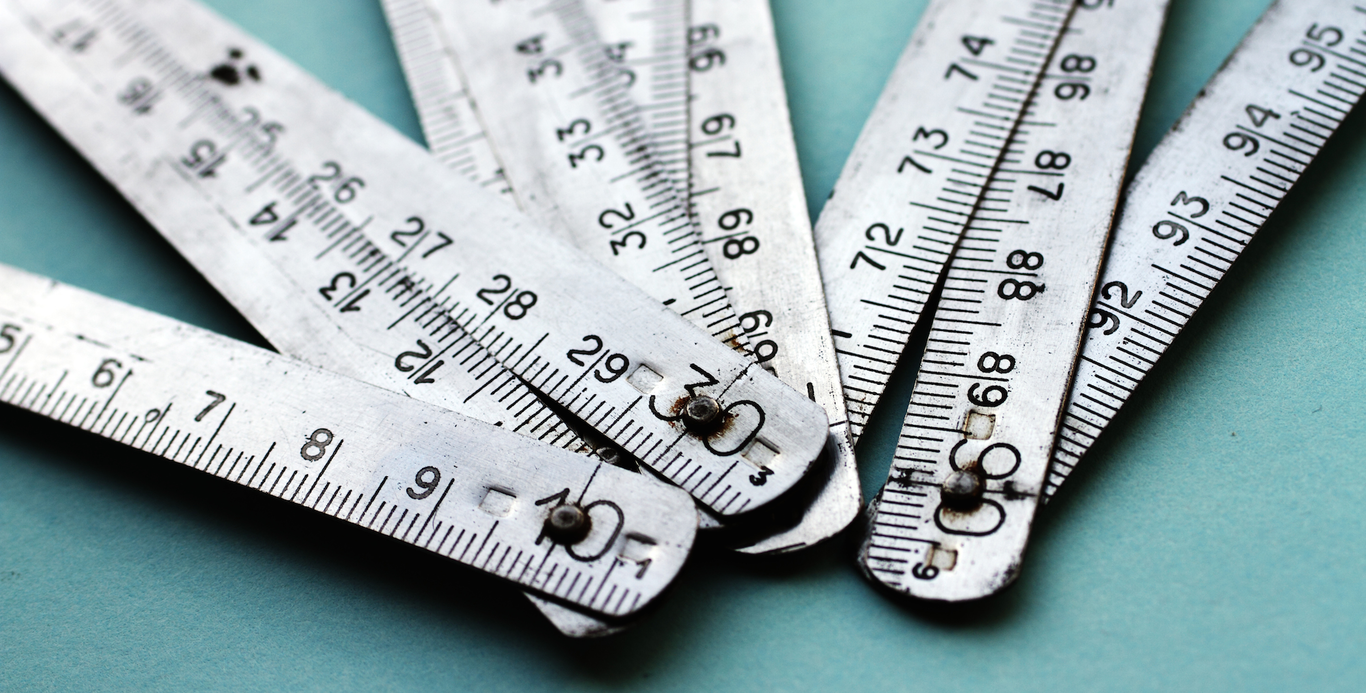
Linear Association and Correlation

Charles C. Igel

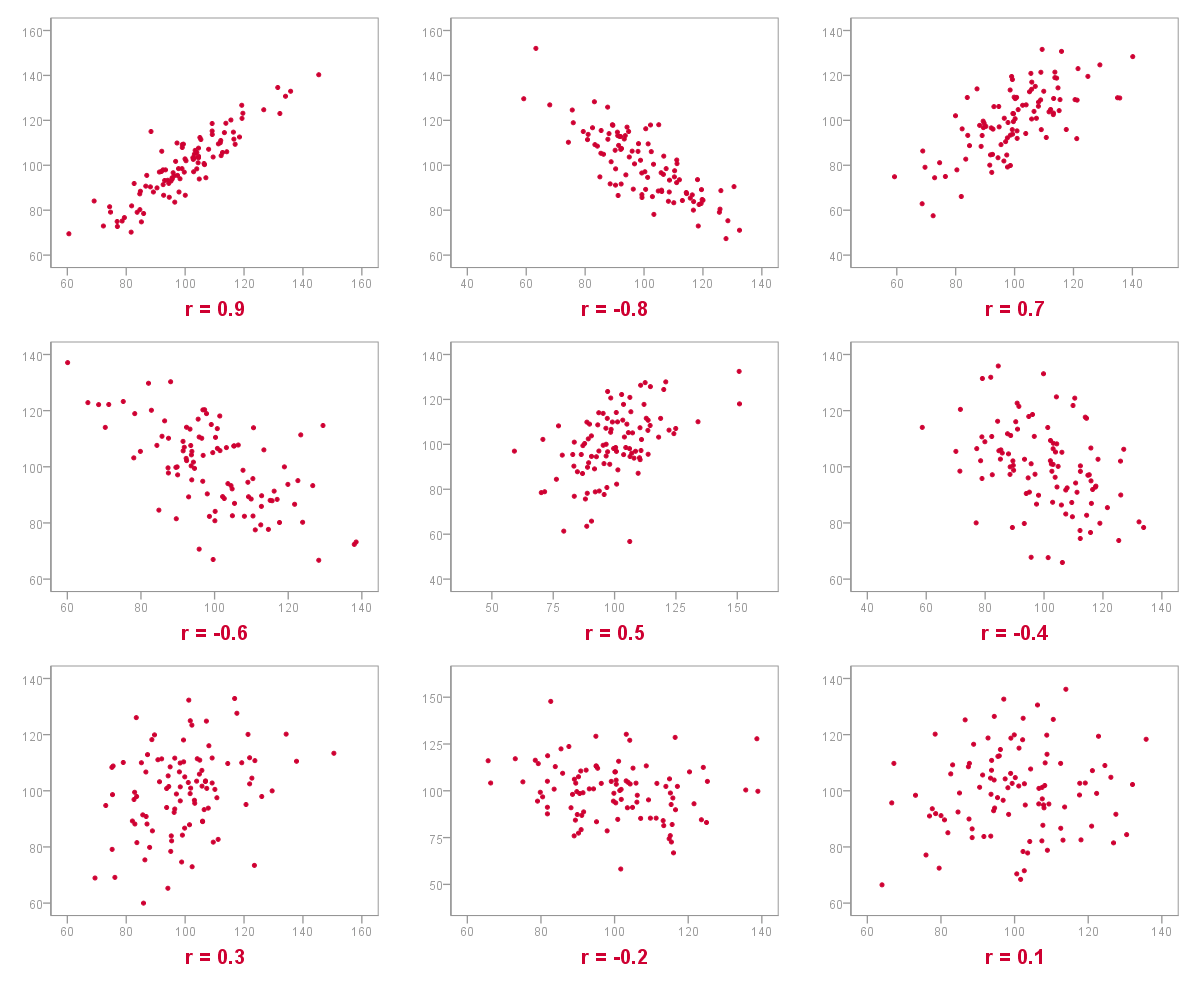


Correlation is a fundamental concept within statistics that, once understood, provides insight into more complex statistical models and ideas. From a conceptual standpoint, correlation summarizes the *measured association* between variables, meaning the extent to which one variable is affected by the other. Put another way, correlation is simply a measure of association.

The term *measured association* carries a lot of meaning here, so let’s unpack it. To calculate the correlation between variables, we first have to measure those variables. The term *measured* *association* rather than simply *association* is a hedge against the possibility that those measures could be inaccurate, and not truly reflect the thing we intend to measure. Science is cautious; and the terms we use reflect that caution. The word *association* refers to how the data points between variables trend relative to each other. If one goes up does the other go up as well? Or maybe it goes down? Or maybe the change in one does not systematically affect the the other. Of course, this means that we need multiple data points across variables to determine correlation; but more on that in a minute.

