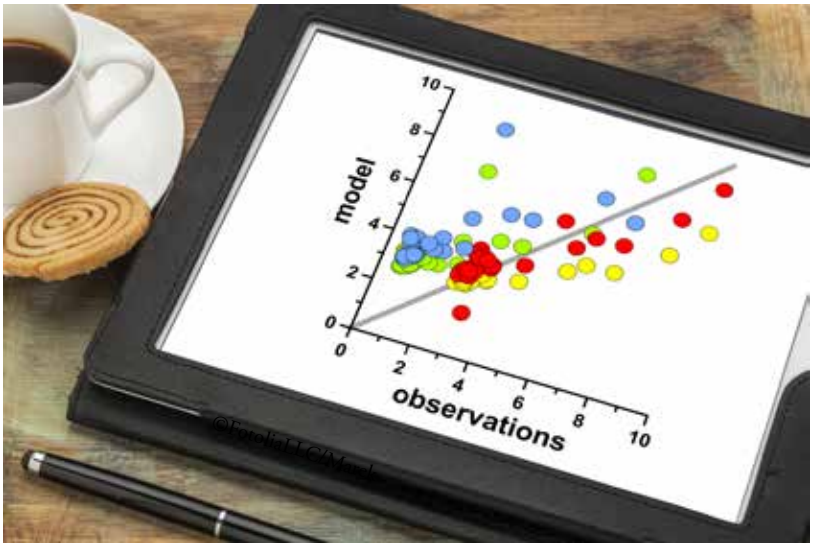
Five Minute Science Lesson: Correlation and Covariance (What is it, and How to Roll-your-own)



by Raymond Nelson

Science is all about trying to understand the universe and reality. What is it? How does it works? What is it made of? How big is it? What is going to happen? It's a big universe and so we do not try to understand everything at once. Instead we take things one little piece at a time. If our knowledge about some small aspect of the universe and reality is correct then we can expect our knowledge to fit together nicely with other pieces of knowledge of other pieces of the universe. Understanding the universe, or anything in it, requires that we organize our knowledge such that our conclusions can be re-analyzed and reproduced at a latter time without having to start at the beginning and learn everything anew. In this way others might begin to develop additional knowledge about other aspects of the universe and reality.

Understanding the universe, or anything in it, requires that we not only systematically differentiate one thing from another but that we also attempt to learn about the relationships between things. Relationships between physical things, much like relationships between people, are abstract and amorphous things that cannot themselves be subject to physical measurement. And yet amorphous

things exist; they are describable and real. Relationships between things can be experienced, observed and even quantified. And so, a goal of science is to understand, describe and quantify things in the real universe and the relationship between those things.

Covariance refers to similarities in the changes in something and something else. The notion of covariance implies that the phenomena under observation are changing and not static. Covariance means that as one thing changes the other thing tends to change in a similar way. Covariance in statistics refers to a unitized description of how much the two things change together.

In general usage *correlation* is a word that describes a mutual relationship or connection between two things. In scientific and statistical usage, correlation also refers to the statistical strength of the relationship. The difference between relationship and correlation is that a relationship is an abstraction that can be observed, experienced or described qualitatively, whereas a correlation implies that we have attempted to quantify a relationship between two things. Correlation implies that the relationship between two things is imperfect. A perfect correlation between any two things would be easy to understand because every aspect of one thing would be mirrored in the other. A perfect correlation would signify redundancy, wherein understand one thing to any degree of satisfaction will provide equivalent knowledge and information about the other. Perfection, however, is rare. Most things, including most relationships and most correlations between things are imperfect. Imperfect relationships and imperfect correlations take more effort to understand and quantify.

To our good fortune, others before us have devoted considerable energy and effort to define the concepts and procedures necessary to begin to describe and quantify the correlation between observable phenomena in the form of a *correlation coefficient* (Katz, 2006; Pearson, 1895), also known as the *Pearson correlation*. The Following is a description of the intuition and procedures for the calculation of a Pearson Correlation coefficient that is often designed with the letter *r*.

