Part 1
Lecture 3a Association

1

Who I am…

Pascal Tyrrell, PhD

Associate Professor

Department of Medical Imaging, Faculty of Medicine

Institute of Medical Science, Faculty of Medicine

Department of Statistical Sciences, Faculty of Arts and Science

LINEAR CORRELATION COEFFICIENT

KARL PEARSON

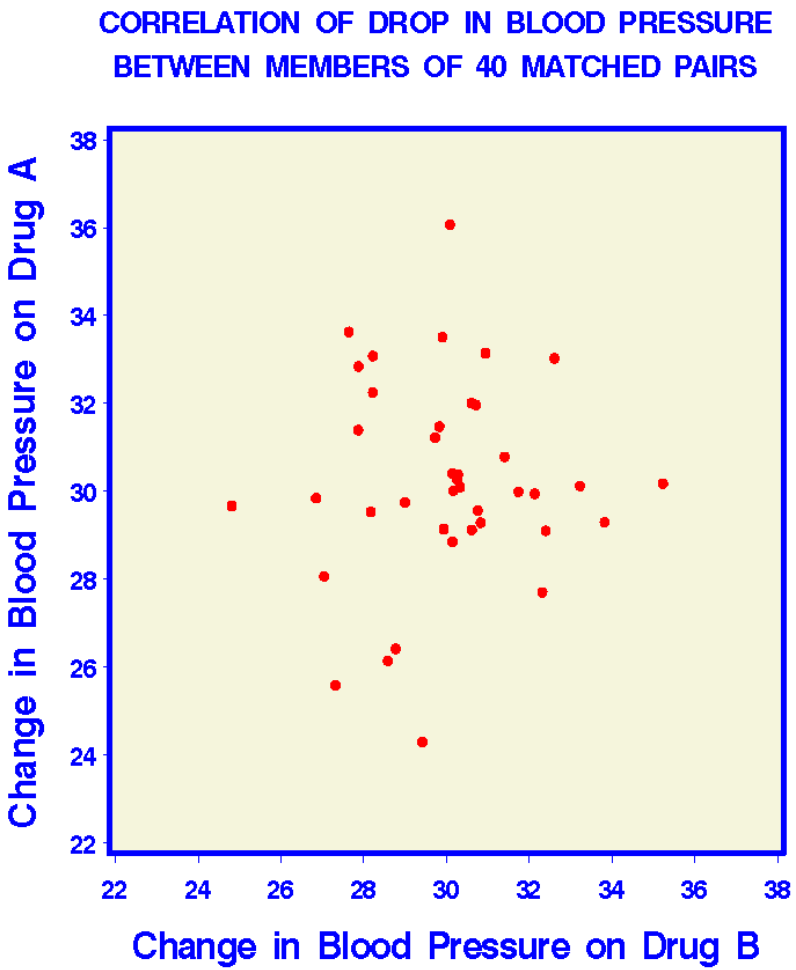
1857 - 1936

A matched pairs design was used in two studies to compare the decrease in blood pressure over six months on drugs A and B.

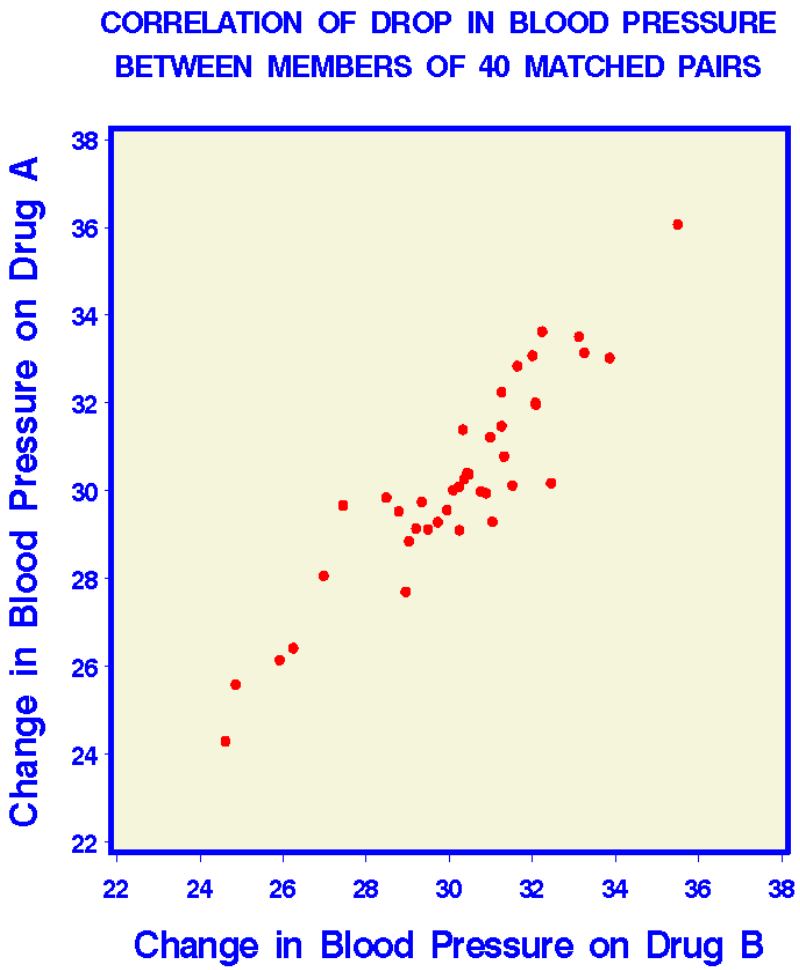
An *association* between blood pressure changes in members of each pair was only apparent in Study 2.

