Pearson's Correlation Coefficient

Using a line of best fit to measure correlation is subjective, and in mathematics we prefer answers to be precise and exact.

Pearson's Correlation Coefficient (*r* – sample or *ρ* - population) is used as a precise measure of the correlation between two random variables.

Given two random variables (X, Y)

$$r_{XY} = \frac{\mathrm{Cov}(X,Y)}{s_X \, s_Y}$$

$$= \frac{\overline{xy} - \overline{x}\,\overline{y}}{s_X \, s_Y}$$

$$-1 \leq r_{XY} \leq 1$$

variance

$$= E[(X - \overline{x})^2]$$

covariance

$$= E[(X - \overline{x})(Y - \overline{y})]$$

$|r| = 1$: perfect correlation

$0.6 \leqq |r| < 1$: strong correlation

$0.4 \leqq |r| < 0.6$: moderate correlation

$0.1 \leqq |r| < 0.4$: weak correlation

$0 < |r| < 0.1$: virtually none correlation

$r = 0$: no correlation

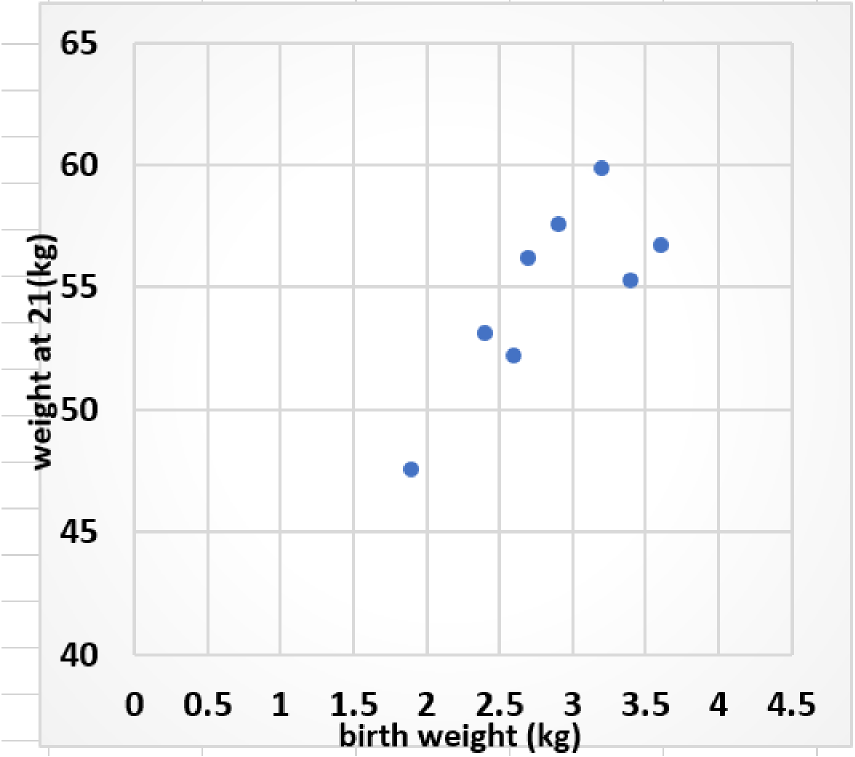
$r > 0$: positive correlation
as X increases Y increases

$r < 0$: negative correlation
as X increases Y decreases

e.g. The birth weight and weight at age 21 of eight people are given in the table below

Birth weight (kg)	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.6
Weight at 21 (kg)	47.6	53.1	52.2	56.2	57.6	59.9	55.3	56.7

(i) Construct a scatterplot of the data and from the plot how would you best describe the association of the data



There appears to be a strong positive linear correlation between the data

(ii) Calculate the correlation coefficient for this bivariate data

Let X = birth weight and Y = weight at 21

									Σ
x	1.9	2.4	2.6	2.7	2.9	3.2	3.4	3.6	**22.7**
y	47.6	53.1	52.2	56.2	57.6	59.9	55.3	56.7	**438.6**
xy	90.44	127.44	135.72	151.74	167.04	191.68	188.02	204.12	**1256.2**
$\boldsymbol{x^2}$	3.61	5.76	6.76	7.29	8.41	10.24	11.56	12.96	**66.59**
$\boldsymbol{y^2}$	2265.76	2819.61	2724.84	3158.44	3317.76	3588.01	3058.09	3214.89	**24147.4**

$$\bar{x} = \frac{22.7}{8}$$
$$= 2.8375$$

$$\bar{y} = \frac{438.6}{8}$$
$$= 54.825$$

$$\overline{xy} = \frac{1256.2}{8}$$
$$= 157.025$$

$$s_x{}^2 = \overline{x^2} - (\bar{x})^2$$
$$= \left(\frac{66.59}{8}\right) - \left(\frac{22.7}{8}\right)^2$$
$$s_x = 0.5219\ldots$$

$$s_y{}^2 = \overline{y^2} - (\bar{y})^2$$
$$= \left(\frac{24147.4}{8}\right) - \left(\frac{438.6}{8}\right)^2$$
$$s_y = 3.5559\ldots$$

$$r_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_x\, s_y}$$
$$= \frac{157.025 - 2.8375 \times 54.825}{0.5219 \times 3.5559}$$
$$\underline{= 0.7862}$$

Least-squares Regression Line

The process of fitting a straight line to bivariate data is known as **linear regression**.

