Research Minute

Statistics 203 — Two Variable Statistics. Correlations.

Issue 18

Sandra Burge, Ph.D.

Most research examines the relationships between two Phenomena. Pearson ***r*** correlation analysis is used when both Phenomena are measured on a continuous numeric scale—like age, A1c, diastolic blood pressure, LDL, BMI, test scores, household size, pill counts, mini-mental status scores, alcohol consumption. Correlations can analyze questions like:

- Are hours of screen time per day (Phenom 1) associated with children's BMI (Phenom 2) ?
- Is BMI (Phenom 1) associated with A1c (Phenom 2) in patients with diabetes?
- Are years of education (Phenom 1) associated with scores on a Mini-Mental Status Exam (Phenom 2) in older adults?

Pearson's *r* Correlations

The *Pearson product-moment correlation coefficient*, denoted by ***r***, measures the strength of a linear relationship between two Phenomena. This relationship can be graphed on a scatterplot. Below, imagine that each dot represents a person's score on two Phenomena, X and Y.

Figure 1: scores on Phenom X appear to bear no relationship at all to scores on Phenom Y; it's simply a cloud of dots.

Figure 2: the cloud of dots leans down to the right– a negative correlation. A person who gets a high score on Phenom X is likely to get a low score on Phenom Y, and vice versa.

Figure 3: the cloud of dots slopes up to the right- a positive correlation. Individuals who score high on Phenom X also score high on Phenom Y, and those who score low on X also score low on Y.

The correlation analysis attempts to draw a line of best fit through these dots and assesses the distance of the dots from this line.

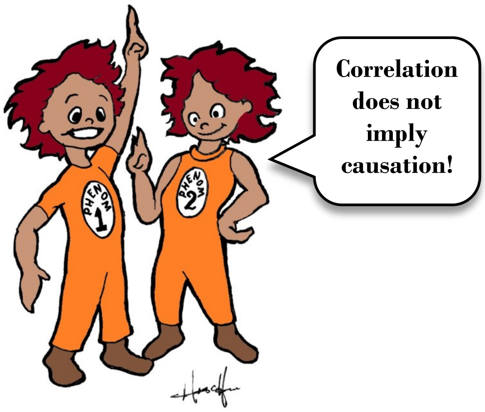
r can range from +1 (a perfect positive correlation) to –1 (a perfect negative correlation). A zero correlation indicates no relationship whatsoever between Phenom X and Phenom Y.

Figure 1 below shows a near-zero correlation, Figure 3 shows a strong positive correlation, and Figure 2 shows a weaker negative correlation. Note the value of ***r*** in each.

How is ***r*** calculated?

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

The distance between the mean of X-scores and each individual's X-score is calculated and divided by the standard deviation of X's; likewise for Y-scores. These are multiplied together (called a cross-product), summed, and divided by the sample size minus 1.



Even when r is statistically significant, it may be clinically insignificant. The square of r (called ***R²***) is the percent of variance in Phenom Y explained by Phenom X. If ***r*** = .4, ***R²*** = .16. So, if I know the value of X, about 16% of the time, I can predict the value of Y. Significant, but not that impressive.

When you interpret correlations, remember: "Correlation does not imply causation." You may find strong associations, but it does not "prove" that Phenom X caused Phenom Y.

You can find an online Pearson correlation calculator here: http://www.socscistatistics.com/tests/pearson/Default2.aspx

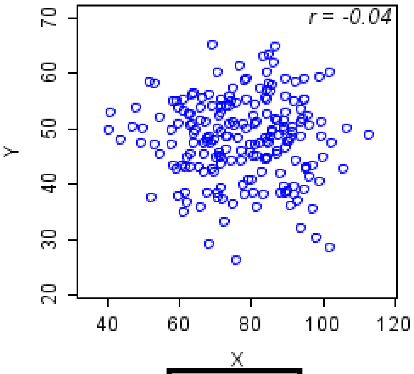


Figure 1

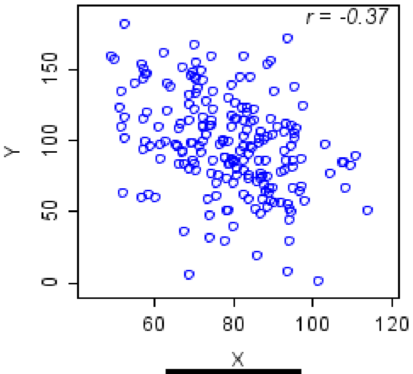


Figure 2

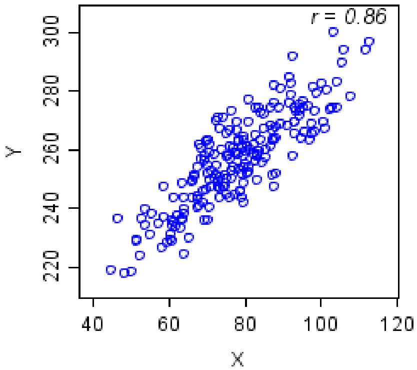


Figure 3

burge.researchminutes@gmail.com

Art by Brian Hischier