Intuition: how to think about correlation

Intuition refers to our ability to think about and understand the concept. The intuition for a correlation coefficient is this: a perfect correlation is signified by the value 1.0, whereas no correlation is signified by the value 0 (similar to 0%). These are similar to the values 100% and 0%, though it would

be incorrect to discuss or describe correlation coefficients as percentages. No correlation (no relationship) can also be thought of as a completely random relationship in which any change on one thing can be accompanied by absolutely any of all possible changes in the other thing. Also, correlations can be either positive or negative (inverse correlation). Inverse or negative correlation means that a change in a particular direction for one thing is generally accompanied by a change in the opposite direction for the other thing. A correlation coefficient is therefore a number on a continuum from -1. 0 to +1.0 (See Figure 1).

Figure 1. Correlation coefficient
-1.0 ---------- 0 ---------- +1.0

How to roll-your-own correlation coefficient

Calculation of a correlation coefficient requires that we first quantify the observed phenomena. And so we begin with any set of data. For this example we will calculate the correlation between two fictitious sets of data, A and B. Calculation of a correlation coefficient requires that the two group have an equal number of items and assumes that the data are normally distributed.

Data A is the set of number 1, 2, 3, 4 and 5. Data B is the set of numbers 6, 7 8, 12, and 10. Notice that there is some describable, though imperfect, relationship between these two sets of numbers. Data A is a simple sequence of five numbers beginning at one and for which each item increases by one. Data B appear to be a similar sequence beginning at six and for which the first three items increase by one while the fourth item is out of sequence before the last item is the expected value 10.

Data A
1
2
3
4
5

Data B
6
7
8
12
10

Step 1: calculate the mean of each data group.

This is done easily by summing the items for each group and then diving those sums by the number (n) of each group.

Mean A = 3
Mean B = 8.6

Step 2: calculate the standard deviation of each data group.

Virtually every computer today can run software such as Excell (proprietary) or LibreOffice (free/open source) that an easily calculate standard deviations using included functions. Or you can roll-your-own standard deviations using the procedures described by Nelson (2017; *Five Minute Science Lesson: An Algorithm to "Roll-Your-Own" Standard Deviations)*. for a description of how to roll-your-own standard deviation.

Standard Deviation A = 1.581
Standard Deviation B = 2.408

Step 3: calculate the deviation for each value and the corresponding group mean

Deviation refers to the difference between each item and the mean. The mathematical deviations give us information about how the data vary around the group means.

Group A Deviations
1 – 3 = -2
2 – 3 = -1
3 – 3 = 0
4 – 3 = 1
5 – 3 = 2

Groups B Deviations
6 – 8.6 = -2.6
7 – 8.6 = -1.6
8 – 8.6 = -0.6
12 – 8.6 = 3.4
10 – 8.6 = 1.4

Step 4: Calculate the products of the paired deviations for the two groups

Multiply each pair of deviations for the two groups. Multiplying the paired deviations gives us a way to quantify the deviation of the two groups.

-2 * -2.6 = 5.2
-1 * -1.6 = 1.6
0 * -0.6 = 0
1 * 3.4 = 3.4
2 * 1.4 = 2.8

Step 5: Sum the products of the deviations

Summing the products will give total amount of deviation around the two group means.

5.2 + 1.6 + 0 + 3.4 + 2.8 = 13

Step 6: Divide the summed product of paired deviations (the result from step 5) by N-1

N is the number of pairs. We divide by N-1 because we cannot observe the entire population and are limited to calculation with a sample. When we have access to all population data we can divide by N.

13 / (5 – 1) =
13 / 4 = 3.25

Step 7: Divide the result from step 6 by the product of the standard deviations for the 2 groups

The product of the standard deviations is obtained by multiplying the results from step 2.

3.25 / (1.581 * 2.408) =
3.25 / 3.807 = .854

Result is the correlation coefficient

r = .854

In this example fictitious data there is an obvious correlation between the changes in the items in the two groups. As the items in Group A increase we can generally, though not always, expect to observe the items in Group B to increase.

Covariance coefficient (what is?)

Covariance is related to correlation. The covariance coefficient is the un-normalized measurement of joint variability between two groups of data. The normalized measure is the correlation coefficient. In the example above we normalize the result in Step 7 by dividing the result of Step 6 by the product of the standard deviations for the two groups. In other words we see the covariance coefficient at Step 6. The covariance coefficient tells us essentially the same conceptual information as the correlation coefficient but does so using units of measure that are the same unit of measure in which the data are expressed (whereas the correlation is constrained or normalized to the range -1.0 to +1.0).

