

Day 07

Measuring associations

- Calculating correlation
- Limitations
- False positives
- Visual inference

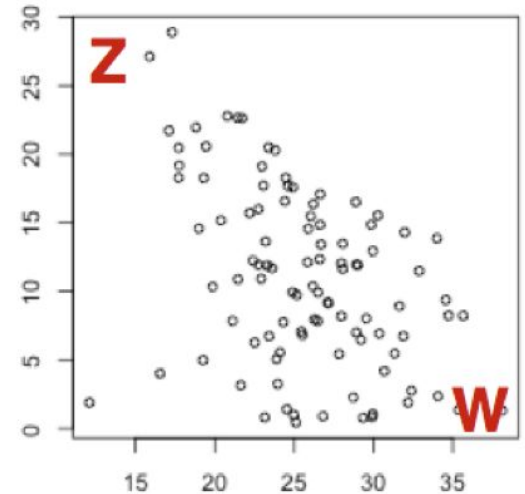
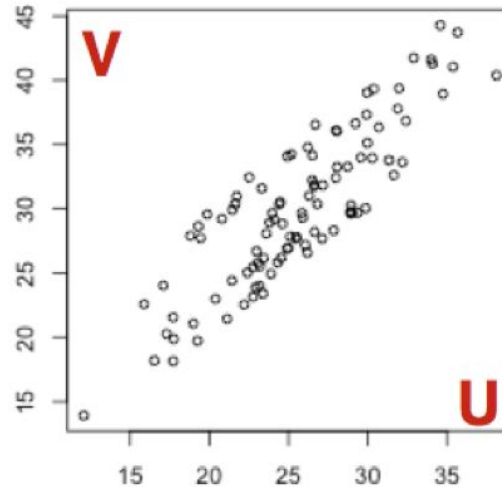
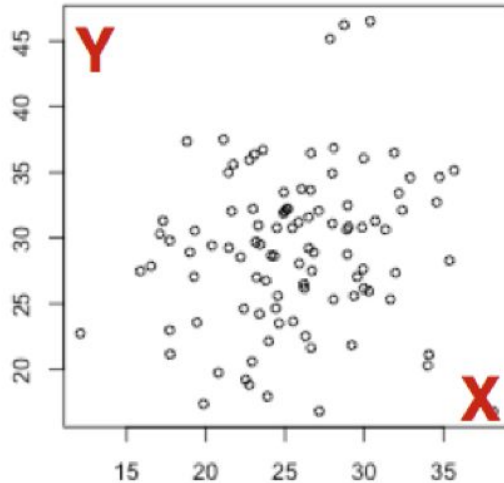
Calculating correlation

Variables

Attributes / Features



x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68



Correlation coefficient

Pearson Correlation Coefficient

- Measures linear relationship between variables.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- n is the sample size
- x_i, y_i are the single samples indexed with i
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (the sample **mean**); and analogously for \bar{y}

