Statistically Speaking: *Correlation is not causation*

DAVID L. TABB, PH.D.

SEPTEMBER 7, 2017

Overview

- Causation is a difficult beast
- Essential concepts
- Magnitudes vs. ranks: Pearson vs. Spearman

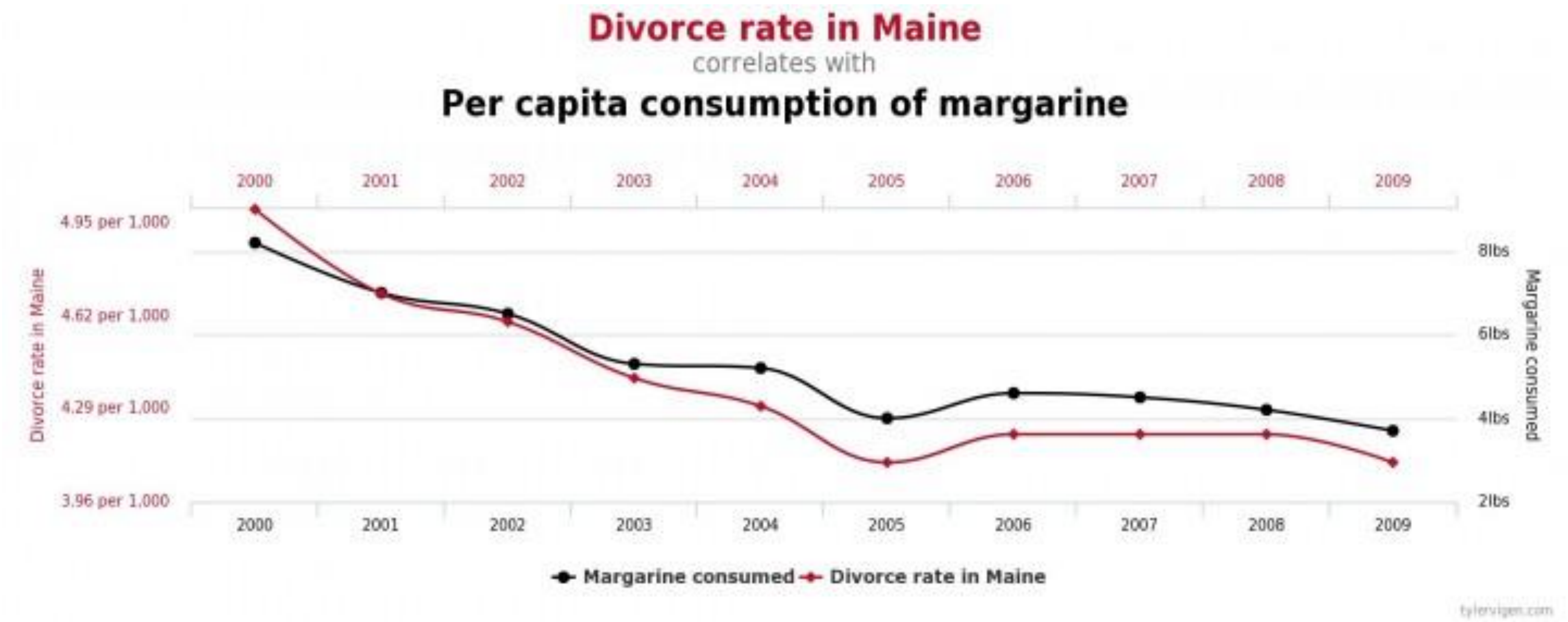
Austin Bradford Hill Causation

- Temporal Relationship
- Strength
- Dose-Response Relationship
- Consistency
- Plausibility
- Consideration of Alternate Explanations
- Experiment
- Specificity
- Coherence