This method assumes that the variables are linearly related, and works best when there are no clear outliers.

It minimises the sum of the squares of the vertical distances of each data plot to the line and ensures that the line passes through $(\overline{x},\overline{y})$

The least-squares regression line has;

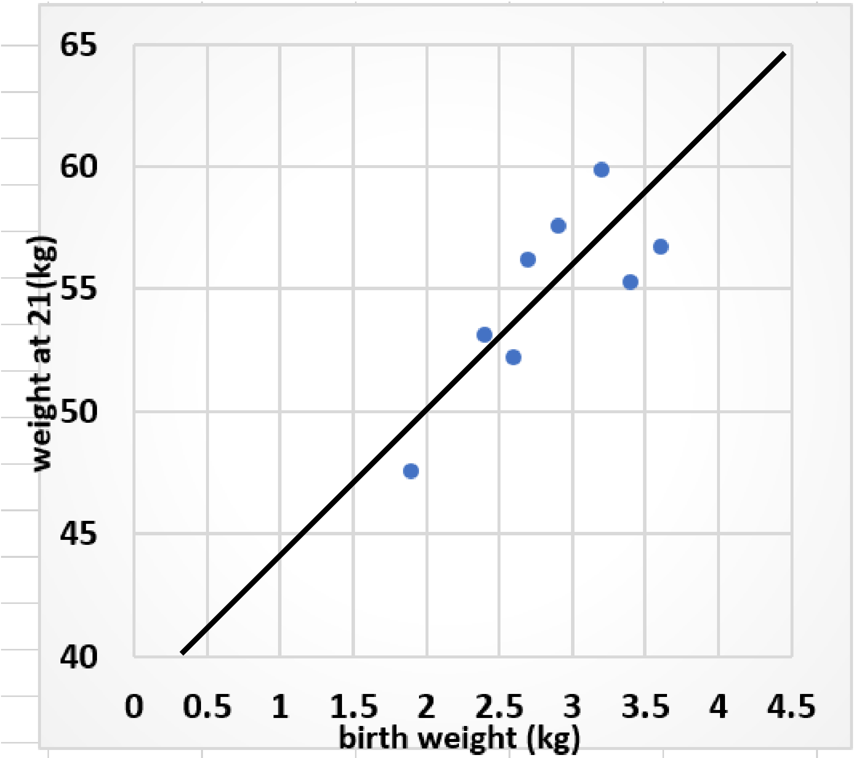
slope

$$m = \frac{\mathrm{Cov}(XY)}{\mathrm{Var}(X)} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{{s_x}^2} = \frac{r_{xy}s_y}{s_x}$$

y-intercept

$$b = \overline{y} - m\overline{x}$$

(iii) Draw a line of best fit and find its equation



Two points on the line are;

(2,50) and (0.275,40)

$$m = \frac{50 - 40}{2 - 0.275}$$
$$= 5.7971$$

$$y - 50 = 5.7971(x - 2)$$
$$y = 5.7971x + 38.4058$$

(iv) Find the least squares regression line and draw it on the scatterplot

$$m = \frac{r_{xy}s_y}{s_x}$$

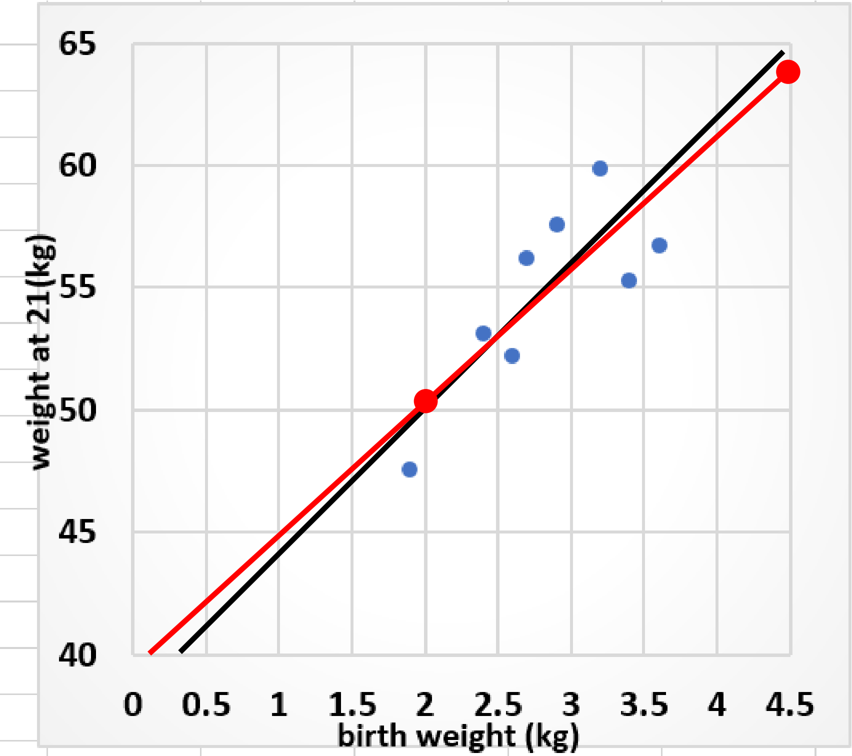
$$= \frac{0.7862 \times 3.5559}{0.5219}$$