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

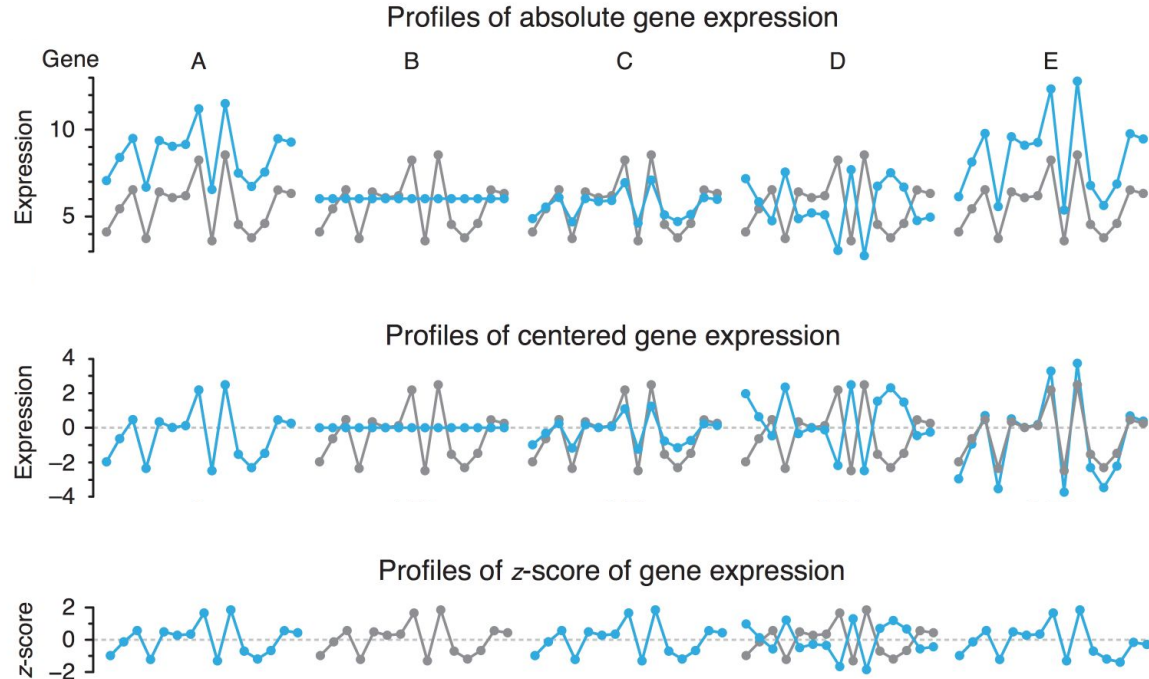
Correlation coefficient

Pearson Correlation Coefficient

- Captures the relationship between 2 vectors after centering each vector by its mean and scaling by its standard deviation.
- The final quantities for each vector are called z-scores.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Diagram showing two arrows pointing from the terms $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ in the equation to the text "The final quantities for each vector are called z-scores."

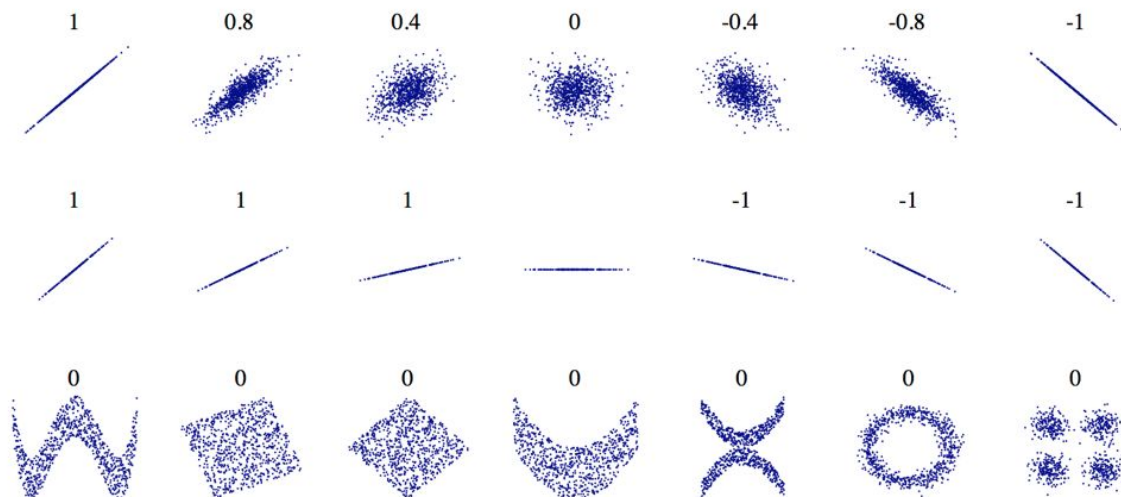


Correlation coefficient

Pearson Correlation Coefficient

- Measures 'linear' relationship between variables.

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$



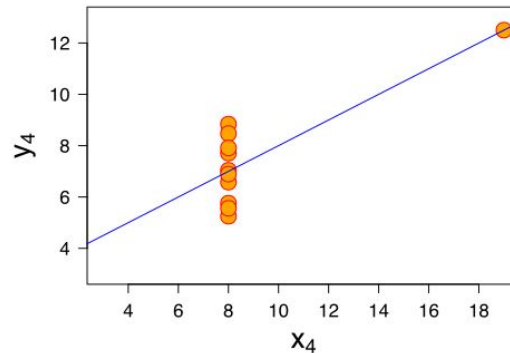
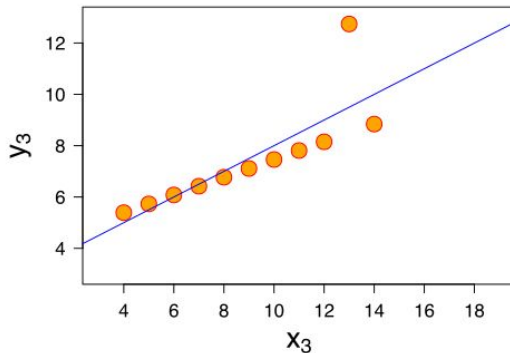
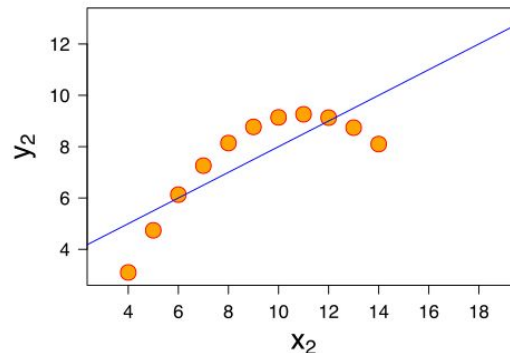
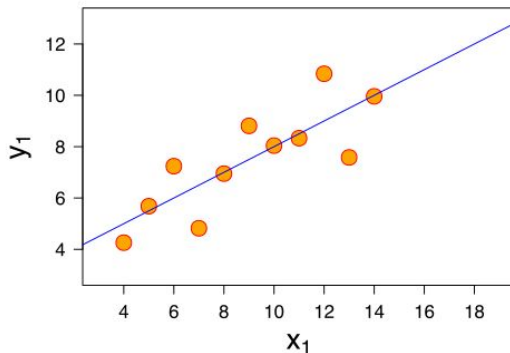
$$-1 \leq r \leq +1$$