Wikimedia Commons

http://www.drabruzzi.com/hills_criteria_of_causation.htm

No, not even when the data lie on top of each other



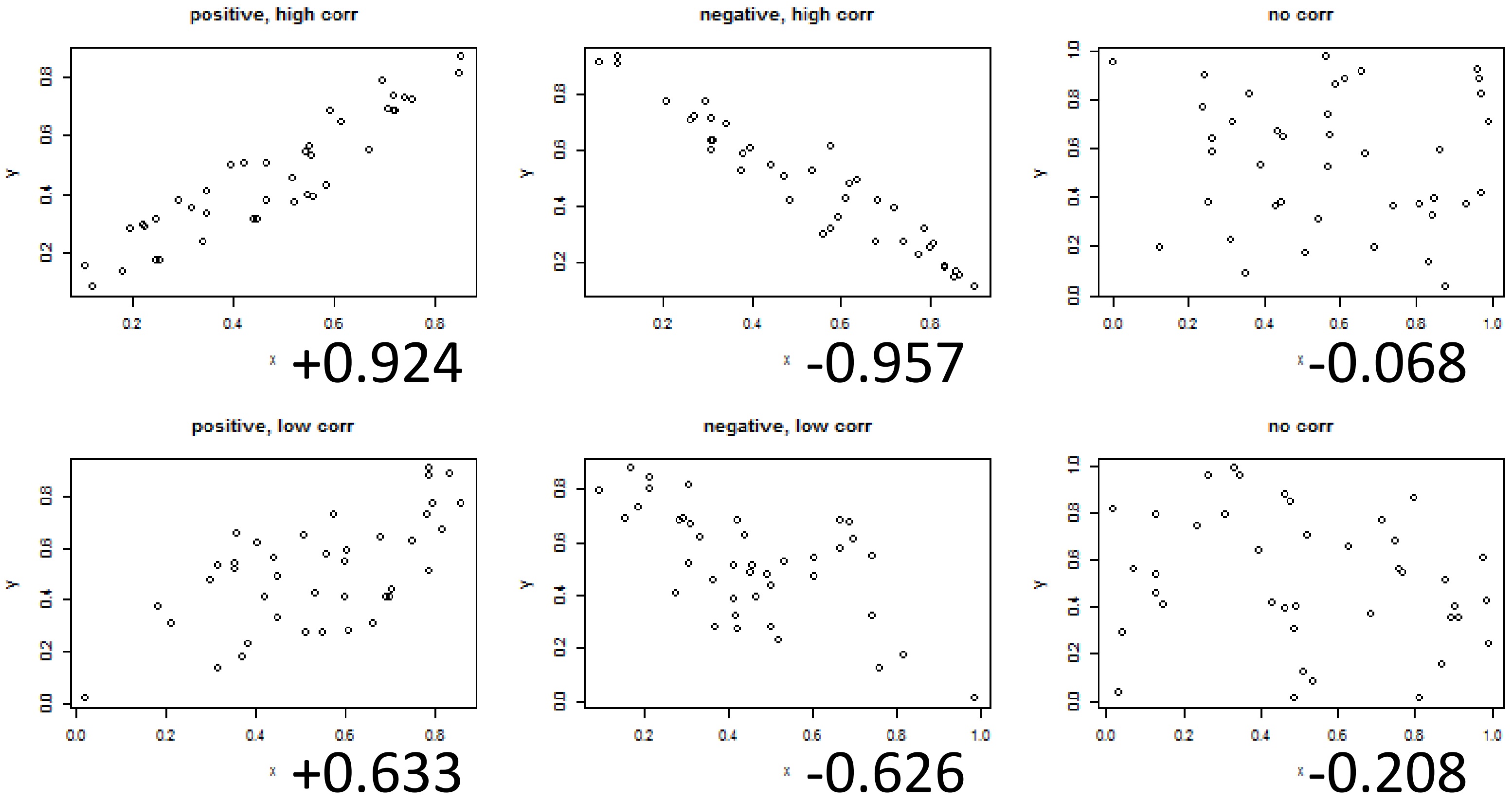
http://www.insightsassociation.org/article/correlation-does-not-equal-causation-and-why-you-should-care

Correlation Concepts

- Positive: higher A values are associated with higher B values.
- Negative: higher A values are associated with lower B values.
- High: we observe a strong association between A and B.
- Low: we observe little association between A and B.

Correlation is commutative; `cor(a,b) = cor(b,a)`.

Examples from random simulations



positive, high corr +0.924

negative, high corr -0.957

no corr -0.068

positive, low corr +0.633

negative, low corr -0.626

no corr -0.208

Pearson correlation values produced through cor function

Partial code for prior slide

```
#Set up for four graphs in one pane
par(mfrow=c(2,3))

#Positive, high correlation
a <- runif(40)
b <- runif(40)
x <- 0.6 * a + 0.4 * b
y <- 0.4 * a + 0.6 * b
plot(x,y,main="positive, high corr")
cor(x,y)
```

Karl Pearson: magnitudes matter

- Technically, $\mathrm{r} = cov(x, y)/(sd(x) * sd(y))$
- "Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together." http://www.statisticshowto.com/covariance/

Wikimedia Commons

Sadly, he was also a eugenicist.

