Comparing Two Continuous Variables

STA 102: Introduction to Biostatistics

Yue Jiang

The following material was used by Yue Jiang during a live lecture.

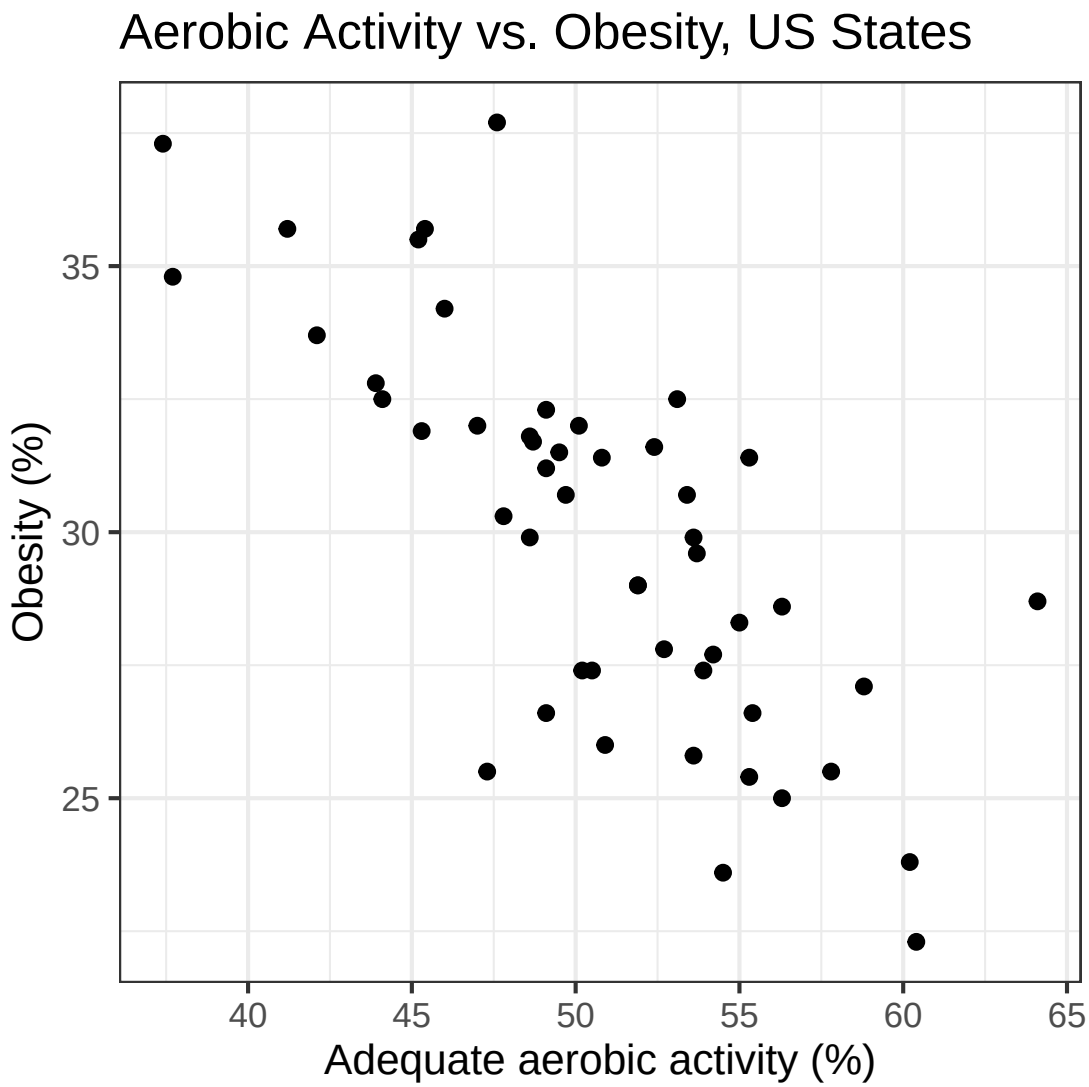
Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Motivation

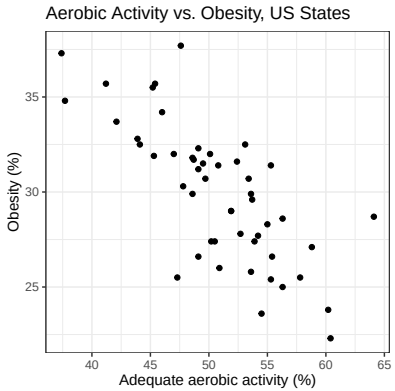
Comparisons of interest:

- Categorical predictor vs. categorical outcome (Fisher's exact test or χ^2 test)
- Categorical predictor vs. continuous outcome (t-tests, ANOVA, and their non-parametric alternatives)
- Continuous predictor and continuous outcome (will begin discussion today!)
- Continuous predictor and categorical outcome (will discuss later)
- ...even more exotic types of comparisons (even later still)

Remember this plot?