-1 is total -ve correlation | 0 is no correlation | +1 is total +ve correlation

Anscombe's quartet: "calculation are exact; graphs are rough!"

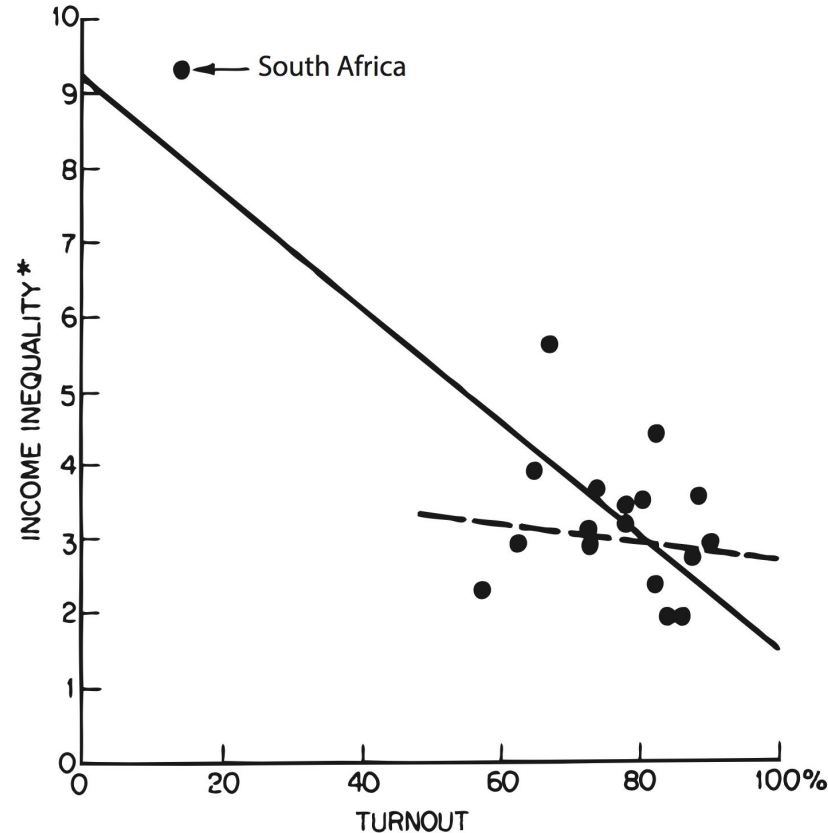
11 data points

- Mean (x) = 9
- Var (x) = 11
- Mean (y) = 7.50
- Var (y) ~ 4.12
- Cor (x, y) = 0.816
- Linear regression line:
 - $y = 3.00 + 0.500x$



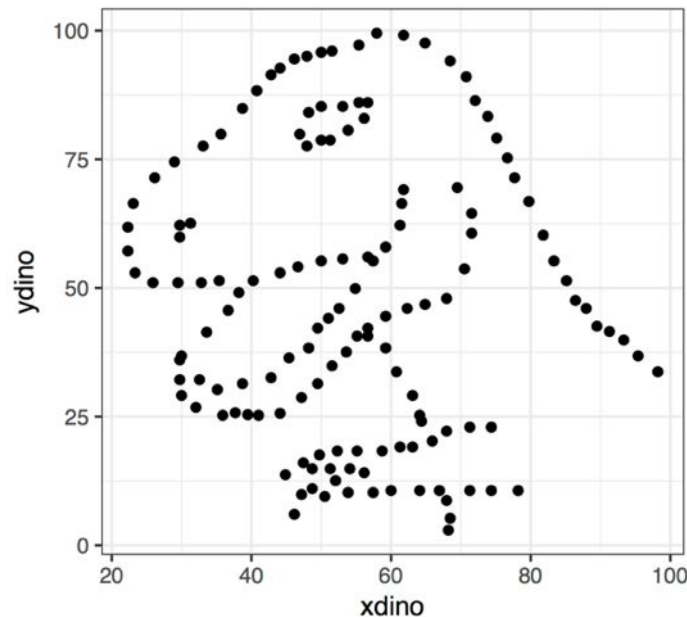
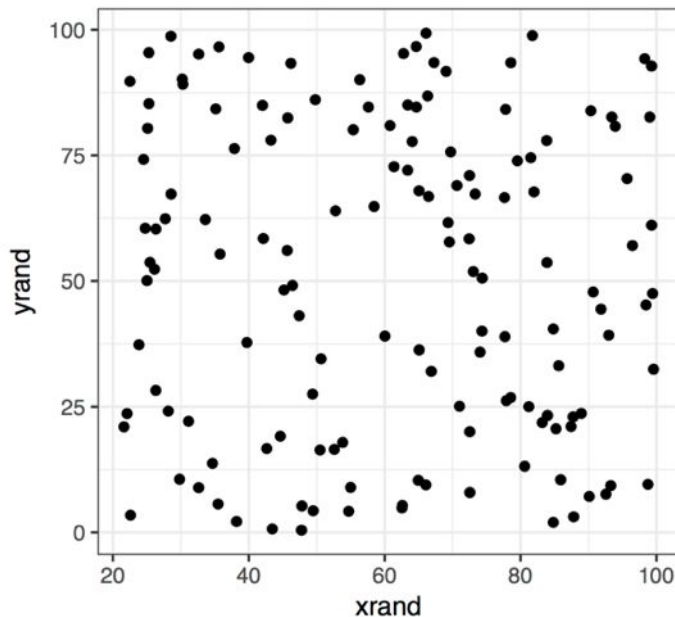
Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