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

Charles Spearman: ranks matter

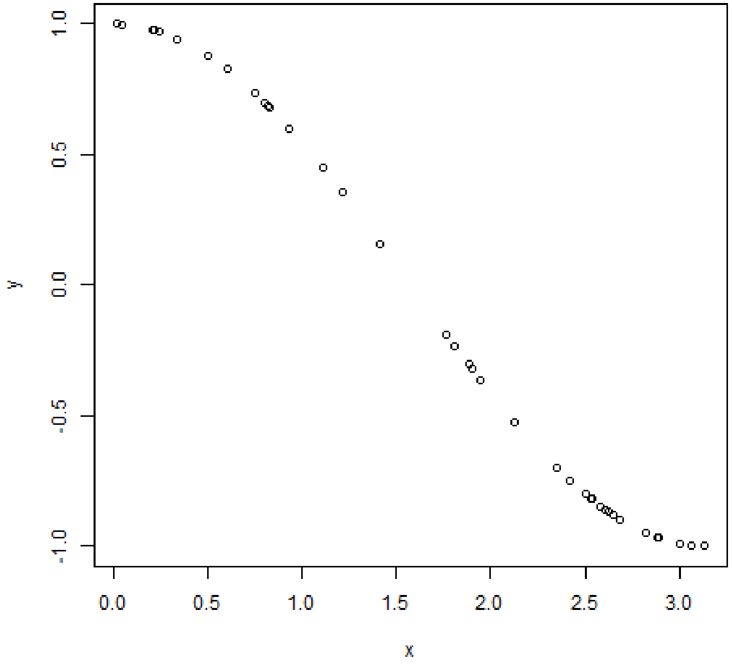
- Technically, $\rho = \frac{cov(ranks(x),ranks(y))}{sd(ranks(x)*sd(ranks(y))}$
- The ranking of values within x or within y, rather than their magnitudes, drives the Spearman correlation.
- Dumping magnitude makes Spearman correlation robust against outliers.

Wikimedia Commons

Argued that general intelligence factor may be measured many ways.

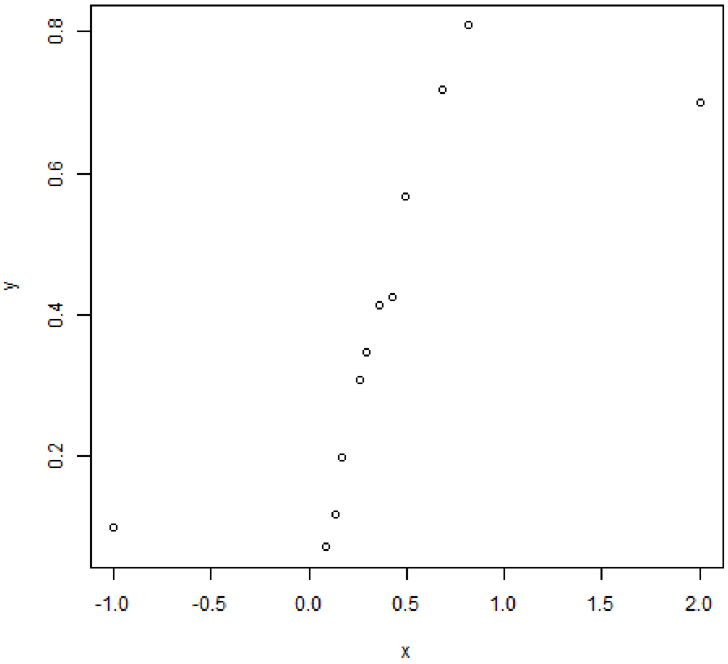
Same input, different outputs

```
x <- pi*runif(40)
y <- cos(x)
plot(x,y)
cor(x,y,
method="pearson")
[1] -0.9926561
cor(x,y,
method="spearman")
[1] -1
```



Resisting outliers

```
a <- runif(10)
b <- runif(10)
x <- c(2,-1,
(0.55*a+0.45*b))
y <- c(0.7,0.1,
(0.45*a+0.55*b))
plot(x,y)
cor(x,y, method="pearson")
[1] 0.7461581
cor(x,y, method="spearman")
[1] 0.972028
```



Takeaways

- Correlation analysis lets us compare the values between two sets for association.
- The two sets are treated as independent of each other; neither is a function of the other.
- The non-parametric Spearman method is preferable when data contain outliers.