Remember this plot?



Some questions of interest may include:

- Direction of relationship: are the variables positively or negatively related?
- Form: is any relationship linear or more complex?
- Strength of relationship: how accurately can one variable predict the other?
- Influential points: are one or a few points driving the relationship we see?

Correlation

The correlation coefficient ρ quantifies the *linear relationship* between two random variables.

In statistics, a correlation coefficient implies a very specific type of association. A correlation coefficient of zero does NOT imply no relationship between two variables, as we shall see in some further examples.

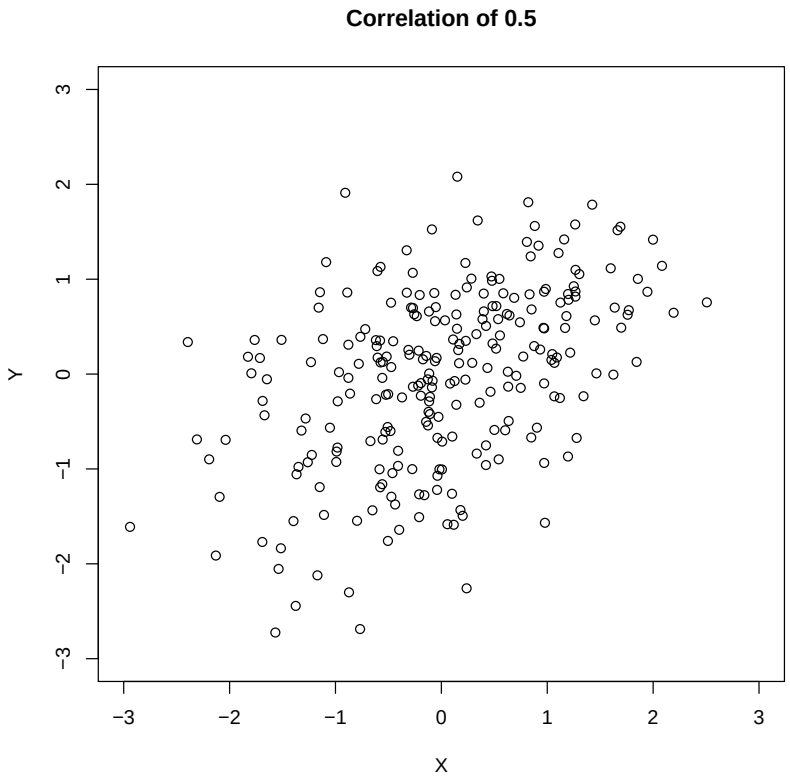
Correlation

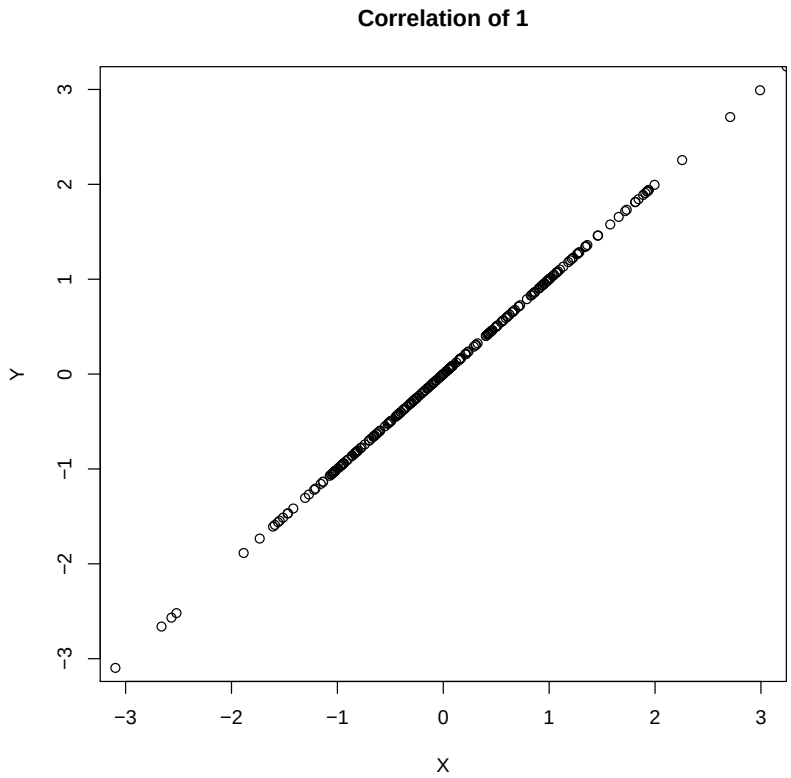
ρ ranges from -1 to 1

- $\rho > 0$ implies positive correlation
- $\rho < 0$ implies negative correlation
- $\rho = 0$ is consistent with no *linear* relationship between variables

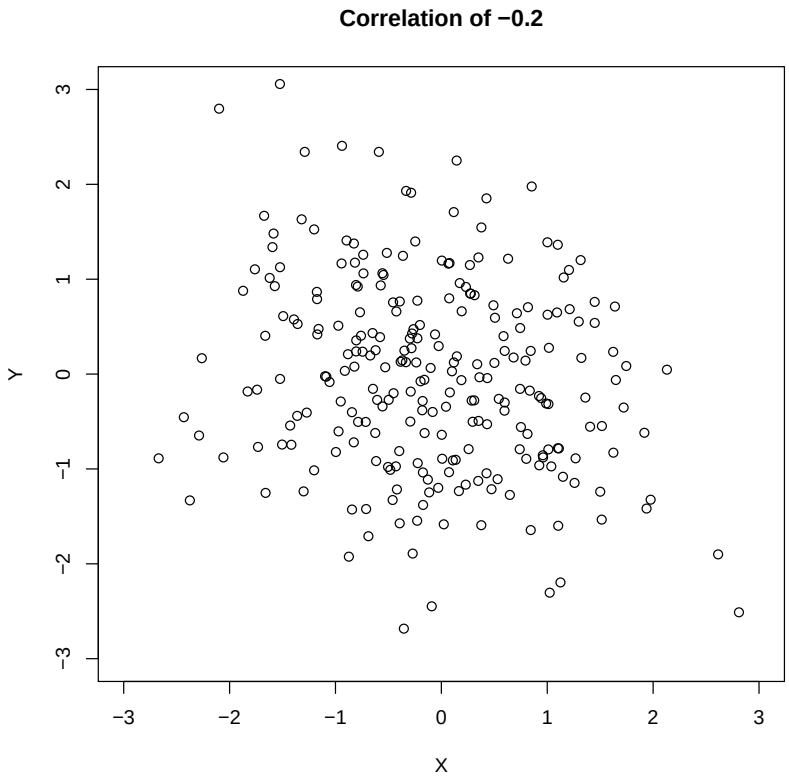
What does it mean to have a correlation of -1 or 1?