Association as a concept is a singular thing, but correlation as a measurement is multiple things. There are a variety of way to calculate correlation; and each is responsive to two important data characteristics. The first characteristic is the type of data being analyzed. All data is not created equal. It comes in levels of measurement that are categorized from least to most detailed as: nominal, ordinal, interval, ratio. Nominal is often something like a discrete category (e.g., Democrat, Republican, Libertarian, Independent) and ratio is a continuous measurement where zero represents an absence of the variable (e.g., height, age). Ordinal and interval are somewhere between. The second characteristic is the trend within the data. Data comes in different types of distributions. Imagine having a list of test scores, placing them in order from lowest to highest, plotting them on a graph, and fitting a line that summarizes the trend of the data. That line may be straight (i.e., linear) or curved (non-linear). Correlation is calculated differently based on this trend within the data begin analyzed.

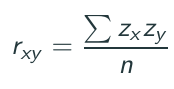
When the data we have is at the interval or ratio level of measurement, and we expect the trend of the data to be essentially linear, correlation is measured with a statistic known as Pearson’s *r*, or sometimes simply *r*. It is measured on a continuous decimal scale from -1 to 1 (Figure 1). The number we arrive upon, called a correlation coefficient, tells us the magnitude (i.e., strength) of the association between the measured variables and the direction (i.e., positive or negative) of that association.