$$= 5.36$$

$$b = 54.825 - 5.36 \times 2.8375$$

$$= 39.616$$

$$\underline{y = 5.36x + 39.62}$$



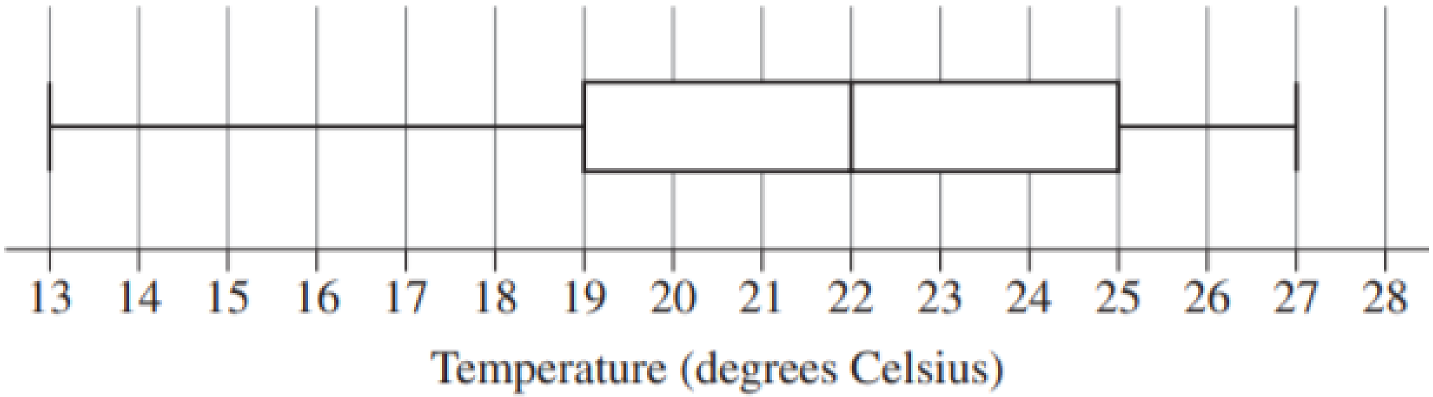
(v) 2020 Mathematics HSC Q27

A cricket is an insect. The male cricket produces a chirping sound.

A scientist wants to explore the relationship between the temperature in degrees Celsius and the number of cricket chirps heard in a 15-second time interval.

Once a day for 20 days, the scientist collects data. Based on the 20 data points, the scientist provides the information below.

* A box-plot of the temperature data is shown



* The mean temperature in the dataset is 0.525°C below the median temperature in the dataset.

* A total of 684 chirps was counted when collecting the 20 data points.

The scientist fits a least-squares regression line using the data (x, y), where x is the temperature in degrees Celsius and y is the number of chirps heard in a 15-second interval. The equation of the line is

$$y = -10.6063 + bx$$

where b is the slope of the regression line.

The least-squares regression line passes through the point $(\bar{x}, \bar{y})$where $\bar{x}$ is the sample mean of the temperature data and $\bar{y}$ is the sample mean of the chirp data.

Calculate the number of chirps expected in a 15-second interval when the temperature is 19℃. Give your answer correct to the nearest whole number.

$$\begin{aligned} \bar{x} &= 22 - 0.525 \\ &= 21.475 \end{aligned} \qquad \begin{aligned} \bar{y} &= \frac{684}{20} \\ &= 34.2 \end{aligned}$$

∴ the regression line passes through (21.475,34.2)

$$y = -10.6063 + bx$$

$$34.2 = -10.6063 + 21.475b$$

$$21.475b = 44.8063$$

$$b = \frac{44.8063}{21.475}$$

$$\text{when } x = 19;\ y = -10.6063 + \frac{44.8063}{21.475}(19)$$

$$= 29.03606088\ldots$$

You would expect 29 chirps

Exercise 15E; abdfghij in all

(use the theory formulae)