Stat 201:
Introduction to Statistics

Standard 10: Variable Association – Scatter Plots & Correlation Coefficient

From *Naked Statistics: Correlation*

- “Correlation measures the degree to which two phenomena are related to one another… The correlation coefficient has two fabulously attractive characteristics. First it is a single number ranging from -1 to 1”
- “A correlation of 1, often described as perfect correlation, means that every change in one variable is associated with a (constant) change in the other variable in the same direction.”
- “A correlation of -1, or perfect negative correlation means that every change in one variable is associated with a (constant) change in the other variable in the opposite direction.”
- “A correlation of 0 (or close to it) means that the variables have no meaningful association with one another.”

From *Naked Statistics: Correlation*

- "Two variables are positively correlated if a change in one is associated with a change in the other in the same direction, such as the relationship between height and weight. Taller people weigh more (on average.)… A correlation is negative if a positive change in one variable is associated with a negative change in the other, such as the relationship between exercise and weight."

From *Naked Statistics: Correlation*

- “The correlation coefficient does a seemingly miraculous thing. It collapses a complex mess of data measured in different units into a single, elegant descriptive statistic (which is unit-free)”

Association of Variables – Two Quantitative Variables

- **Response Variable –** this is our dependent variable, the outcome variable on which comparisons are made
- **Explanatory Variable –** this is our independent variable, the groups to be compared with respect to values on the response variable
- **Think "we use the explanatory variable to EXPLAIN what's going on with the response variable."**

More Definitions

- An **correlation** exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable

Response Variable	this is our dependent variable, the outcome variable on which comparisons are made
Explanatory Variable	this is our independent variable, the groups to be compared with respect to values on the response variable
Correlation	exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable

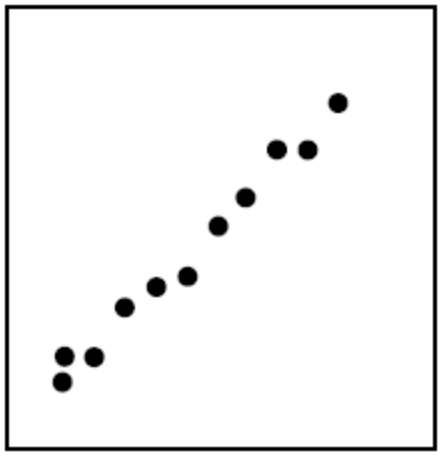
What About Two Quantitative Variables?

- We can **summarize** each quantitative variable with the following things we have learned
  - **Graphics:** Boxplots, Histograms, Dot plots
    - Look at the shape of the distribution (CH 2)
      - Is it skewed?
  - **Numerical Summaries:** Mean, median, mode, standard deviation, variance, IQR, etc
    - Compare the averages and spread of the data
      - Mean and Standard Deviation

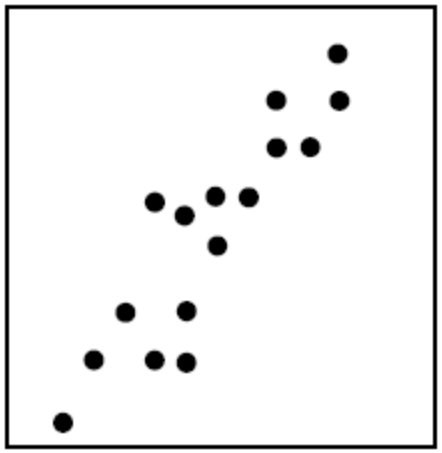
Scatterplots

- We can **compare <u>two quantitative variables</u>** and explore their association or correlation with a **scatterplot**

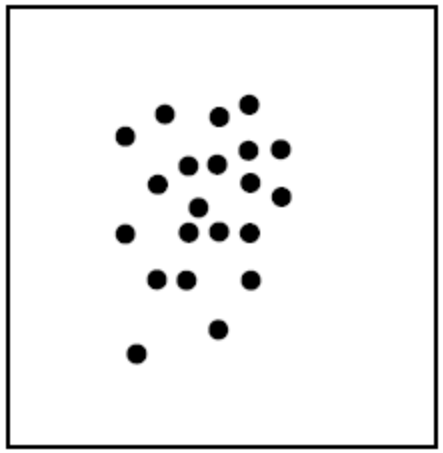
- To form a **scatterplot** we let the **response** variable be the y variable and the **explanatory** variable be the x variable and plot the points