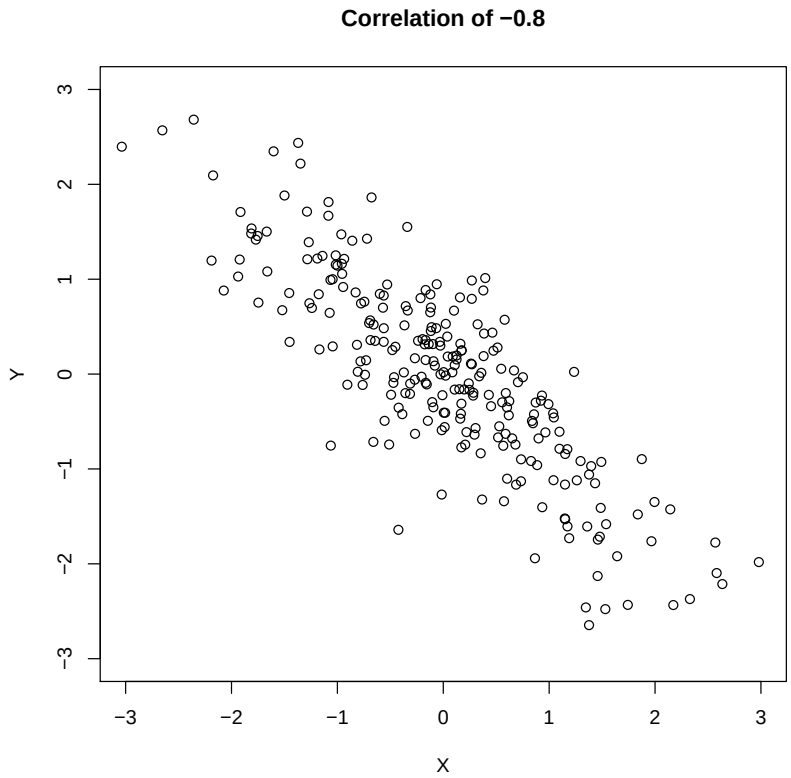
Visualizing ρ



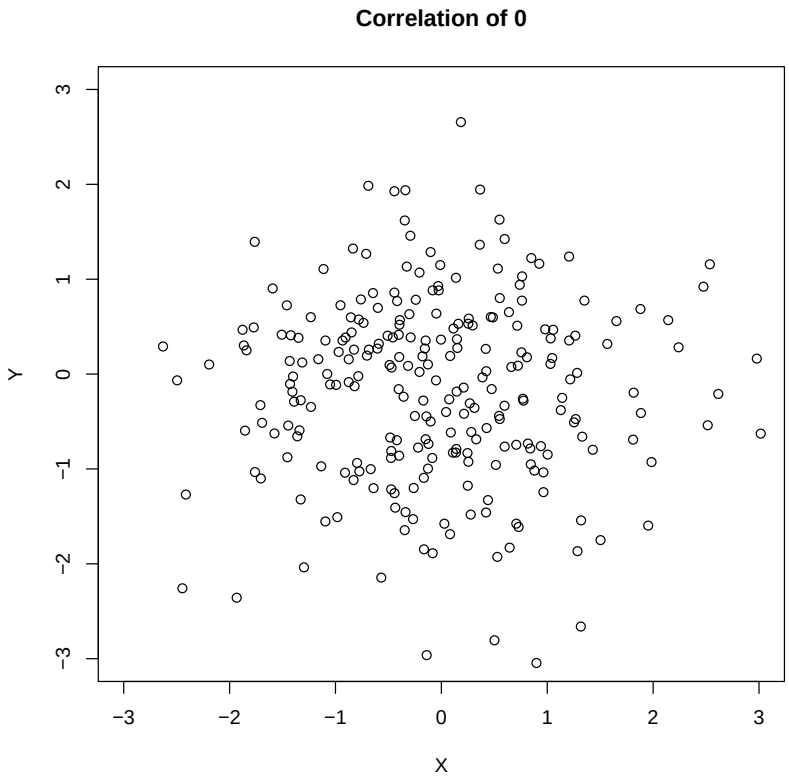


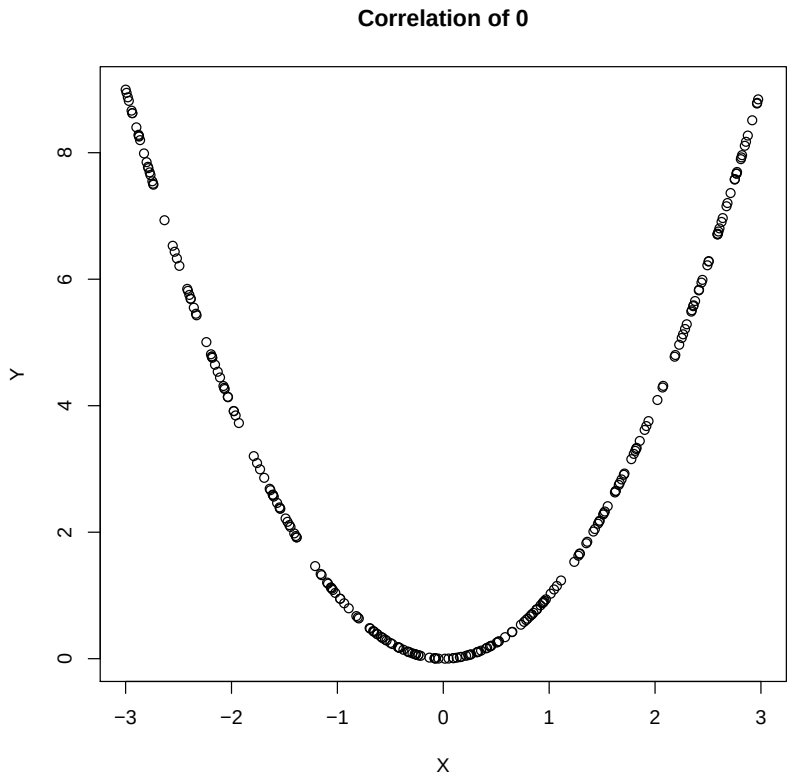
Visualizing ρ





Visualizing ρ





Pearson's correlation coefficient

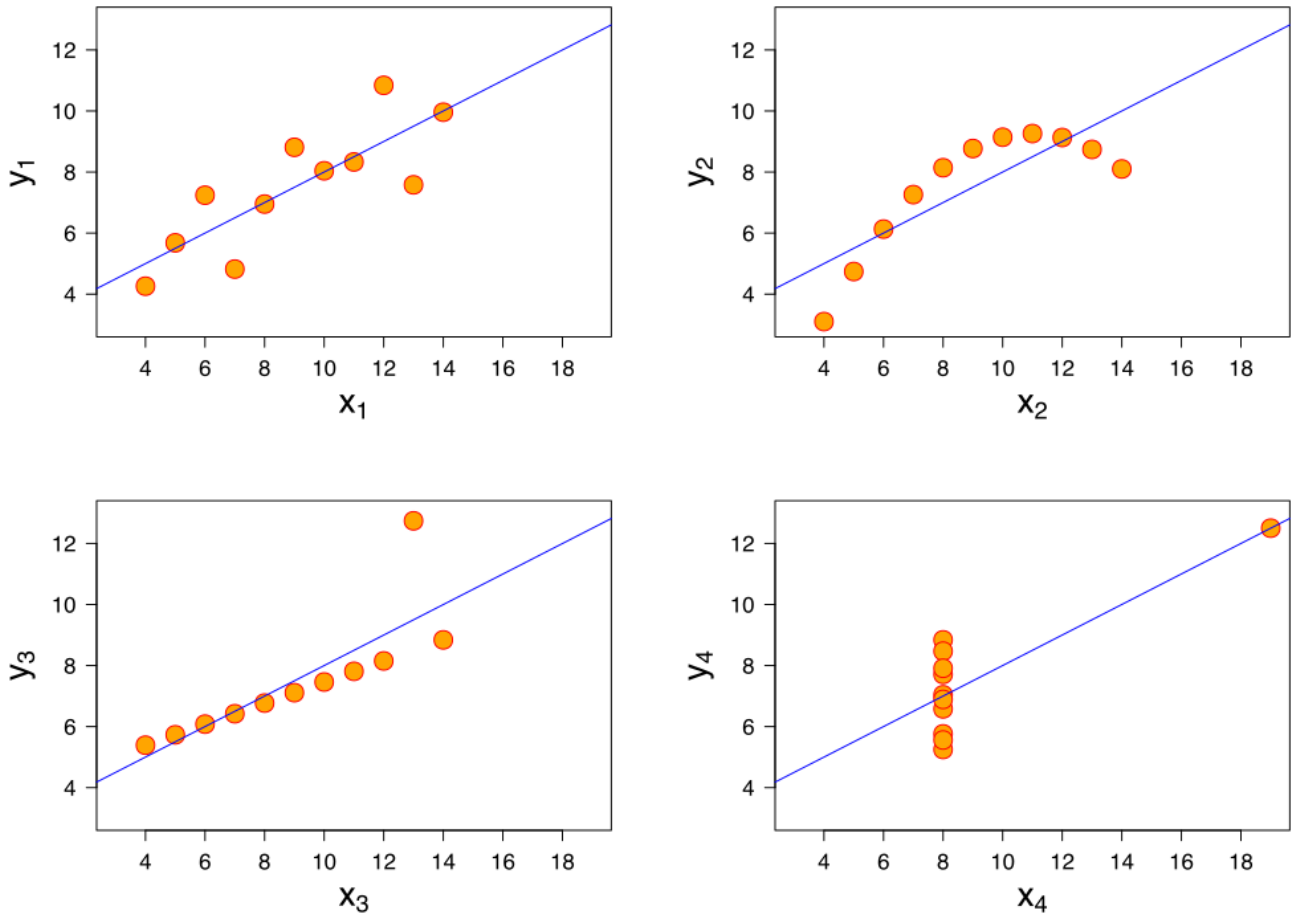
r (Pearson's correlation) gives an estimate of ρ for the pairs $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$:

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{S_x} \right) \left(\frac{y_i - \bar{Y}}{S_y} \right)$$

