ISTA 116: Statistical Foundations for the Information Age

Multivariate Numeric Data

26 and 28 September 2011

Outline

Reminders/Announcements

- Web Quiz 4 due Wednesday.
- Lab 3 due Friday by 5 P.M.

Multivariate Data: Three Cases

- The kinds of relationships we can identify depend on the types of variables we have
- Three Cases:
  - All categorical variables ✓
  - A mix of categorical and numeric ✓
  - **All numeric**

Multiple Univariate vs. One Multivariate

- What's the difference between this...

Person	Sex
1	M
2	F
3	F
4	M
5	F
6	F
7	M
8	M

Person	Height (in.)
A	64
B	74
C	72
D	68
E	61
F	70
G	68
H	69

Multiple Univariate vs. One Multivariate

- and this?

Person	Sex	Height (in.)
1	M	74
2	F	64
3	F	61
4	M	68
5	F	70
6	F	69
7	M	72
8	M	68

Grouped vs. Paired Numeric Data

- What's the difference between this...

Sex	Height (in.)
M	74
F	64
F	61
M	68
F	70
F	69
M	72
M	68

$\Longleftrightarrow$

Females
64
61
70
69

Males
74
68
72
68

Grouped vs. Paired Numeric Data

- and this?

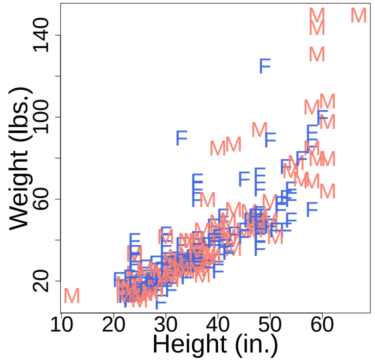
Person	Height Age 10	Height Age 12
1	58	62
2	49	54
3	52	53
4	55	60

- Instead of a numeric variable and a grouping variable, we now have two numeric variables observed *from the same people*
- Numeric-numeric relationships are more complicated, since we can no longer just compare groups.

The Scatterplot

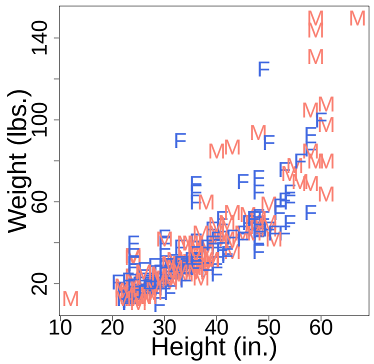
- With two numeric values *from the same source* (e.g. from the same person), we can represent each person (say) as a point in 2D space.
- If we plot all of these points, we obtain a **scatterplot**

Example: Height and Weight of Children



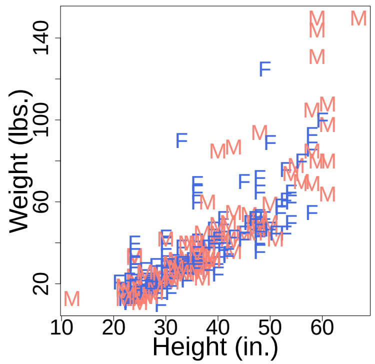
- Shows a clear relationship between Height and Weight
- Here we can depict a third, categorical variable by the plotting color/character

Example: Height and Weight of Children



- What does the relationship look like?
- $BMI = \frac{\text{Weight}(kg)}{(\text{Height}(m))^2}$
- Supposed to give a value that tells you whether you are over- or under-weight, independent of how tall you are.

Example: Height and Weight of Children



- Both scatterplots and QQ Plots place points in 2D space, but they serve very different functions. How are they different?