What does a correlation coefficient tell you about the data?



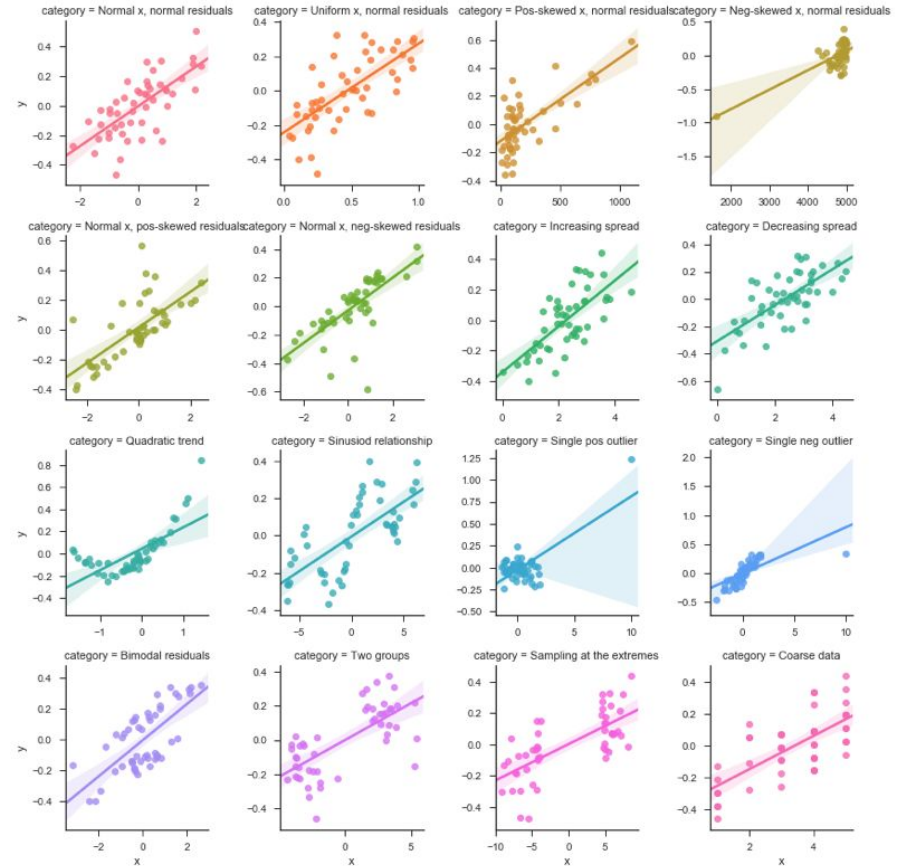
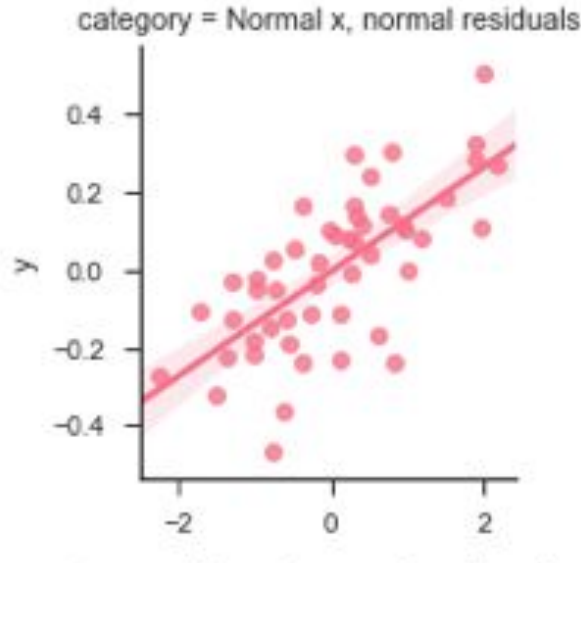
What does a correlation coefficient tell you about the data?