Interpreting a correlation coefficient

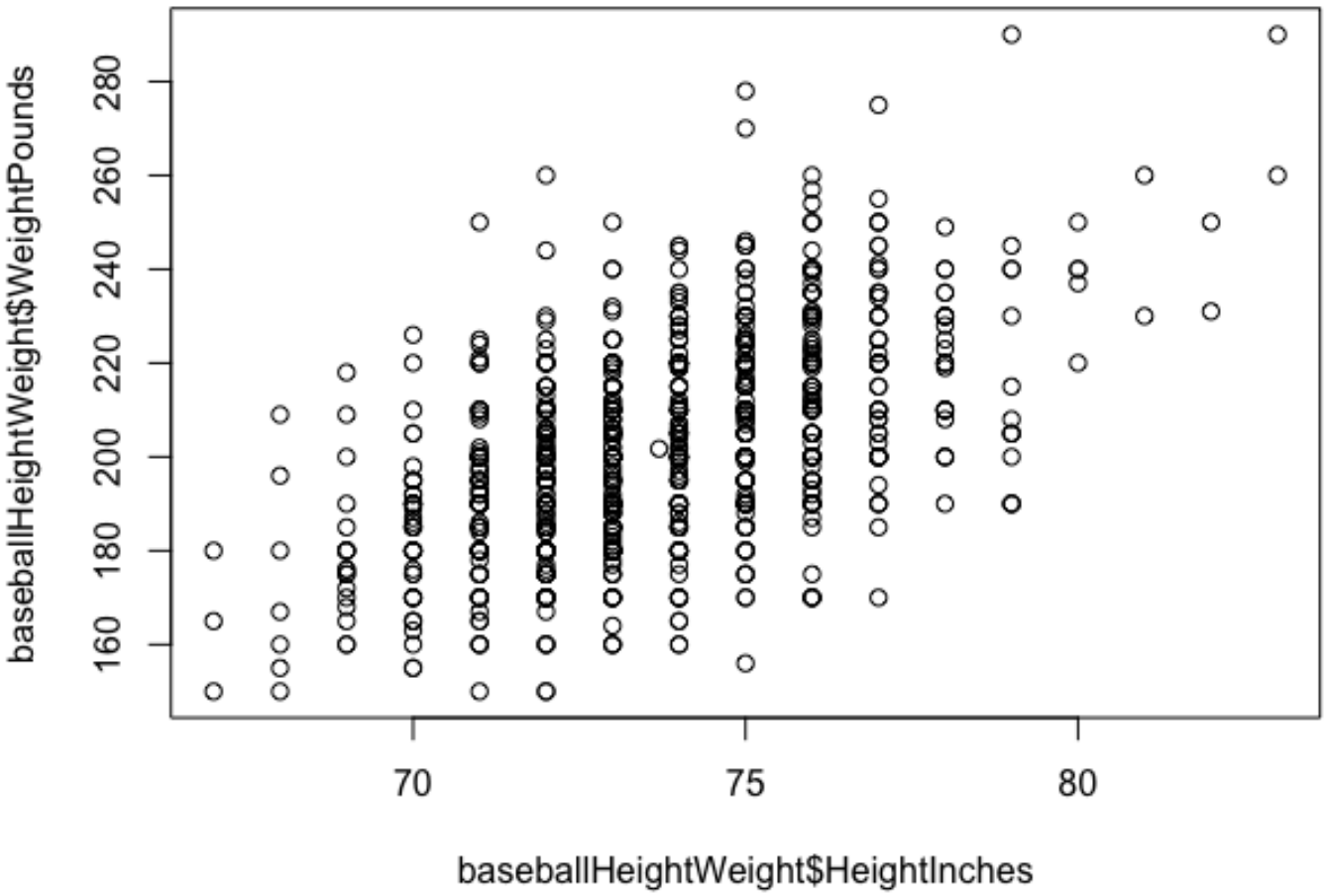
Statistical correlations are often interpreted described qualitatively using words such as *very weak*, *weak, moderate, strong,* or *very strong*. A desire for concrete interpretation guidelines has prompted some to suggest numerical thresholds for different qualitative adjectives, though these are increasingly viewed as arbitrary, unwise and problematic because they distract from an adequately nuanced understanding

of the information. To simplify our understanding and interpretation of meaning of a correlation statistic it is generally sufficient to remember that correlation coefficients close to 1.0 are approaching a linear or perfect relationship. Similarly, a relationship that approaches the value -1.0 is approaching linearity or perfection though as one thing changes the other will be observed to change in the opposite direction. Values close to 0 signify no relationship – also thought of as a random relationship. It is also helpful to develop our understanding and intuition for correlation coefficients by thinking about examples.