Example: Height and Weight of Children

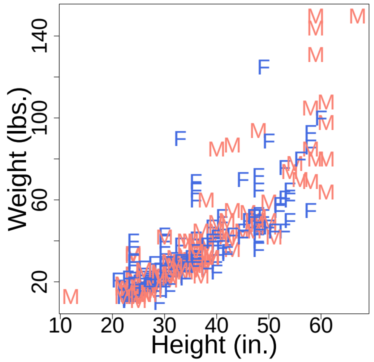


Figure: Scatterplot

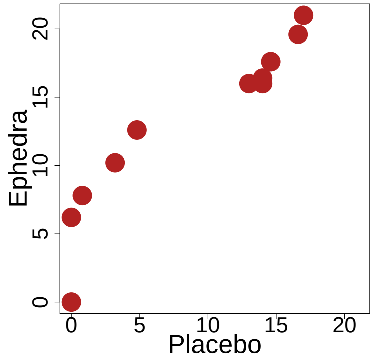
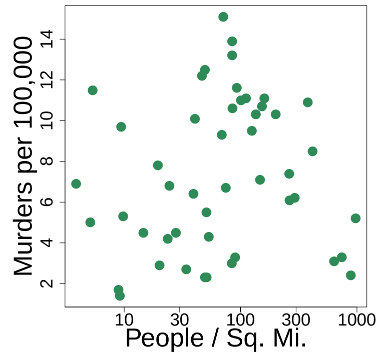


Figure: QQ Plot

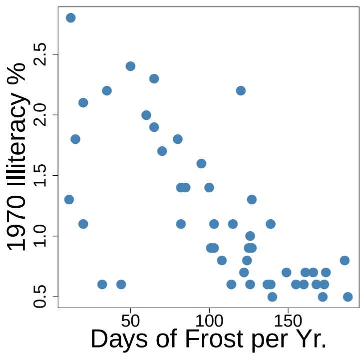
Example: Population Density and Murder

- What do you expect a scatterplot relating a state's population density and its murder rate to look like?



Example: Population Density and Murder

- What do you expect a scatterplot relating a state's frost rate (days of frost per year) and its illiteracy rate (% of population who cannot read) to look like?

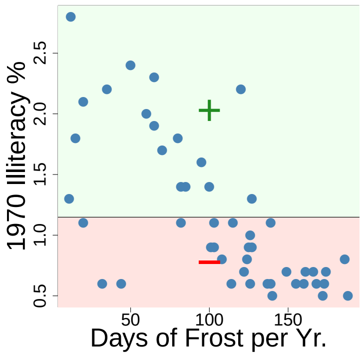


Measuring Relationships

- How can we quantify the relationship between two numeric variables?
- What is correlation?
- The idea: Two variables have a "positive" relationship if one has "high values" at the same time the other is high, and "low values" when the other is low
- If the opposite is true, there is a "negative" relationship.
- What counts as "high" or "low"?

What's High and Low?

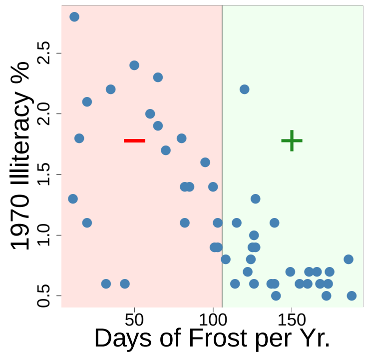
- Intuition: count "above average/center" as "high"; "below average/center" as "low".



ISTA 116: Statistical Foundations for the Information Age
Measuring Association

What's High and Low?

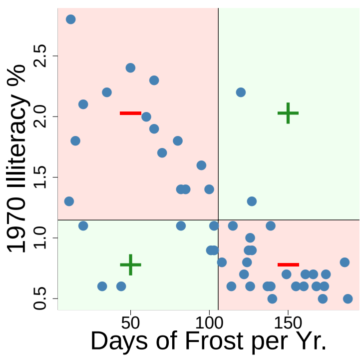
- Intuition: count "above average/center" as "high"; "below average/center" as "low".