Correlation = -0.06



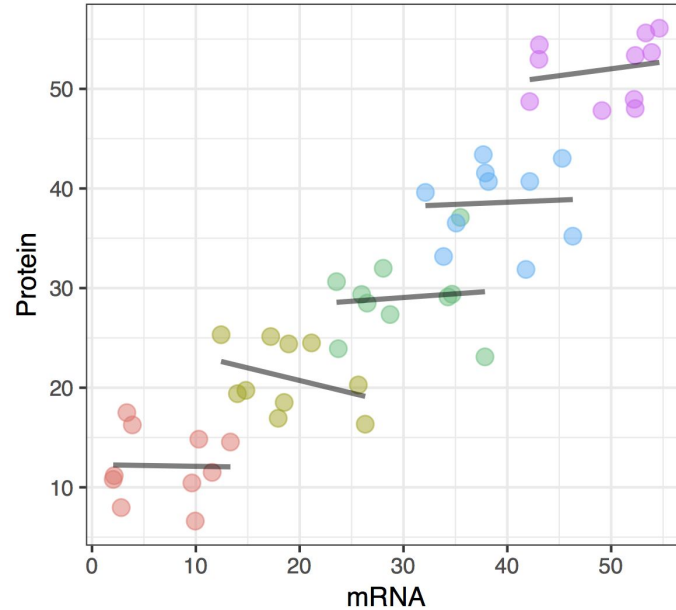
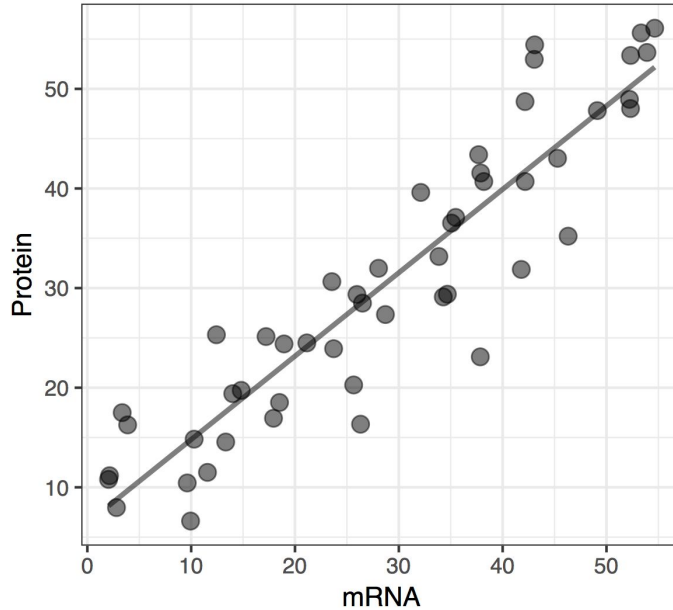
What does a correlation coefficient tell you about the data?

Correlation = 0.7



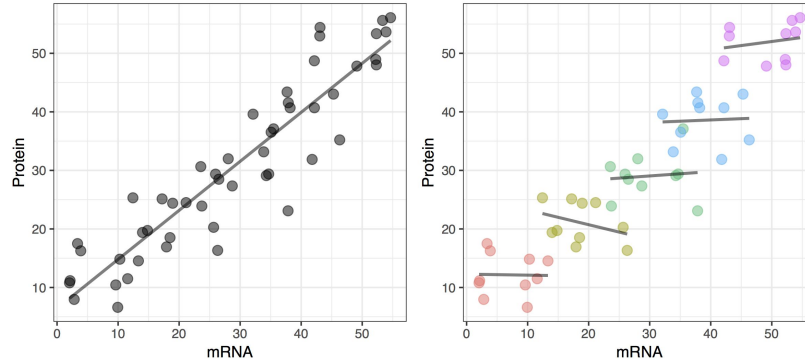
What does a correlation coefficient tell you about the data?

Simpson's Paradox



What does a correlation coefficient tell you about the data?

