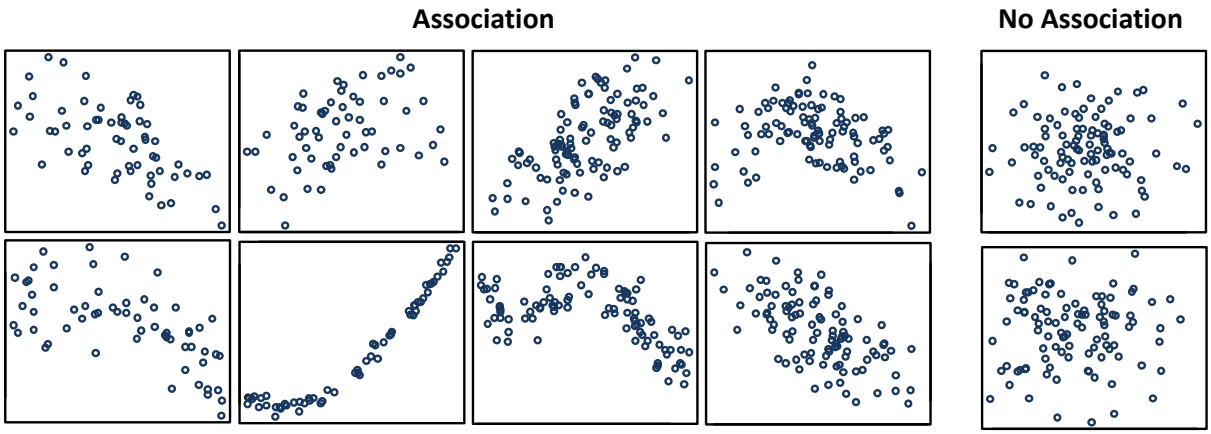
# Association and Correlation

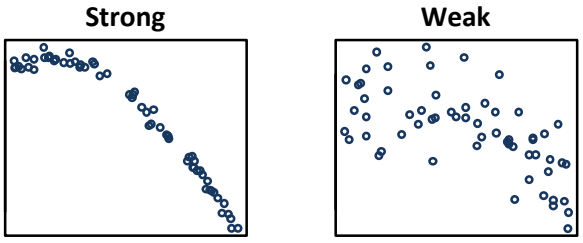
*Chris Wild, University of Auckland*

**This article explains terms that are often used to describe a relationship between two numeric variables.**



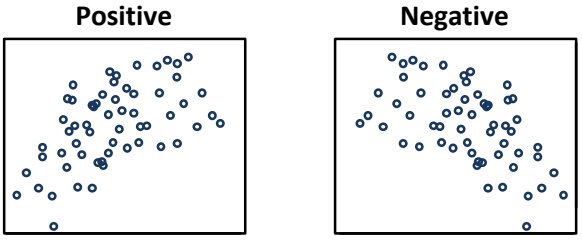

Statisticians say two variables are ***associated*** if there is if there is a pattern in the scatterplot that is too strong to be likely to arise simply by chance. Otherwise there is ***no association***.

**Strong *versus* Weak Association**




The association can be **strong** (very little scatter compared to the movement in the trend) or **weak** (lots of scatter around the trend).