ISTA 116: Statistical Foundations for the Information Age
Measuring Association

Positive vs. Negative Association

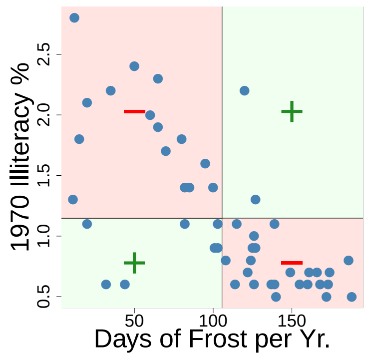
- Then we can say that if a data point is above average (or below average) on both variables at once, it contributes a positive to the relationship.
- If it's above on one and below on the other, it contributes a negative.



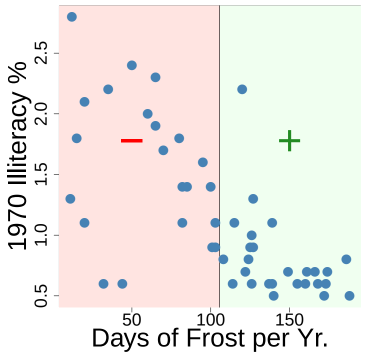
ISTA 116: Statistical Foundations for the Information Age
Measuring Association

Positive vs. Negative Association

- Here, most of the data is in the negative regions; hence, negative relationship.



What measure?



- What have we used to measure whether a data point is above or below center?
- The **deviation scores** are positive for above average values and negative for below average values.

x and y deviations

- Each data point has two coordinates. We can write the i^{th} data point as (x_i, y_i).
- Then its deviation in the x direction is:

$$\text{Deviation}_{x_i} = x_i - \bar{x}$$

- Similarly, the deviation in the y direction:

$$\text{Deviation}_{y_i} = y_i - \bar{y}$$

From Deviations to Association

- What can we do so that we get a positive number when both deviation scores have the same sign, and a negative otherwise?
- The product of the two deviations is called a **cross-product**
- We can sum these up to get a measure of association.

$$\text{Sum of Cross Products} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

But Wait!

- There are (at least) two problems with just reporting the sum of cross products as a standard measure of association. What are they?
  1. Depends on how much data we have
  2. Depends on the choice of units
- Ideally, we want a measure that is independent of both units and sample size.
- Suggestions?

Some Solutions

- There are (at least) two problems with just reporting the sum of cross products as a standard measure of association. What are they?
  1. Depends on the choice of units
     - Solution: Instead of raw deviations, use z-**scores**!
     - The product of z-scores is a **standardized cross product**
  2. Depends on the choice of units
     - Solution: Instead of the sum, use an average standardized cross product.
     - For the same reason as with variance, we use $n-1$ rather than n in the denominator.

Announcements/Reminders

- Midterm coming up on Wednesday, October 12$^{\text{th}}$!
  - Next Monday's lecture will be partly a review/problem session, so come equipped with questions or exercises you want to see worked through.
- See the Schedule pdf for an updated schedule of topics and due dates.
  - Lab 4 pushed back to after the midterm.
  - Web Quiz 5 will take its place; due this Friday.
- See your lab instructor at the end of class to get Quiz 2 back.

Pearson's Correlation Coefficient

- We define **Pearson's Correlation Coefficient** as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

where z_{x_i} represents the z-score for the i^{th} data point in the x variable.

Pearson's Correlation Coefficient

- We define **Pearson's Correlation Coefficient** as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

- Note that since $z_{x_i} = \frac{(x_i - \bar{x})}{s_x}$, and s_x and s_y are the same for every data point, we can factor them out of the sum:

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

Pearson's Correlation Coefficient

- We can go further: since $s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$, we have

$$\begin{aligned} s_x s_y &= \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum(y_i-\bar{y})^2} \\ &= \frac{1}{n-1}\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2)(\sum_{i=1}^{n}(y_i-\bar{y})^2)} \end{aligned}$$

- In the formula for r, the $(n-1)$s cancel to give:

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

Three Equivalent Formulas

- Conceptually:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$

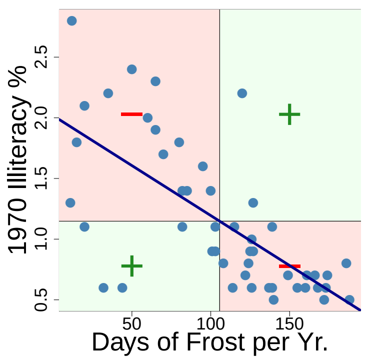
- If you already have standard deviations:

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

- Most efficient to compute from scratch:

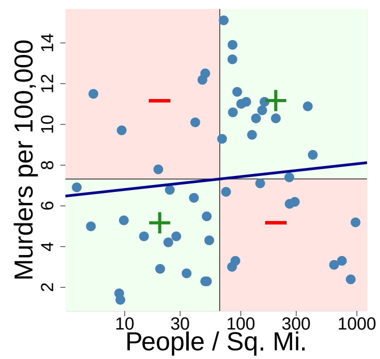
$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i-\bar{x})^2)(\sum_{i=1}^{n}(y_i-\bar{y})^2)}}$$

Linear Association



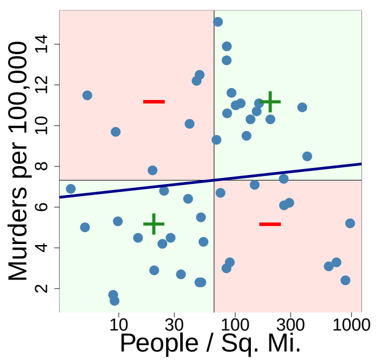
- Pearson's correlation measures how well the data fits a *straight line*.
- Always takes values between -1 and $+1$, with $r = 1$ when the data falls *exactly* on a line with positive slope; $r = -1$ when *exactly* on a line with a negative slope
- Here, $r = -0.68$

Little Association



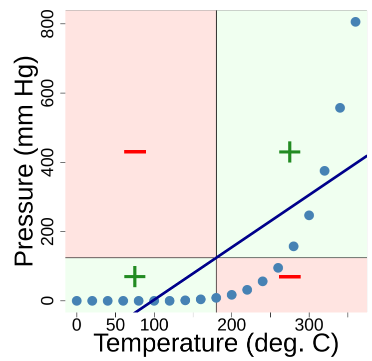
- $r = 0.1$

Little Association



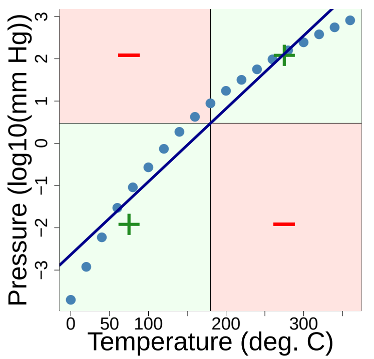
- $r = 0.1$

Nonlinear Association



- $r = 0.76$

With Log Transformation



- $r = 0.97$

What Matters for Pearson's Correlation?

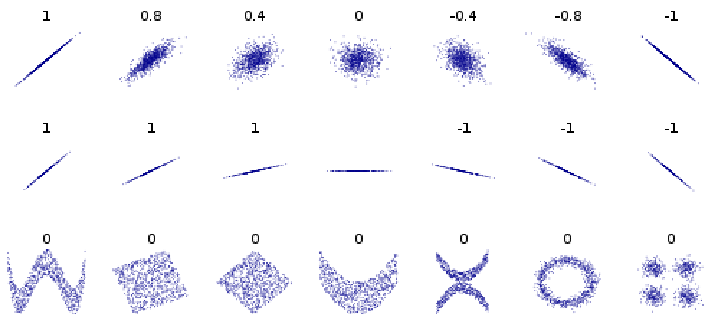


Figure: Hypothetical bivariate data and the corresponding Pearson's correlation coefficient

Correlation $\neq$ Causation

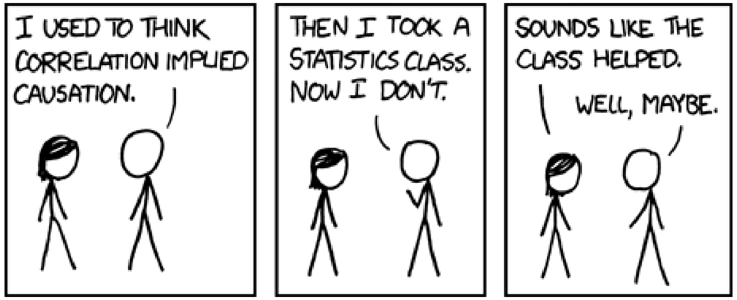


Figure: `xkcd.com/552/`

How can we capture nonlinear associations?