Correlation Scatterplots

Source: [https://spss-tutorials.com](https://spss-tutorials.com/img/correlation-coefficient-multiple-scatterplots.png)

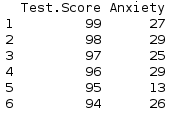
It’s worth restating that *r* is used to detect association between variables when the data is interval or ratio and we anticipate a linear association between the variables. Under these circumstances, correlation is calculated as,



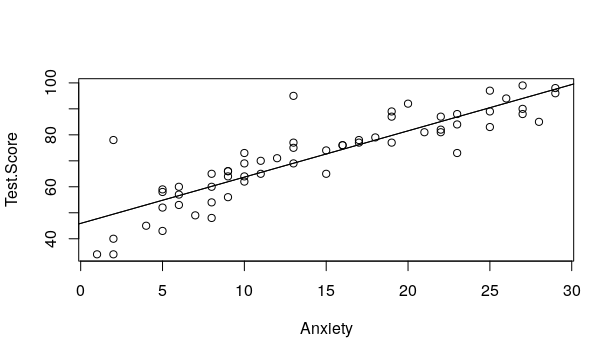
where *Zx*, *Zy* are the standard scores for a single case of the variables *x* and *y* respectively, and *n* is the sample size of the data set used for this calculation. If the concept of a standardized score (*z*-score) is unfamiliar, [this site](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/) explains it well. Conceptually, the use of a standardized score means that we do not have to know the scale of the measures used to interpret the correlation coefficient. You actually could calculate correlation by hand given a small data set or enough time and patience with a large data set, but hand calculations are no longer done aside from illustrative purposes. Thankfully we have statistical packages to do the heavy lifting for us. Still, it’s important to understanding what’s happening when we tell those packages to compile.

Let’s see what this looks like in practice. Assume that we want to know if there is an association between performance on math tests (*y*) and the test taker’s anxiety level (*x*). To do this we sample 50 students, measuring their anxiety level, then have them take a math test. Table 1 shows the a truncated list of the first six cases (i.e., study participants) showing each participants anxiety level matched to their math score.

*Table 1*. Matched Anxiety-Math Dataset

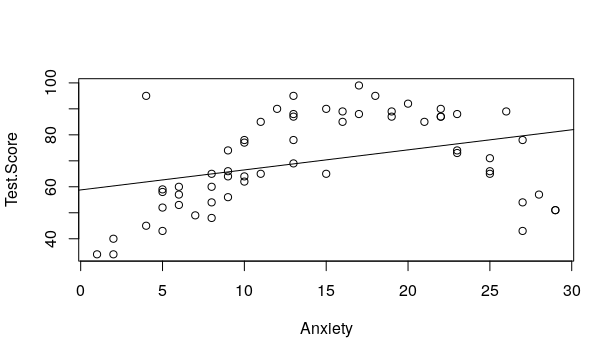


The calculated correlation for this sample is *r*(*xy*)=0.88, indicating a strong positive association between test scores (*y*) and anxiety (*x*). This is plotted in Fig. 2. That’s a great thing about correlation, you can actually see it when you plot the data. Notice that as anxiety increases along the [x-axis](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) of the plot, test scores generally increase along the [y-axis](https://en.wikipedia.org/wiki/Cartesian_coordinate_system). There are a few outliers, but these have not affected the correlation coefficient too much.



Correlation with Linear Data

Our assumption about linearity has held in this example, but what happens when the association between our variables is non-linear. Fig.3 shows familiar data that fits this situation.



Correlation with Non-Linear Data

In this example test scores (*y*) rise alongside anxiety (*x*) to a point, but then begin to fall as anxiety continues to rise. This scenario seems plausible enough - some anxiety is good and keeps us from getting complacent, but too much can inhibit our performance. What does this do to our correlation though? The calculated correlation for this non-linear sample is *r*(*xy*)=0.35, indicating a very weak correlation between test scores (*y*) and anxiety (*x*). This is when correlation can be misleading. Visual inspection of Fig. 2 shows a clear association; however, a linear measure of correlation does not detect it. Later posts will cover options for detecting non-linear associations between variables.

To conclude, when we talk about correlation, we are speaking of a number that represents the true association between variables. That number may be more or less accurate depending on whether or not we have used the appropriate measure of correlation for the data we have. If correlation is detected we know there is some association; however, we cannot assume that variables are independent of each other (i.e., not associated) just because no correlation is detected.

# Key Ideas:

* Correlation measures the association between variables.
* Correlation is represented within a range of -1 to 1.
* Linear correlation is typically measured with Pearson’s *r*.
* It cannot be assumed that variables are independent just because *r*=0.
* Pearson’s *r* is insufficient for detecting non-linear correlation.