household's daily energy consumption when the average daily temperature is 2°C. (2)

Solution

$$y = 28.1 - 2.14 \times 2 = \underline{\underline{23.82}}.$$

The scientist wants to use the linear regression model to predict her household's energy consumption in the summer.

(f) Discuss the reliability of using this model to predict her household's energy consumption in the summer. (2)

Solution

The regression model is based on temperatures from the winter; therefore, $\underline{\underline{\text{not reliable}}}$ during the summer months.

36. A biologist is studying the behaviour of bees in a hive. Once a bee has located a source of food, it returns to the hive and performs a dance to indicate to the other bees how far away the source of the food is. The dance consists of a series of wiggles. The biologist records the distance, d metres, of the food source from the hive and the average number of wiggles, w, in the dance.

d	30	50	80	100	150	400	500	650
w	0.725	1.210	1.775	2.250	3.518	6.382	8.185	9.555

You may use

$$\Sigma\, w = 33.6,\ \Sigma\, dw = 13\,833,\ S_{dd} = 394\,600,\ \text{and } S_{ww} = 80.481 \text{ (to 3 decimal places).}$$

(a) Show that $S_{dw} = 5\,601$. (2)

Solution

$$\Sigma\, d = 1\,960$$
$$S_{dw} = \Sigma\, dw - \frac{(\Sigma\, d)(\Sigma\, w)}{n} = 13\,833 - \frac{(1\,960)(33.6)}{8} = \underline{\underline{5\,601}}.$$

(b) State, giving a reason, which is the response variable. (1)

Solution

w is the <u>response variable</u> since <u>the number of wiggles depends on the distance</u>.

(c) Calculate the product moment correlation coefficient for these data. (2)

Solution

$$\begin{aligned} r &= \frac{S_{dw}}{\sqrt{S_{dd}S_{ww}}} \\ &= \frac{5\,601}{\sqrt{394\,600 \times 80.481}} \\ &= 0.993\,894\,644 \quad \text{(FCD)} \\ &= \underline{\underline{0.994 \text{ (3 dp)}}}. \end{aligned}$$

(d) Calculate the equation of the regression line of w on d, giving your answer in the form $w = a + bd$. (4)

Solution

$$b = \frac{S_{dw}}{S_{dd}} = \frac{5\,601}{394\,600} = 0.014\,194\,120\,63 \text{ (FCD)}$$
$$a = \overline{w} - b\overline{d} = \tfrac{33.6}{8} - 0.014\ldots \times \tfrac{1\,960}{8} = 0.722\,440\,446 \text{ (FCD)}.$$

Hence,

$$\underline{\underline{w = 0.722 + 0.014\,2d \text{ (3 sf)}}}.$$

A new source of food is located 350 m from the hive.

(e) (i) Use your regression equation to estimate the average number of wiggles in the corresponding dance. (2)

Solution

$$w = 0.722 + 0.0142 \times 350 = \underline{\underline{5.692}}.$$

(ii) Comment, giving a reason, on the reliability of your estimate.

Solution
It should be $\underline{\underline{\text{reliable}}}$, since 350 m is inside the range $30 \leqslant d \leqslant 650$.

37. Before going on holiday to *Seapron*, Tania records the weekly rainfall (x mm) at *Seapron* for 8 weeks during the summer. Her results are summarised as

$$\Sigma\, x = 86.8 \text{ and } \Sigma\, x^2 = 985.88.$$

(a) Find the standard deviation, σ_x, for these data. (3)

Solution

$$\sigma_x = \sqrt{\frac{985.88}{8} - \left(\frac{86.8}{8}\right)^2} = \underline{\underline{\tfrac{21\sqrt{5}}{20}}}.$$

Tania also records the number of hours of sunshine (y hours) per week at *Seapron* for these 8 weeks and obtains the following

$$\overline{y} = 58,\ \sigma_y = 9.461 \text{ (correct to 4 significant figures)},\ \text{and } \Sigma\, xy = 4\,900.5.$$

(b) Show that $S_{yy} = 716$, correct to 3 significant figures. (1)

Solution

$$S_{yy} = 8\sigma_y^2 = 8 \times 9.461^2 = \underline{\underline{716 \text{ (3 sf)}}}.$$

