1. The Pearson Correlation Coefficient

Correlation

You've likely heard before about how two variables may be *correlated*. While we use this word in an informal sense, there is actually a very specific meaning of the term in statistics. Correlation means that, given two variables X and Y measured for each case in a sample, variation in X corresponds (or does not correspond) to variation in Y, and vice versa. In other words, extreme values of X are associated with extreme values on Y, and less extreme X values with less extreme Y values. The degree of this correspondence is called statistical correlation.

Correlation and causation

If one variable causally influences a second variable, then we would expect a strong correlation between them. However, a strong correlation could also mean, for example, that they are both causally influenced by a third variable. Therefore a strong observed correlation can suggest a causal connection, but it doesn't per se indicate the direction or nature of that causation.

$X \rightarrow Y$	$X \leftarrow Y$	$X \leftrightarrow Y$	$X \leftarrow A \rightarrow Y$
X influences Y	Y influences X	X and Y influence each other	A influences X and Y

Alternative Explanations for Strong Observed Correlation

Important: Correlation between two variables does not prove X causes Y or Y causes X. Example: There is a statistical correlation between the temperature of sidewalks in New York City and the number of infants born there on any given day.

Pearson r

There is a simple and straightforward way to measure correlation between two variables. It is called the Pearson correlation coefficient (***r***) – named after Karl Pearson who invented it. It's longer name, the Pearson product-moment correlation, is sometimes used.

The formula for computing the Pearson *r* is as follows:

$$r = \frac{1}{n-1} \sum_i \frac{(x_i - \overline{X})(y_i - \overline{Y})}{s_x s_y}$$

The value of *r* ranges between +1 and -1:

- r > 0 indicates a positive relationship of *X* and *Y*: as one gets larger, the other gets larger

- r < 0 indicates a negative relationship: as one gets larger, the other gets smaller
- r = 0 indicates no relationship

Let's try to intuitively understand how this formula works. We start by subtracting the means from *X* and *Y*, and then multiplying the results. When we subtract the mean from a variable, some of the resulting values will be positive and some negative. When we subtract the mean from both *X* and *Y*, that will happen with both variables.

If there is no association between *X* and *Y*, there will be no systematic relationship between $(x_i - \overline{X})$ and $(y_i - \overline{Y})$. Therefore the positive values of one will match up with positive and negative values of the other randomly, and the same with negative values of the first variable. Therefore when we take the sum of $(x_i - \overline{X})(y_i - \overline{Y})$, all these positive and negative results will tend to cancel each other out, making r close to 0.

However if two variables are strongly associated, then positive values of $(x_i - \overline{X})$ will match up with positive values $(y_i - \overline{Y})$, and negative values with negative values. The sum of $(x_i - \overline{X})(y_i - \overline{Y})$ will produce a positive r.

In a reverse relationship, positive values of $(x_i - \overline{X})$ will match up with negative values of $(y_i - \overline{Y})$, and vice versa. Then the sum of $(x_i - \overline{X})(y_i - \overline{Y})$, and r, will be negative.

If we calculate the Pearson correlation of X with itself, the result will be 1:

$$r = \frac{1}{n-1}\sum_i \frac{(x_i - \overline{X})^2}{s_x s_y} = \frac{s_x^2}{s_x^2} = 1.$$

Computational shortcut

We may rewrite our original formula as:

$$r = \frac{1}{n-1}\sum_i \left[\frac{(x_i - \overline{X})}{s_x} \times \frac{(y_i - \overline{Y})}{s_y} \right]$$

Recalling the formula for a z score:

$$z = \frac{(x - \overline{X})}{s}$$

we get:

$$r = \frac{1}{n-1}\sum_i z_x z_y$$

Therefore all we need to do is to convert our original *X* and *Y* values into *z*-scores, then multiply these for each case, sum, and divide by *n* – 1.

Spreadsheet calculation

Pearson correlation calculator

X	Y	X-Xbar	Y-Ybar	z_x	z_y	(z_x)(z_Y)
1	1	-4.5	-4.5	-1.4863	-1.4863	2.2091
2	2	-3.5	-3.5	-1.1560	-1.1560	1.3364
3	3	-2.5	-2.5	-0.8257	-0.8257	0.6818
4	4	-1.5	-1.5	-0.4954	-0.4954	0.2455
5	5	-0.5	-0.5	-0.1651	-0.1651	0.0273
6	6	0.5	0.5	0.1651	0.1651	0.0273
7	7	1.5	1.5	0.4954	0.4954	0.2455
8	8	2.5	2.5	0.8257	0.8257	0.6818
9	9	3.5	3.5	1.1560	1.1560	1.3364
10	10	4.5	4.5	1.4863	1.4863	2.2091

Xbar	5.5
Ybar	5.5
N	10
N-1	9
sd_s X	3.0277
sd_s Y	3.0277
r	1.00

We'll construct the above spreadsheet calculator in class.

Video

Khan Academy – Correlation and causation
http://www.youtube.com/watch?v=ROpbdO-gRUo