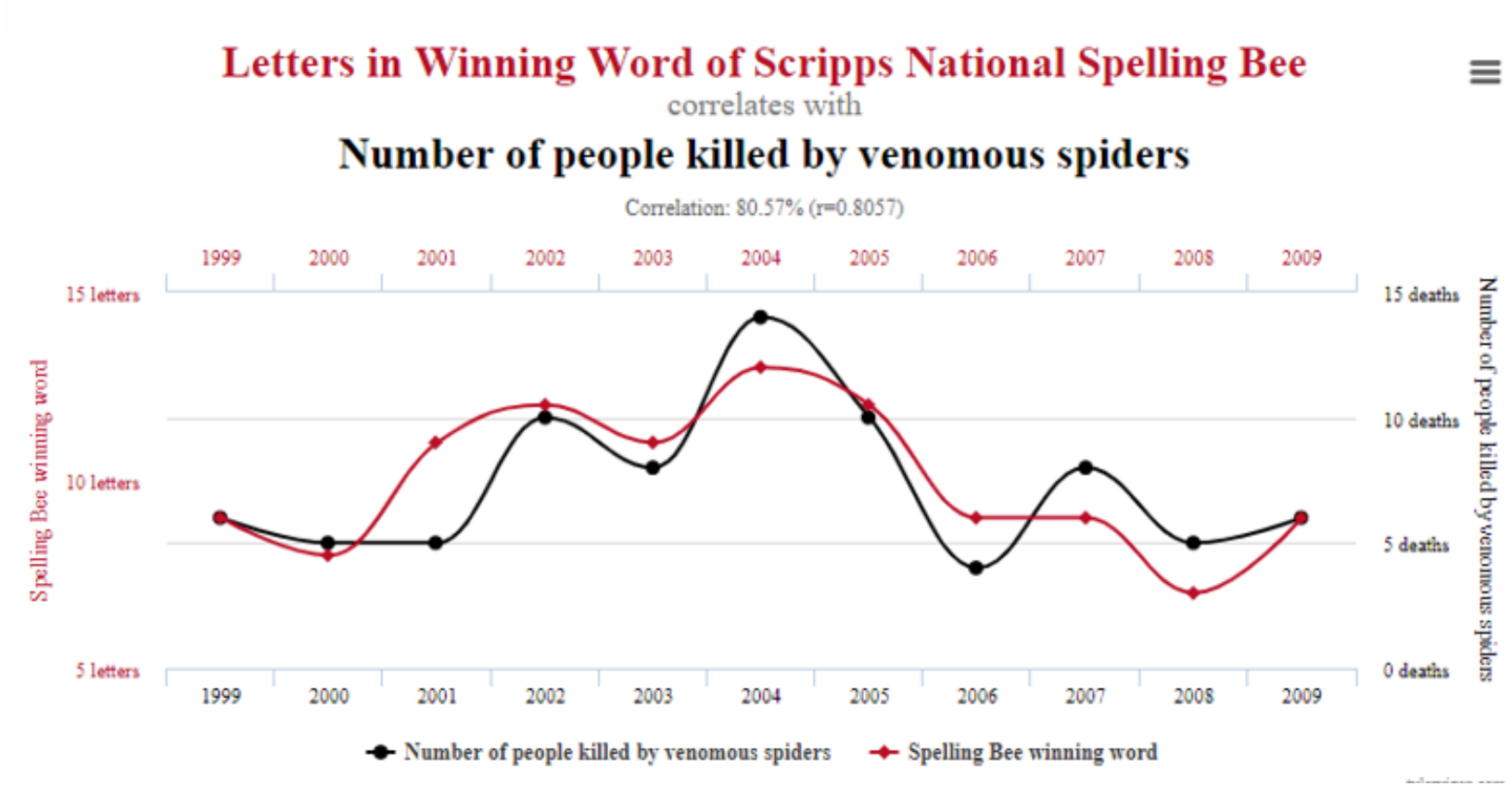
$$= \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{X})^2 \sum_{i=1}^{n} (y_i - \bar{Y})^2}}$$

...we'll just use the `cor()` in R.

Anscombe's quartet



Correlation does not imply causation



Source: Tyler Vigen, Spurious Correlations

Confounding

Many of spurious correlations are due to confounding – when a third lurking variable responsible for the observed relationship.

Testing correlation

Suppose we wish to test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$. If we assume that our two variables are normally distributed, then we can use a t-statistic to test this hypothesis (don't worry about the exact details; we'll do this using R).

Spearman's correlation coefficient

Remember, Pearson's correlation is sensitive to outliers, and only measures the *linear* association. What can we do when we have outliers or are interested in non-linear association?

Spearman's correlation is simply the Pearson's correlation calculated on the *ranks* of the data instead of the original dataset.