PLOTS OF CHANGE IN BLOOD PRESSURE FOR DRUGS A AND B

STUDY 1

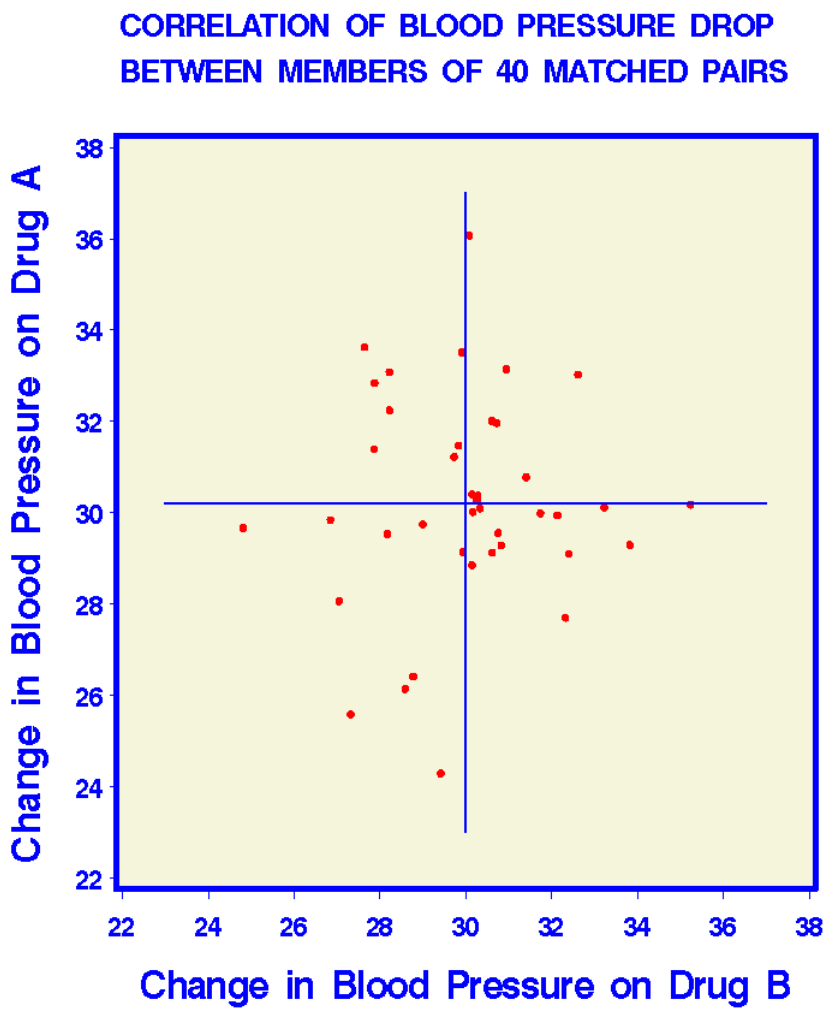


STUDY 2

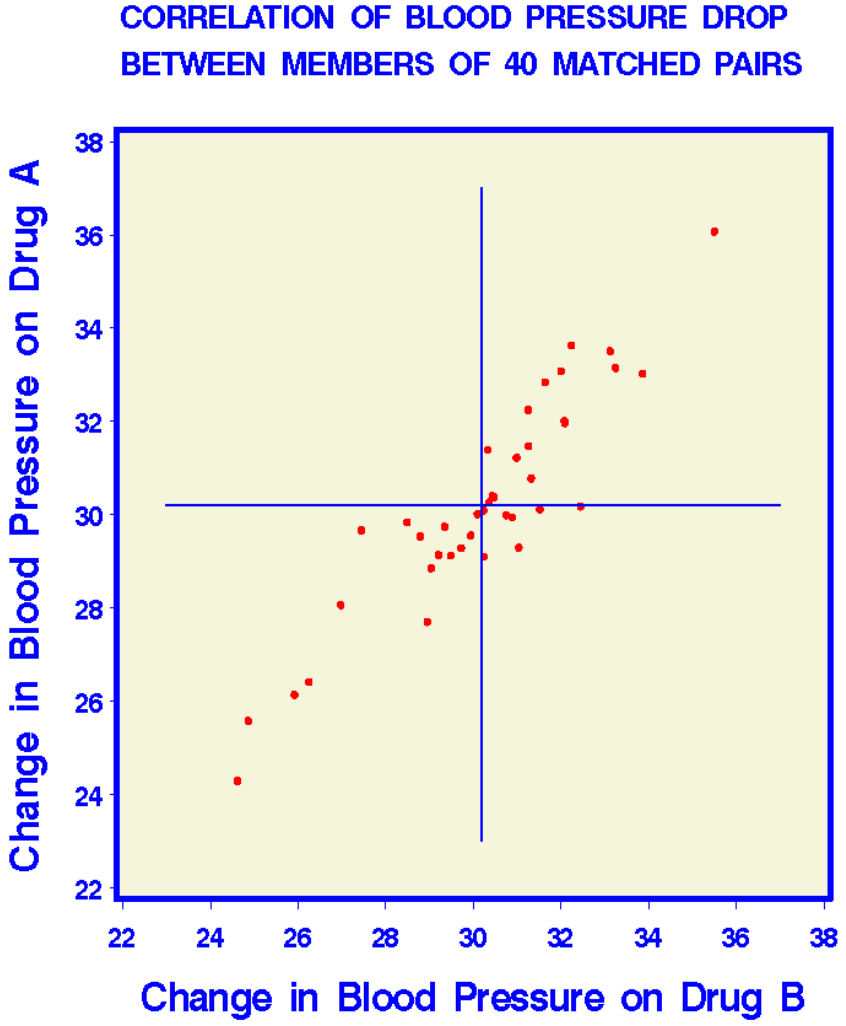


Horizontal and vertical lines are drawn through the plot means.

STUDY 1



STUDY 2



How might you quantify this association? Vertical and horizontal lines were drawn through the group means.

The number of points in each of the four quadrants for the left and right plots are

DRUG A	**Top**	**Left**	**=**	**8**	**Top**	**Right**	**= 9**
	Bottom	**Left**	**=**	**10**	**Bottom**	**Right**	**= 13**
DRUG B	**Top**	**Left**	**=**	**0**	**Top**	**Right**	**= 17**
	Bottom	**Left**	**=**	**16**	**Bottom**	**Right**	**= 7**

Our first stab at coming up with a measure of association might be something as simple as the proportion of points in the top right and bottom left quadrants.

Denoting this measure by R1 we have

$$\text{Study 1} \quad R1 = 19 / 40 = 0.475$$

$$\text{Study 2} \quad R1 = 33 / 40 = 0.825$$

A weakness of this definition is that it can only be positive because we have not taken into account the number of points in the top left and bottom right quadrants.

A large proportion of points in these quadrants would indicate a negative association. Therefore we will subtract the number of points in the top left and bottom right quadrants from the total in the top right and bottom left quadrants and divide the difference by the sample size.

Study 1 $R2 = (19 - 21) / 40 = -0.05$

Study 2 $R2 = (33 - 7) / 40 = 0.65$

If all points were in the left bottom and right top quadrants then R2 = 1

and if all points were in the left top and right bottom then R2 = -1