An example using height and weight with major league baseball players

A sample of baseball statistics can be obtained from (http://wiki.stat.ucla.edu/socr/). Intuitively we understand that there is a coherent relationship between player height and weight. Players who are taller will tend to weigh more than those who are less tall. Using a sample of N=1035 players we know that their heights rage from 67 inches to 83 inches and their weights range from 150 to 290 pounds. We can calculate the correlation between height and weight as r = .542. This is generally regarded as a moderate to strong correlation. We know that height and weight are determined largely by genetics, though nutrition and lifestyle (and growth hormones) may also play a role. We will forgo the details of the calculations and encourage readers to obtain the data and work it out for themselves. Figure 2 shows the scatterplot of players' height and weight. The correlation between player height can be observed in the apparent linearity of the data points; as player height increases player weight also generally increases.

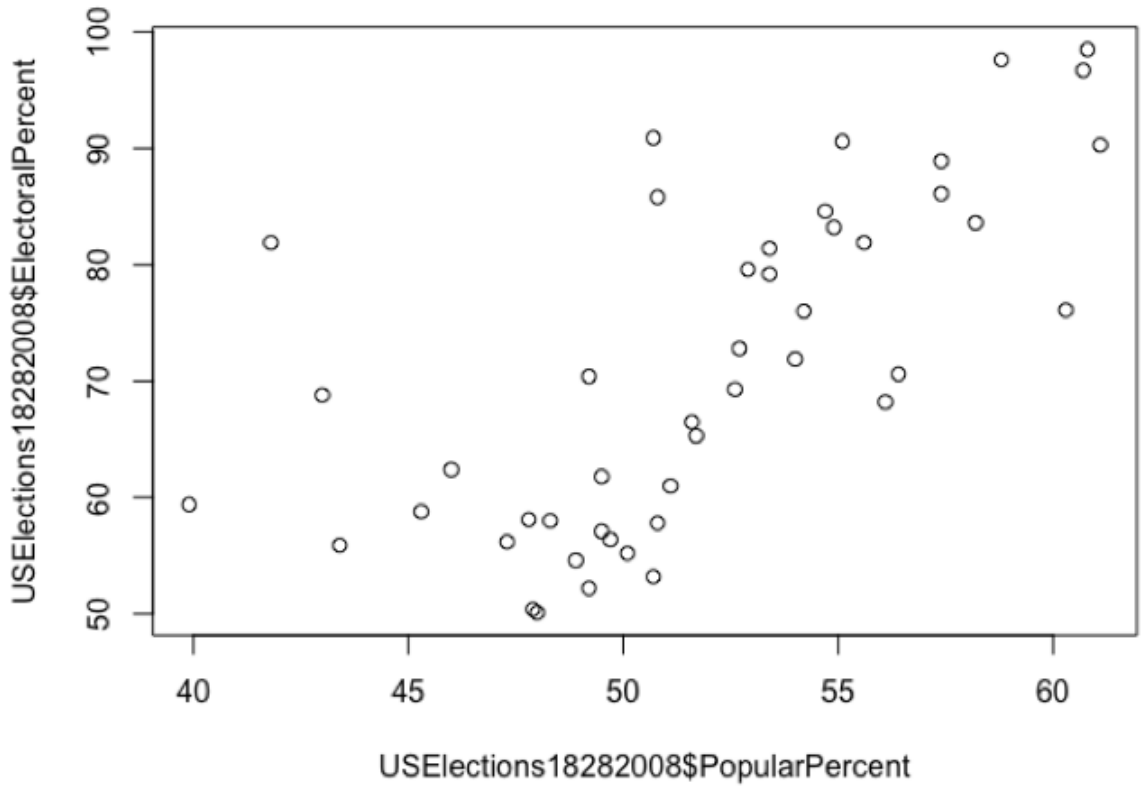
Figure 2. Scatterplot of height and weight for N=1035 major league baseball players.



