# Statistical Thinking for the 21st Century

# Chapter 13

# Modeling continuous relationships

Most people are familiar with the concept of *correlation*, and in this section of the chapter we will provide a more formal understanding for this commonly used and misunderstood concept.

## 13.3 Covariance and correlation

One way to quantify the relationship between two continuous variables is by calculating their *covariance*.

Remember that variance for a single variable is computed as the average squared difference between each data point and the mean:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N - 1}$$

The variance tells us how far each observation deviates from the mean, on average, in squared units.

The covariance tells us whether there is a relation between the deviations of two different variables across observations. Covariance is defined as the following:

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

The covariance will be far from zero when a variable x and another variable y are both highly deviant from the mean.

If both variables, x and y, are deviant in the same direction, then the covariance is positive. In contrast, if both variables, x and y, are deviant in opposite directions the covariance is negative.

The covariance is simply the mean of the crossproducts. We don't usually use the covariance to describe relationships between variables, because it varies with the overall level of variance in the data.

Instead, we would usually use the *correlation coefficient* (often referred to as *Pearson's correlation* after the statistician Karl Pearson). The correlation is computed by scaling the covariance by the standard deviations of the two variables:

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

The correlation coefficient is useful because it varies between -1.000 and 1.000 regardless of the nature of the data. For that reason, as we already discussed, the correlation coefficient can be considered an effect sizes. A correlation of 1.000 indicates a perfect linear relationship, a correlation of -1.000 indicates a perfect negative relationship, and a correlation of 0.000 indicates no linear relationship.

### 13.3.1 Hypothesis testing for correlations

Imagine we calculated a correlation between two continous variables and observed a correlation coefficient of 0.423. That correlation coefficiant seems to indicate a reasonably strong relationship between our two variables, but we can also imagine that this could occur by chance even if there is no relationship.

Therefore, we can test the null hypothesis that the correlation is zero, using a simple equation that lets us convert a correlation value into a *t* statistic:

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

If this test shows that the likelihood of an r value this extreme or more is quite low, we can reject the null hypothesis that $r = 0.000$.