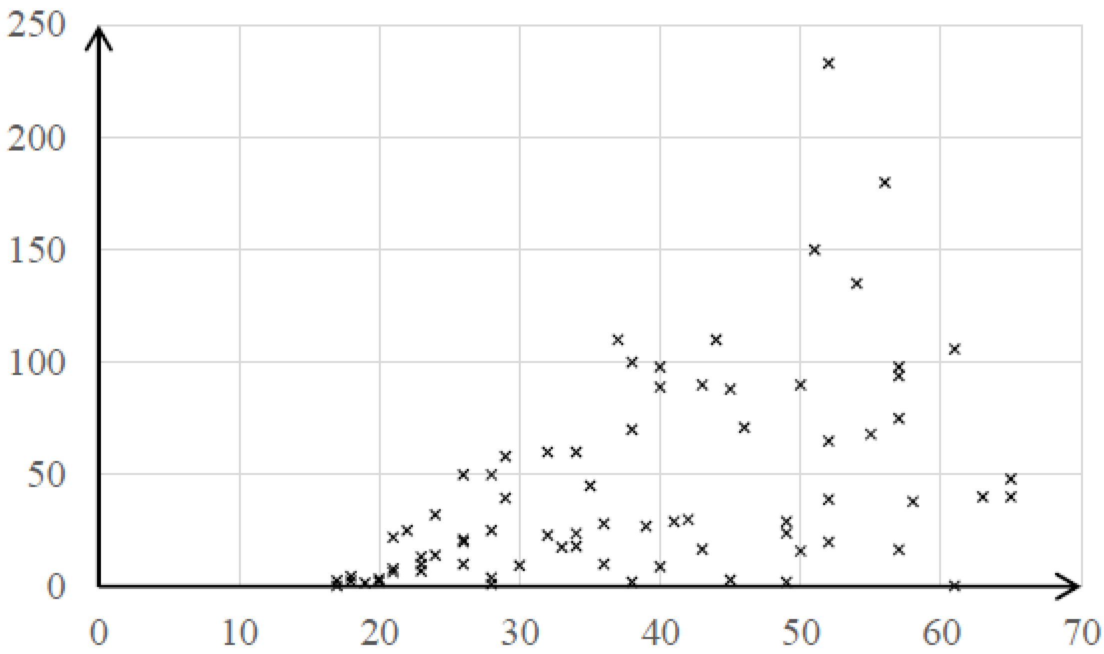
Correlation – Exercises (with Sol'ns) (8 pages; 14/2/26)

(1) Data has been collected for two variables (as shown below), and possible association between the variables is to be investigated. What issues should be considered, to ensure that the test is valid?



Solution

- Is this fresh data (rather than the data that prompted the investigation)?

- Are both variables random?

(Necessary for either rank correlation or PMCC).

- Should the outlier at (approx.) $(52, 235)$ be excluded

(if either an error, or atypical)?

- The data points don't seem to form an ellipse (concentrated at its centre), and so a bivariate normal distribution is unlikely to be appropriate (so that a PMCC wouldn't be valid)

- Is there a linear association? If not, rank correlation would be appropriate, instead of PMCC.

(2) Show that the formula $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ can be written as

$r_s = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$ when the data items are ranks.

[In other words, instead of using the formula for Spearman's coefficient, it is theoretically possible to use the standard formula for r.]

(2) Show that the formula $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ can be written as

$r_s = 1 - \frac{6\sum d_i{}^2}{n(n^2-1)}$ when the data items are ranks.

[In other words, instead of using the formula for Spearman's coefficient, it is theoretically possible to use the standard formula for r.]

Solution

$S_{xx} = \sum x_i{}^2 - \frac{(\sum x_i)^2}{n}$

As the x_i are just the numbers 1 to n in some order,

$\sum x_i{}^2 = \frac{n}{6}(n+1)(2n+1)$ and $\sum x_i = \frac{n}{2}(n+1)$

and so $S_{xx} = \frac{n}{6}(n+1)(2n+1) - \frac{n(n+1)^2}{4}$

By the same reasoning, S_{yy} will also have this value,

so the denominator of r is

$\frac{n}{6}(n+1)(2n+1) - \frac{n(n+1)^2}{4} = \frac{n}{12}(n+1)\{4n+2-(3n+3)\}$

$= \frac{n}{12}(n+1)(n-1)$

Then $S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$

Now $\sum d_i{}^2 = \sum(x_i - y_i)^2 = (\sum x_i{}^2) + (\sum y_i{}^2) - 2\sum x_i y_i$,

so that $\sum x_i y_i = \frac{(\sum x_i{}^2) + (\sum y_i{}^2) - \sum d_i{}^2}{2}$

$= \frac{1}{2}\left\{2.\frac{n}{6}(n+1)(2n+1) - \sum d_i{}^2\right\}$

Hence $S_{xy} = \frac{n}{6}(n+1)(2n+1) - \frac{1}{2}\sum d_i{}^2 - \frac{\left(\frac{n}{2}(n+1)\right)^2}{n}$

$$= \frac{n}{12}(n+1)\{4n+2-(3n+3)\} - \frac{1}{2}\sum {d_i}^2$$

$$= \frac{n}{12}(n+1)(n-1) - \frac{1}{2}\sum {d_i}^2$$

and $r = \dfrac{\frac{n}{12}(n+1)(n-1) - \frac{1}{2}\sum {d_i}^2}{\frac{n}{12}(n+1)(n-1)} = 1 - \dfrac{6\sum {d_i}^2}{n(n^2-1)}$

(3) Find the PMCC for the data below, using formulae.

x_i	1	2	3	4	5
y_i	1	4	9	16	25

Comment on the suitability of the PMCC in this case.

(3) Find the PMCC for the data below, using formulae.

x_i	1	2	3	4	5
y_i	1	4	9	16	25

Comment on the suitability of the PMCC in this case.

Solution

						$\sum$
x_i	1	2	3	4	5	15
y_i	1	4	9	16	25	55
${x_i}^2$	1	4	9	16	25	55
${y_i}^2$	1	16	81	256	625	979
$x_i y_i$	1	8	27	64	125	225

$$S_{xx} = \sum {x_i}^2 - n\bar{x}^2 = 55 - 5 \times \left(\frac{15}{5}\right)^2 = 10$$

$$S_{yy} = \sum {y_i}^2 - n\bar{y}^2 = 979 - 5 \times \left(\frac{55}{5}\right)^2 = 374$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 225 - 5 \times \left(\frac{15}{5}\right)\left(\frac{55}{5}\right) = 60$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{60}{\sqrt{10 \times 374}} = 0.981$$

Suitability?

- values taken by x are non-random, so PMCC not suitable (also, the y values are obviously artificial, and clearly non-random)

- plotted values form a curve; linear relation probably not appropriate

- plotted values don't give elliptical pattern