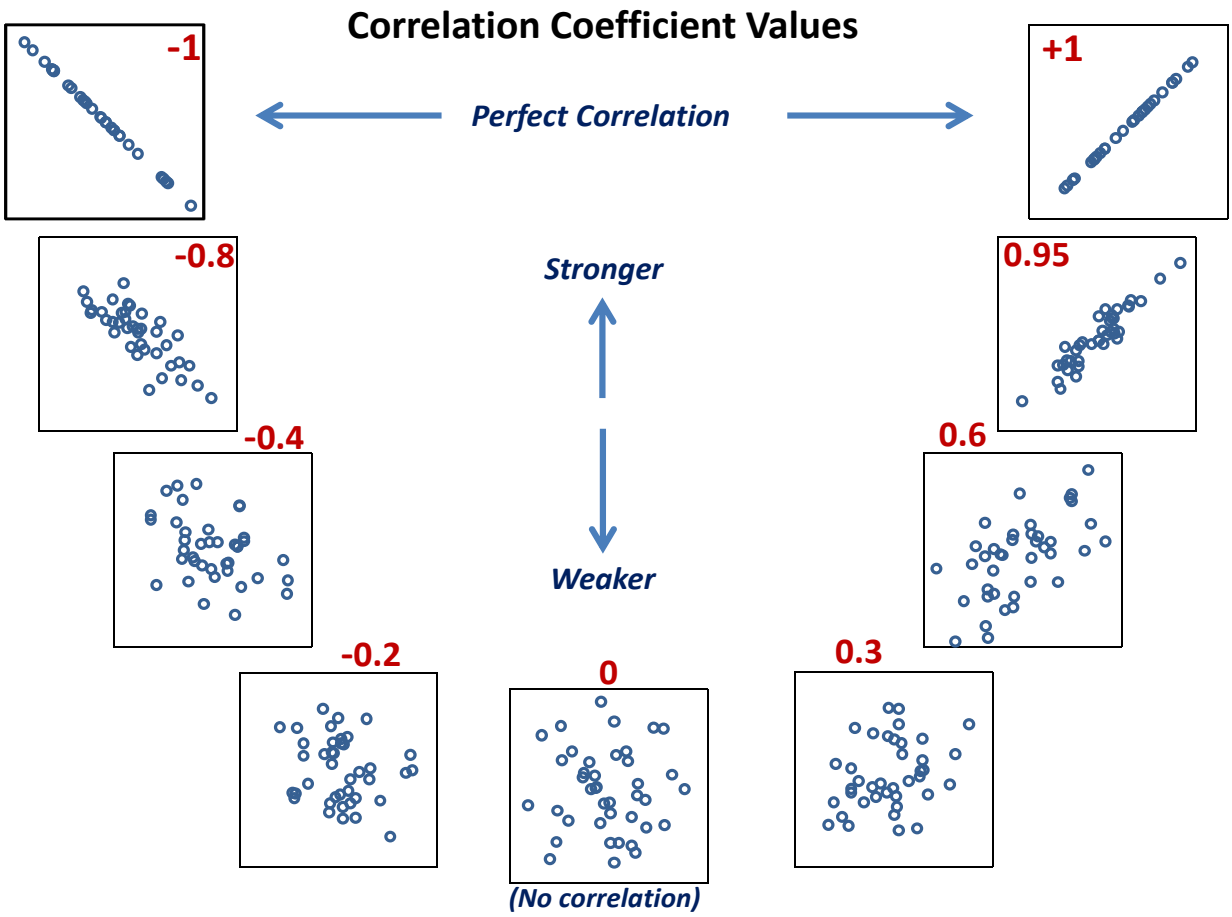
**Positive *versus* Negative Association**




An association is called **positive** if ***y*** tends to ***get bigger*** when ***x gets bigger*** and **negative** if ***y*** tends to ***get smaller*** as **x *gets bigger***.

## CORRELATION

Correlation measures a specific form of association. The most often quoted ***correlation*** is the "Pearson correlation" which is ***relevant*** to relationships with ***a linear trend***. It is a measure of how close the points are to lying on a straight line. This picture shows how it works:




Correlations take values between -1 and +1. A correlation of **+1** occurs when all of the points lie ***exactly on a line*** that has a ***positive*** slope. A correlation of **-1** occurs when all of the points lie ***exactly on a line*** that has a ***negative*** slope. The correlation is ***zero*** if there is ***no relationship***. The picture shows the intermediate values. As the relationship gets stronger, the correlation gets closer to either -1 or +1. As the relationship gets weaker, the correlation gets closer to zero. (iNZight gives you the correlation if you put a line on the scatterplot and then click "Get Summary").

The form of correlation relevant to variables that have a ***curved trend***, is called "***Spearman's rank correlation***". It measures the strength of any positive or negative association. The rank correlation again falls between -1 and +1. If every time *x* gets bigger, *y* also gets bigger, then the rank-correlation will be +1. It is -1 if whenever *x* gets bigger, *y* gets smaller. When there is no discernible upward or downward drift the rank correlation will be close to zero.

**"Correlation does not imply causation"**

When there is a strong relationship in a scatterplot, people tend to jump to a premature and often false conclusion that changes in the predictor are actually causing changes in the outcome. You can never reliably conclude "this is what did it" just by seeing a strong correlation (or any other form of strong relationship) between variables. The first few paragraphs of **"Causation and Correlation"** from the Statistical Assessment Service provides a very good discussion. Their main example concerns a strong correlation between the rise of the use of Facebook and the deterioration in the Greek economy.

The site **Spurious Correlations** searches through huge numbers of data series to find things that have tended to go up and down at approximately the same times and calculates correlations between them. At the time of writing the headline example is a 0.99 correlation between "US spending on science, space, and technology" and "Suicides by hanging, strangulation and suffocation". There is also a 0.99 correlation between "Divorce rate in Maine" and "Per capita consumption of margarine (US)".

We pick up this issue in Week 5.



- **See also** **Guessing Correlations**

  Guess-and-check way of learning to interpret the sizes of correlation coefficients (this only applies to scatter plots with linear trends)

- Alternative guess the correlation game
- xkcd cartoon (look - then hover your mouse over the last frame)
- How summary statistics can tell less than half the story …