Let us create a formula that captures what was done in the calculations of this measure.

Replace X by a new variable SX , the sign of X, that is equal to +1 if X is to the right of the vertical axis and -1 if it is to the left of the vertical axis. Similarly Y, is replaced by a new variable SY, the sign of Y, that is equal to +1 for values above the horizontal and by -1 for values below the horizontal axis. Our general formula for R2 would be

$$R_2 = \frac{\sum_{j=1}^{j=n} SX_j \times SY_j}{n}$$

One defect of this measure is that it gives a value of +1 or –1 to SX and SY depending on which quadrant they are in but neglects any information in the size of X and Y which, in our case, is the change in blood pressure in groups A and B

A possible improvement would be to replace SX and SY by the difference from their means

Our third suggested measure of association between the X and Y variables is R3.

$$R_3 = \frac{\sum_{j=1}^{j=n} (X_j - \bar{X}) \times (Y_j - \bar{Y})}{n}$$

A defect of this measure is that its value would increase if we changed our units of measurement say from inches to cm

This would be an undesirable property for a measure of association. We must standardize these two deviations from their sample means: each difference is divided by their standard deviations

(in this case the sum of squared deviations are divided by N instead of N-1)

$$R_4 = \frac{\sum_{j=1}^{j=n} \frac{(X_j - \overline{X})}{S_X} \times \frac{(Y_j - \overline{Y})}{S_Y}}{n}$$

$$= \frac{\sum_{j=1}^{j=n} (X_j - \overline{X}) \times (Y_j - \overline{Y})}{n \times S_X \times S_Y}$$