(c) Find S_{xy}. (2)

Solution

$$S_{xy} = \Sigma\, xy - \frac{(\Sigma\, x)(\Sigma\, y)}{n} = 4\,900.5 - \frac{(86.8)(464)}{8} = \underline{\underline{-133.9}}.$$

(d) Calculate the product moment correlation coefficient, r, for these data. (2)

Solution

Now,

$$S_{xx} = 8\sigma_x^2 = 44.1$$

and

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \frac{-133.9}{\sqrt{44.1 \times 716}} \\ &= -0.753\,537\,328\,5 \text{ (FCD)} \\ &= \underline{\underline{-0.754 \text{ (3 dp)}}}. \end{aligned}$$

During Tania's week-long holiday at *Seapron* there are 14 mm of rain and 70 hours of sunshine.

(e) State, giving a reason, what the effect of adding this information to the above data would be on the value of the product moment correlation coefficient. (2)

Solution

$r = -0.754$ (3 dp) means "high sunshine and low rain". The 8-days in the question are "high sunshine and high rain" so $\underline{\underline{r \text{ is closer to } 0}}$.

38. A clothes shop manager records the weekly sales figures, £s, and the average weekly temperature, t°C, for 6 weeks during the summer. The sales figures were coded so that $w = \frac{s}{1000}$. The data are summarised as follows:

$$S_{ww} = 50,\ \Sigma\, wt = 784,\ \Sigma\, t^2 = 2\,435,\ \Sigma\, t = 119, \text{ and } \Sigma\, w = 42.$$

(a) Find S_{wt} and S_{tt}. (3)

Solution

$$S_{wt} = \Sigma\, wt - \frac{(\Sigma\, w)(\Sigma\, t)}{n} = 784 - \frac{(42)(119)}{6} = \underline{\underline{-49}}$$

$$S_{tt} = \Sigma\, t^2 - \frac{(\Sigma\, t)^2}{n} = 2\,435 - \frac{119^2}{6} = \underline{\underline{74.8\dot{3}}}$$

(b) Write down the value of S_{ss} and the value of S_{st}. (2)

Solution
$S_{ss} = 50 \times 1\,000^2 = \underline{\underline{50\,000\,000}}$ and $S_{st} = \underline{\underline{-49\,000}}$

(c) Find the product moment correlation coefficient between s and t. (2)

Solution

$$\begin{aligned} r &= \frac{S_{st}}{\sqrt{S_{ss}S_{tt}}} \\ &= \frac{-49\,000}{\sqrt{50\,000\,000 \times 74.8\dot{3}}} \\ &= -0.801\,057\,207\,9 \text{ (FCD)} \\ &= \underline{\underline{-0.801 \text{ (3 dp)}}}. \end{aligned}$$

The manager of the clothes shop believes that a linear regression model may be appropriate to describe these data.

(d) State, giving a reason, whether or not your value of the correlation coefficient supports the manager's belief. (1)

Solution
PMCC shows strong negative correlation ($r = -0.801$); hence, does support the belief.

(e) Find the equation of the regression line of w on t, giving your answer in the form $w = a + bt$. (3)

Solution

$$\begin{aligned} b &= \frac{S_{wt}}{S_{tt}} = \frac{-49}{74.8\dot{3}} = -0.654\,788\,418\,7 \text{ (FCD)} \\ a &= \overline{w} - b\overline{t} = \tfrac{42}{6} + 654.788\ldots \times \tfrac{119}{6} = 19.986\,636\,910 \text{ (FCD)}. \end{aligned}$$

Hence,

$$\underline{\underline{w = 20.0 - 0.655t \text{ (3 sf)}}}.$$

(f) Hence find the equation of the regression line of s on t, giving your answer in the form $s = c + dt$, where c and d are correct to 3 significant figures. (2)

Solution

$$\frac{s}{1000} = 20.0 - 0.655t \Rightarrow \underline{\underline{s = 20\,000 - 655t}}.$$

(g) Using your equation in part (f), interpret the effect of a 1°C increase in average weekly temperature on weekly sales during the summer. (1)

Solution

The effect is a decrease in sales by £655.