- We'd like a measure that captures association in cases where the data isn't captured by a straight line.
- Example: Pressure and Temperature data lie exactly on an increasing curve. Higher $\Longleftrightarrow$ Higher
- What could we use instead of numerical deviations from the mean?
- **Possibility:** Ordinal position relative to median

Ranks

- Remember order statistics for univariate data?
- With parentheses around the index, denotes the i^{th} *smallest value* in the data set. Called the i^{th} **order statistic**.

$$\begin{aligned} x_{(1)} &= \text{minimum value} \\ x_{(2)} &= \text{next lowest (may be same)} \\ &\ldots \\ x_{(n)} &= \text{maximum value} \end{aligned}$$

- Each data point corresponds to some order statistic. The **rank** of a data point is just the little number in parentheses.

Ranks

The Median

The **median** is written Q_2, and defined as:

$$Q_2 = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ Mean(\{x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}\}) & \text{if } n \text{ is even} \end{cases}$$

- So, when n is odd, the median is a data point, with rank $\frac{n+1}{2}$.
- When n is even, the median is halfway between the data points with ranks $\frac{n}{2}$ and $\frac{n}{2}+1$
- So we can say the median has rank $\frac{n}{2}+\frac{1}{2}$ (same for odd)

Rank Distance from Q_2

- For each data point, its "rank distance" from the median is its rank, minus $\frac{n+1}{2}$.
- For bivariate data, each data point has an x rank distance from the x median, and a y rank distance from the y median.

Spearman's Rank Correlation

- We can define **Spearman's Rank Correlation** in the exact same way as Pearson's, just using the ranks instead of the values:

$$\rho = \frac{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})(\text{Rank}(y_i) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})^2 \sum_{i=1}^{n}(\text{Rank}(y_i) - \frac{n+1}{2})^2}}$$

Spearman's Rank Correlation

- Note that the $\frac{n+1}{2}$ is the rank of the median; it is also the mean of the ranks (i.e., the numbers 1 through n)! Compare:

$$\rho = \frac{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})(\text{Rank}(y_i) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})^2 \sum_{i=1}^{n}(\text{Rank}(y_i) - \frac{n+1}{2})^2}}$$

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

Example: Frost and Illiteracy

- Consider the Frost and Illiteracy example from earlier:

	Frost.	Illit.	Rank(Frost)	Rank(Illit.)
Alabama	20	2.1	4.5	43.0
Arizona	15	1.8	3.0	39.5
Arkansas	65	1.9	11.5	41.0
California	20	1.1	4.5	29.0
Colorado	166	0.7	42.0	16.0
Connecticut	139	1.1	35.5	29.0
Delaware	103	0.9	21.5	23.5
Florida	11	1.3	1.0	32.5
Georgia	60	2.0	10.0	42.0
Idaho	126	0.6	30.5	8.5

Example: Frost and Illiteracy

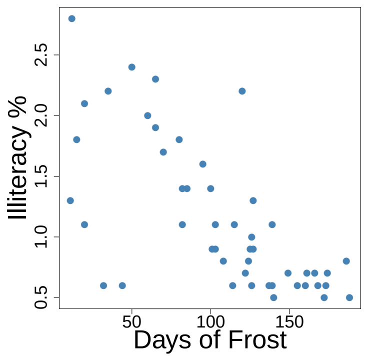


Figure: Numeric Values

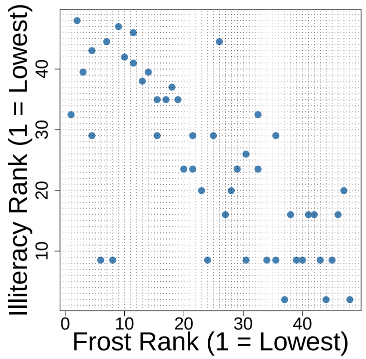


Figure: Rank Data

Spearman's Rank Correlation

- Question: What would the rank plot look like if y always increased when x increased?
- What will the corresponding Spearman's rho be?