Correlation of popular vote and electoral college votes for US elections from 1826 to 2008

The population of elected U.S. Presidents from 1826 to 2008 can also be obtained from (http://wiki.stat.ucla.edu/socr/). Calculation of the relationship between popular votes totals and electoral college totals revealed a correlation coefficient of r = .681. This relationship is approaching what is generally regarded as a strong though imperfect relationship. Again, readers are encouraged to obtain the data and practice the calculation of the correlation coefficient. Figure 3 shows the scatterplot of popular vote percentages and electoral college percentages, with some obvious linearity despite the presence of some noisy data points.

Figure 3. Scatterplot of popular and electoral vote percentages for US Presidents 1828 to 2008.



Caution

Correlation is not itself an effect size

The correlation coefficient does not tell us the size of the effect in terms of a linear proportion or rate of agreement between two things. For that we will need to calculate the *coefficient of determination* which is simply r^2 (r-squared). The coefficient of determination tells us how much of a change in each can be attributed to a change in the other thing.

Correlation is not causation

Correlation is not causation is a phrase that might easily be repeated as a mantra. It is easily overlooked despite its importance. Simply identifying a relationship, and even the direction of the association, between two phenomena does not automatically tell us about the underlying causal relationship. In a now infamous and entertaining example, Messerli (2012) published a study in the New England Journal of Medicine showing a strong correlation ($r = .791$) between chocolate con-

sumption per capita and the number of Nobel laureates per 10 million persons in 23 countries. A causal inference from the reported data indicated that an increase of 0.4kg of chocolate per capita per year would increase the number of Nobel prizes by one. The humor and entertainment value of that publication was as useful as the point: correlation is not causation. Conclusions about causation require controlled studies. In the case of the chocolate/Nobel study it is likely that some intervening variables such as per capita income, time and resources for sustained study activity, and education in general may play some more plausible causal role.

Why bother to learn these calculations

Once upon a time computers were not widely available as they are today. In that epoch it was necessary for both scientists and field practitioners – anyone who wished to claim expertise or achieve any productivity – to learn to execute mathematical and statistical calculations using only pencils and paper (and perhaps a slide-rule). Today computers are everywhere are calculations are rarely completed without the use of a computer. Why then should scientists and field practitioners today have to be exposed to any suggestion that they learn to calculate statistics such as the correlation coefficient?

My answer: there is no expectation that scientists or field practitioners will ever again manually calculate any statistics in field practice. However, those who take the time to familiarize themselves with basic statisticians and basic calculations, and those who take the time to work through a few simple examples, will possess much stronger intuition and much greater expert understanding of the meaning of probabilistic results. And many, if not most, professional and scientific conclusions are, in reality, probabilistic conclusions – including when those conclusions are reduced or simplified to categorical conclusions.

Professionals who do not take the time to develop their knowledge and intuition about science and basic statistics will present a limited and flawed form of expertise at best, and will be less able to help other professionals and members of the public or information media to correctly understand their work and their conclusions. Moreover, there is a risk that pseudo-expertise will be accompanied by an obvious undertone of insecurity and phobic avoidance of concepts having to do with science, statistics or probabilities. In the absence of real expertise there is the very real danger that the void will

be filled with pretense and over-statement around the strength of one's conclusion, and this can lead only to professional embarrassment when faced with the need for real discussion with other educated professionals. Moreover, in an age where artificial-intelligence can begin to invite us to more completely outsource a wider variety of skills and judgments that formerly required trained and experienced human expertise, there is potentially grave risk in neglecting to develop mathematical, statistical and analytic skills and intuition because this would require that we complete surrender our expertise to robots and machines. Said differently, if human professionals want to continue to enjoy the role of expert then it is our responsibility to develop our expertise and competence. Professionals who take the time to develop their intuition for science and skills for statistics, as the mathematical language of science, are likely enjoy and important advantages in marketplaces that are expected to enjoy an increasing range options to replace non-expert practitioners with autonomous systems while continuing to require the involvement of human experts.

©FotoliaLLC/Makibestphoto

References

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58,* 240–242.

Katz, M. H. (2006) *Multivariable Analysis – A Practical Guide for Clinicians. 2nd Edition.* Cambridge University Press.

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine, 367(16)*, 1562–1564.