Simpson's Paradox

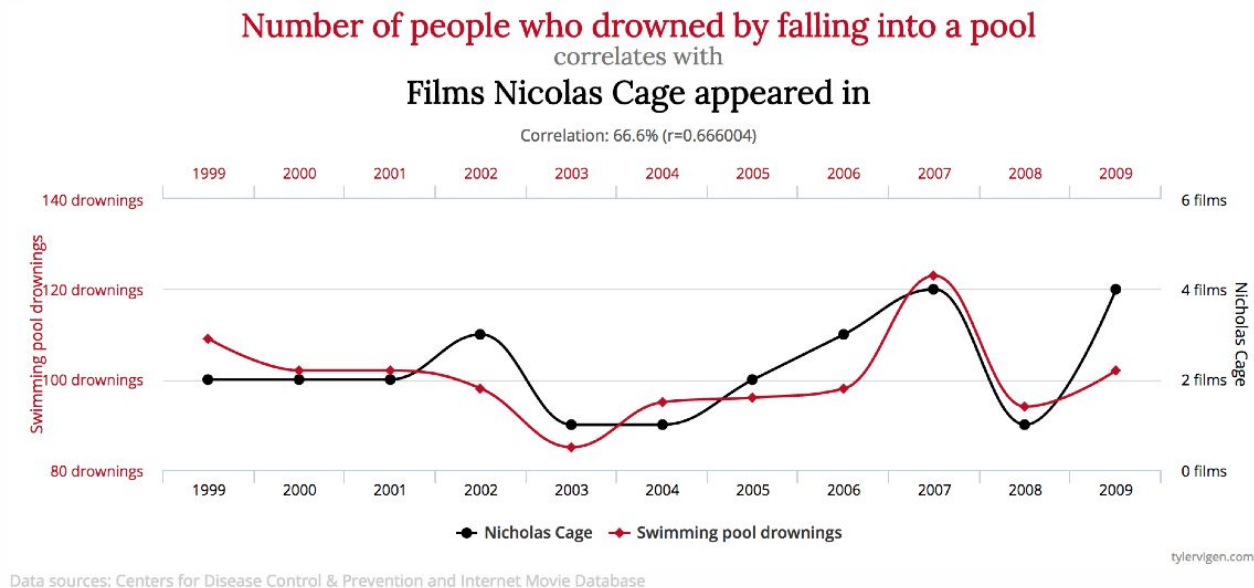


Success rates of kidney stone removal surgeries

Treatment	Diameter < 2 cm	Dia. \geq 2 cm	Overall
Open surgery	93%	73%	78%
Percutaneous nephrolithotomy	87%	69%	83%

Spurious correlations

What does Nicholas Cage have to do with people drowning in swimming pools?



Checkout <https://www.google.com/trends/correlate>

Spurious correlations

Simulate fluctuations in correlation coefficients

- Repeat 10,000: Calculate correlation coefficients of $n = 10$ samples of two independent uniformly distributed variables between (0, 1). Plot a histogram.
- Mark statistically significant coefficients ($\alpha = 0.05$).

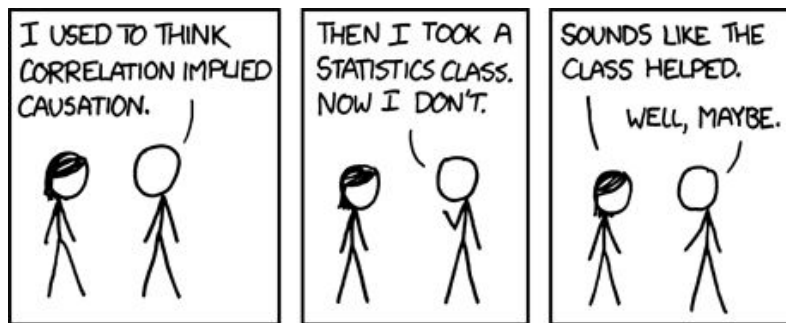
Vary sample size $n = \{5, 10, 20, 50\}$.

- For each, simulate the distribution of correlation coefficients and mark the coefficient corresponding to $\alpha = 0.05$.

Correlation does not imply causation

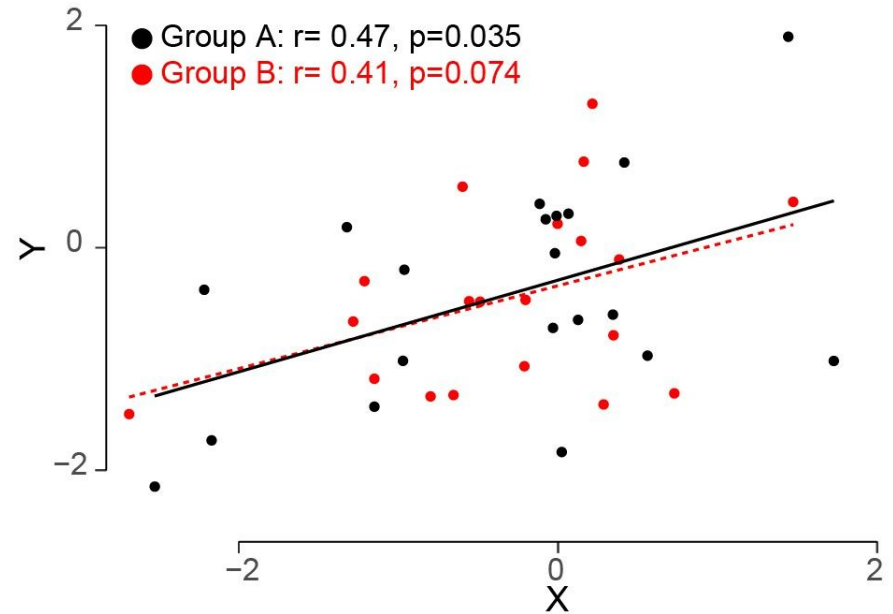
There is a significant correlation between annual chocolate consumption and number of Nobel laureates for different countries ($r(20)=.79$; $p<0.001$) → chocolate intake provides nutritional ground for sprouting Nobel laureates.

