Pearson's Product Moment Correlation Coefficient

Starter

1. **(Review of last lesson)** The number of accidents per week at a factory is a Poisson random variable with parameter 2.
   (a) Find the probability that in any week chosen at random exactly 1 accident occurs.
   (b) The factory is observed for 100 weeks. Determine the expected number of weeks, to 1 d.p., in which 5 or more accidents occur.

Video : **Calculating mean and variance using a Classwiz calculator**
Enter data >> AC >> OPTN >> 2

2. For the following data, use the 6: Statistics function on your calculator to find:
   (a) the mean for X and the mean for Y
   (b) the standard deviation of X and the standard deviation of Y.

X	2130	2424	2328	2394	2280	2232	2082	2274
Y	2244	2430	2376	2388	2226	2166	2202	2208

2. Eight runners ran the same 10 km route twice — once in the dry and once when it was wet. Their times, in minutes, are below:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Dry, D	35.5	40.4	38.8	39.9	38.0	37.2	34.7	37.9
Wet, W	37.4	40.5	39.6	39.8	37.1	36.1	36.7	36.8

Use the 6: Statistics function on your calculator to find:
(a) the mean time for the dry and the wet.
(b) the standard deviation for the dry and the wet.

Notes
We are trying to find how closely related, or correlated, each set of data from the starter is i.e. is there a correlation between X and Y and between times in the dry, D, and times in the wet, W?

Please do not copy.
We could first of all find the deviations from the mean i.e. $X - \bar{X}$ and $Y - \bar{Y}$:

X	138	−156	−60	−126	−12	36	186	−6
Y	36	−150	−96	−108	54	114	78	72

What do the sum of the devotions add up to?
Zero (this is the start of the calculation for standard deviation)
Next we find the product $(x - \bar{x})(y - \bar{y})$.

$(x - \bar{x})(y - \bar{y})$	4968	23400	5760	13608	−648	4104	14508	−432

By summing them up and dividing by how many data values there are (i.e. 8) we get what is known as the ***covariance of x and y***, which is denoted by S_{xy}.

$$S_{xy} = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y}) = 8158.5$$

If we do the same for the dry vs. wet data we get:

D	−3.1	−2.3	−0.6	0.1	0.2	1	2.1	2.6
W	−1.3	−0.6	−1.9	−1.2	−0.9	1.6	1.8	2.5
$(d - \bar{d})(w - \bar{w})$	4.03	1.38	1.14	−0.12	−0.18	1.6	3.78	6.5

So the covariance of D and W is $S_{dw} = \frac{1}{n}\sum (d - \bar{d})(w - \bar{w}) \approx 2.266$

For our two sets of data we have:

$$S_{xy} = 8158.5$$
$$S_{dw} \approx 2.266$$

Are these values a good measure of the correlation of the data?

The data given in the first table is the times from the second table but in seconds rather than minutes. Therefore, we would expect that the value for the correlation is the same for both sets of data. Therefore, it is not a good value.

What happens if we divide the covariance by the product of the individual standard deviations?

$$\frac{S_{xy}}{s_x \times s_y} = \frac{8158.5}{111.2 \times 94.77} \approx 0.774$$
$$\frac{S_{dw}}{s_d \times s_w} = \frac{2.266}{1.853 \times 1.580} \approx 0.774$$

Now we have a value that agrees for both data — this value is called Pearson's product moment correlation coefficient and it is given the letter, r. It is named after the English mathematician Karl Pearson (1857-1936).

Please start copying again.

Product moment correlation coefficient

The product moment correlation coefficient, r, is used with ***bivariate*** data as a measure of ***how closely related*** two variables are ***when it is evident there could be a linear relationship***.
i.e. don't bother calculating the PMCC when it is clear the relationship is not linear.

Range of PMCC, r

The range of values of the PMCC is $-1 \leq r \leq 1$.
$r = 1$ means the points lie exactly along a straight line with a positive gradient
$r = -1$ means the points lie exactly along a straight line with a positive gradient
So r is a measure of how closely the points on the scatter graph are to the line of best fit.

N.B. It is unlikely to have $r = 1$ or $r = -1$ with real-life data.
Roughly: $0.7 \leq r < 1$ means strong positive correlation
$0.5 \leq r < 0.7$ means weak positive correlation
$-1 < r \leq -0.7$ means strong negative correlation
$-0.7 < r \leq -0.5$ means weak negative correlation

A lower value of r would be accepted in the social sciences compared to the exact sciences.
The value of r does not change with linear coding (i.e. adding or multiplying by a constant)

Formula for PMCC, r

$$r = \frac{\text{Covariance of X and Y}}{\text{Standard deviation of X} \times \text{Standard deviation of Y}}$$

It is expected that you can calculate r using your calculator.
Menu >> 6: Statistics >> 2: y=a+bx >> (Enter data) >> AC >> OPTN >> 3: Regression calc

Video (Classwiz PMCC): https://www.youtube.com/watch?v=D0EBJAQ7mUI

Notation
Different books express the formula in different ways.

Variance of x: $$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Variance of y: $$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

Covariance of x and y: $$S_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Product moment correlation coefficient, r: $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

These all appear in your formula book.

E.g. 1 Without using the special function on your calculator, find the PMCC for $\sum x_i = 22.09$, $\sum y_i = 49.7$, $\sum x_i^2 = 45.04$, $\sum y_i^2 = 244.83$, $\sum x_i y_i = 97.778$ and $n = 12$.

Working: $$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 97.778 - \frac{22.09 \times 49.7}{12} = 6.2886 \text{ (4 d.p.)}$$
...

E.g. 2 In the study of a city, the population density p (people/hectare) and the distance from the city centre d (km) was investigated by picking a number of random areas with the following results.

Area	A	B	C	D	E	F	G	H	I	J
Distance, d (km)	0.6	3.8	2.4	3.0	2.0	1.5	1.8	3.4	4.0	0.9
Population density, p (people/hectare)	50	22	14	20	33	47	25	8	16	38

Calculate the PMCC and comment on your findings.

Video: Pearson's PMCC
Video: PMCC on calculators
Video : Calculating mean and variance using a Classwiz calculator

Exercise
p75 5A Qu 1i, 2i

Summary
The product moment correlation coefficient, r, is used with ***bivariate*** data as a measure of ***how closely related*** two variables are ***when it is evident there could be a linear relationship***.
i.e. don't bother calculating the PMCC when it is clear the relationship is not linear.
The range of values of the PMCC is $-1 \leq r \leq 1$.
So r is a measure of how closely the points on the scatter graph are to the line of best fit.

Variance of x: $$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Variance of y: $$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

Covariance of x and y: $$S_{xy} = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Product moment correlation coefficient, r: $$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$