Example: Temperature and Pressure

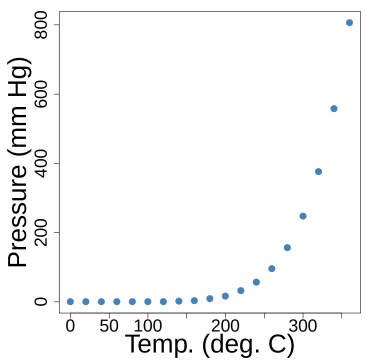


Figure: Numeric Values

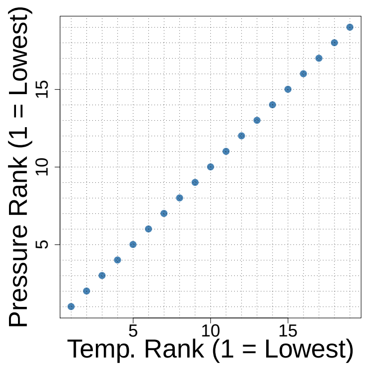


Figure: Rank Data

Example: Temperature and Pressure

- What will happen to Spearman's rho after pressure is transformed to the log scale?

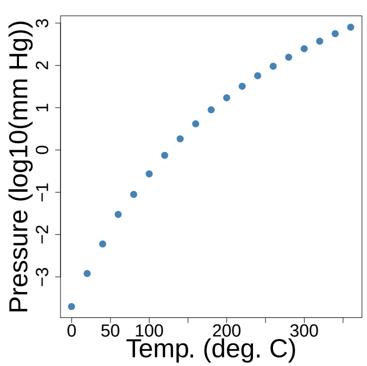


Figure: Numeric Values

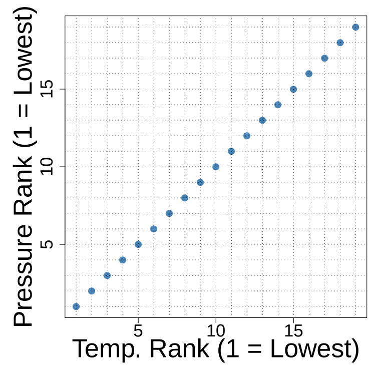


Figure: Rank Data

Correlation $\neq$ Causation

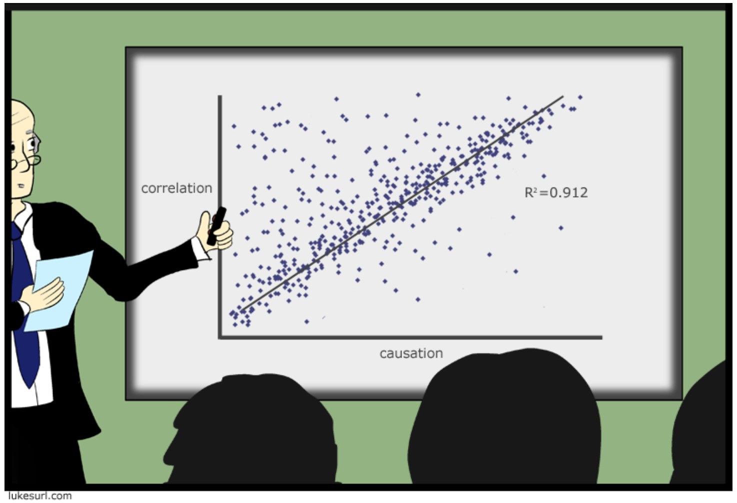


Figure: http://www.lukesurl.com/