- Correlation can occur by random chance.
- Confounding variables could lead to correlation.
- Even when there is causation, there might not be obvious correlation.

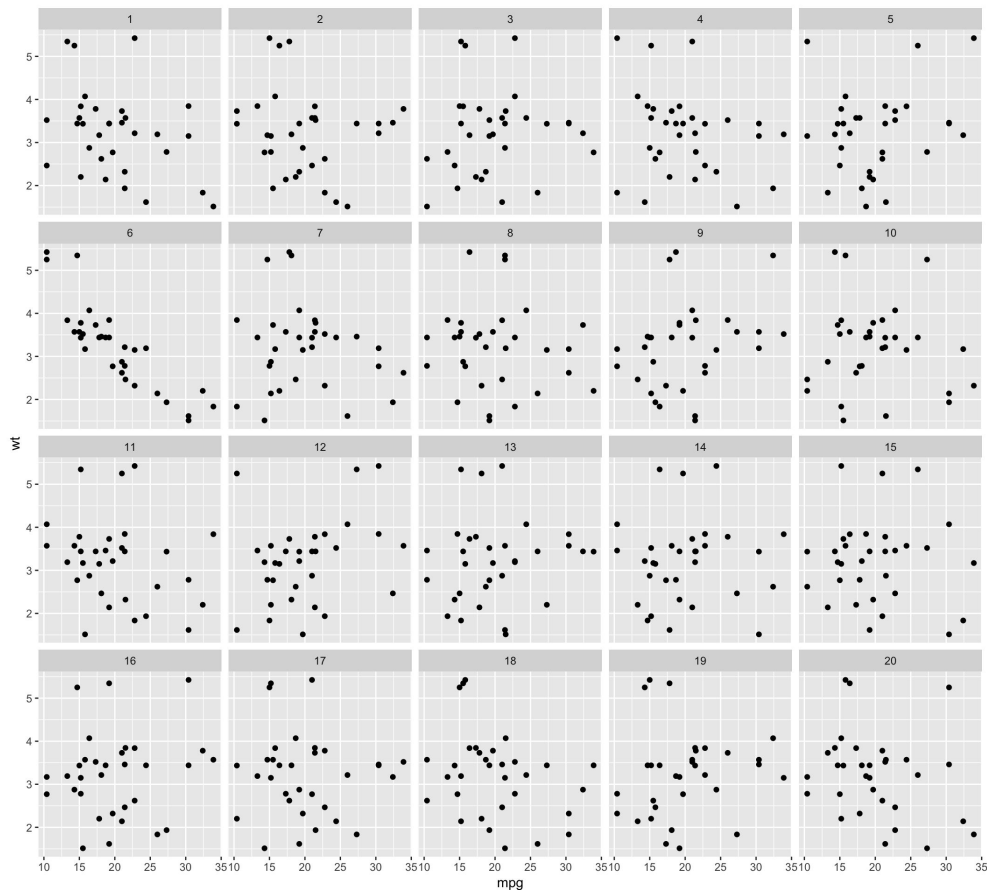


Interpreting comparison b/w two correlation w/o comparing them

Conclusion regarding the impact of an intervention based on correlation in treatment group vs. correlation in control group.



Spurious correlations – But it *looks* associated!

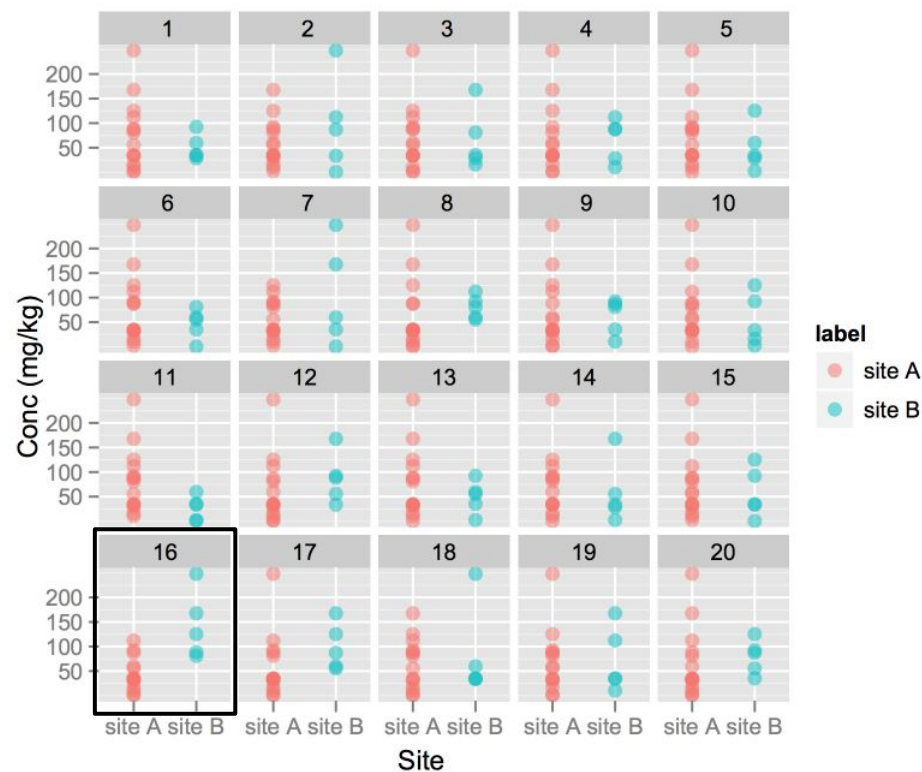
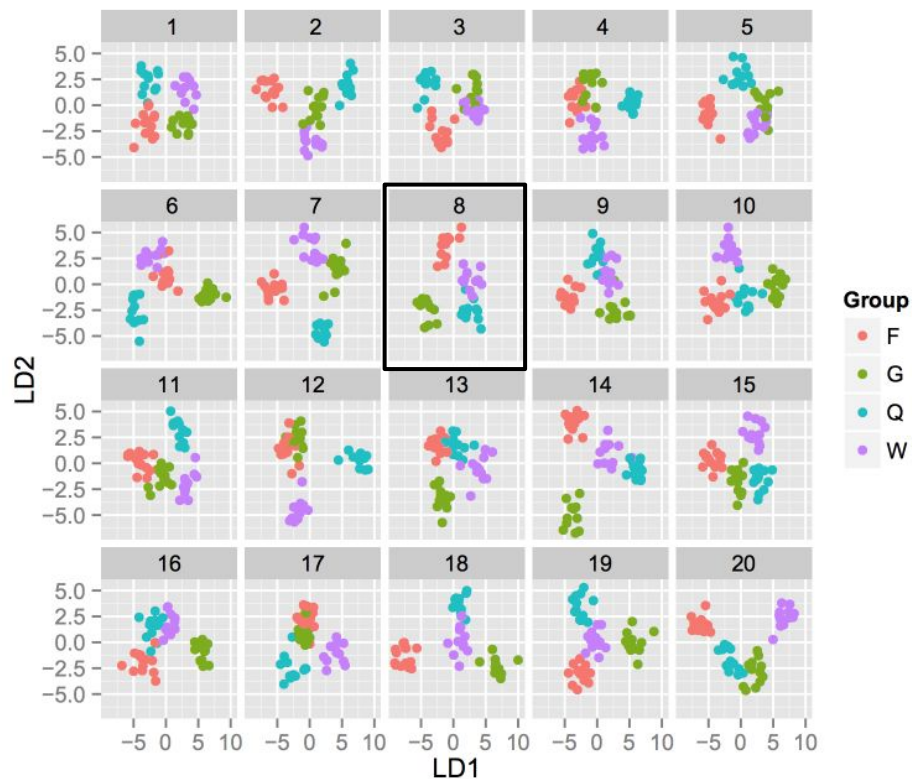


Create a lineup for visual inference

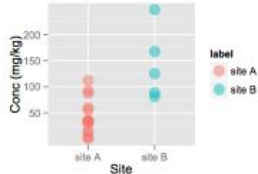
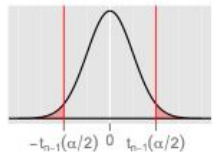
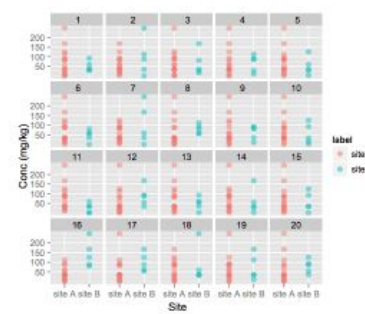
Place the plot of the real data amongst a set of null plots to create a lineup; Null plots are generated in a way consistent with the null hypothesis.

If the observer can pick the real data as different from the others, this puts weight on the statistical significance of the structure in the plot.

Spurious correlations – But it *looks* associated!



Spurious correlations – But it *looks* associated!

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T(y) =$ 
Sampling Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
Reject H_0 if	observed T is extreme	observed plot is identifiable