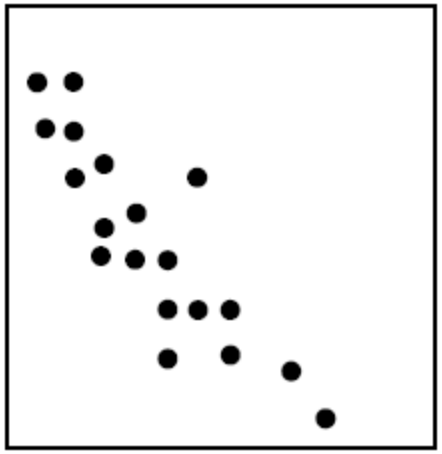
Strong positive correlation



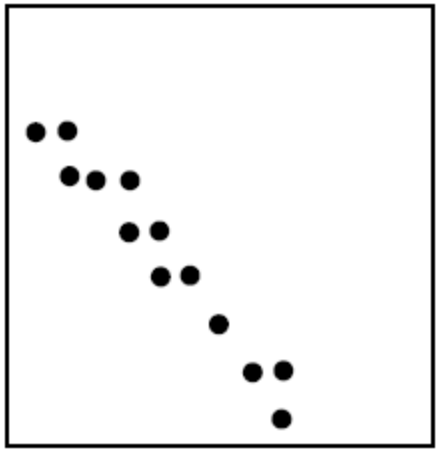
Moderate positive correlation



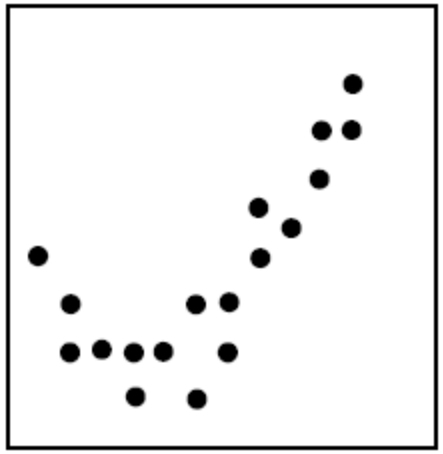
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

Coefficient of Correlation (r)

- **r** measures the **LINEAR** relationship between x and y [linear, linear, linear, linear!!!]
- **r > 0** → positive correlation or association
- **r < 0** → negative correlation or association
- **r=1** → perfect positive correlation or association
  - Here, all points would fit on a line
- **r=-1** → perfect negative correlation
  - Here, all points would fit on a line
- **r=0** → no correlation

Properties

- $-1 \leq r \leq 1$
- The closer r is to 1 the stronger the evidence for positive association
- The closer r is to -1 the stronger the evidence for a negative association
- The closer r is to 0 the weaker the evidence for association
- **Affected by outliers so we have to be careful**

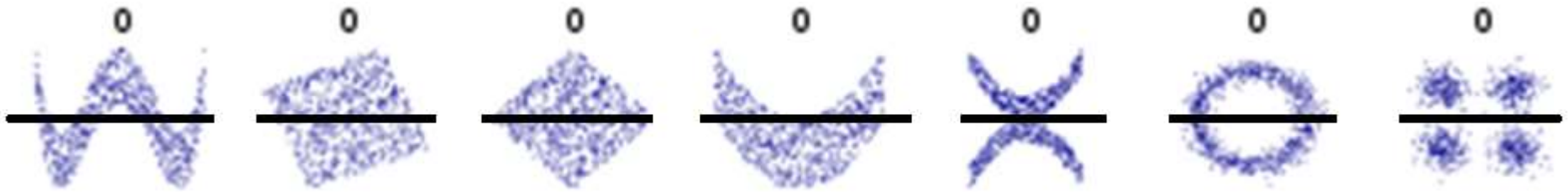
Coefficient of Correlation Examples

- The perfect lines have r=1 or r=-1 **depending on the sign of their slope <u>not</u> the magnitude of the slope**
- You might think that the plot in the middle has r=1 too because it too is fit perfectly by a line. **The catch** is that the line would be horizontal, thus having a slope value of zero (no sign.) These dots actually show that the explanatory variable provides no explanation for the response variable.

Coefficient of Correlation Examples

- Again, the perfect lines have r=1 or r=-1
- The points that don't make perfect lines have decimal values depending on how close they are to a perfect line.
  - The closer r is to 1 the stronger the evidence for positive association
  - The closer r is to -1 the stronger the evidence for a negative association
  - The closer r is to 0 the weaker the evidence for association
- **Note: their value changes based on how close they are to forming a line <u>not</u> the magnitude of the slope**

Coefficient of Correlation Examples



- Here there are obvious patterns here **but** they are not linear!
- Since, r, measures the **linear** relationship between two variables r=0 even though there are patterns!
  - In each case the best balanced line would be horizontal

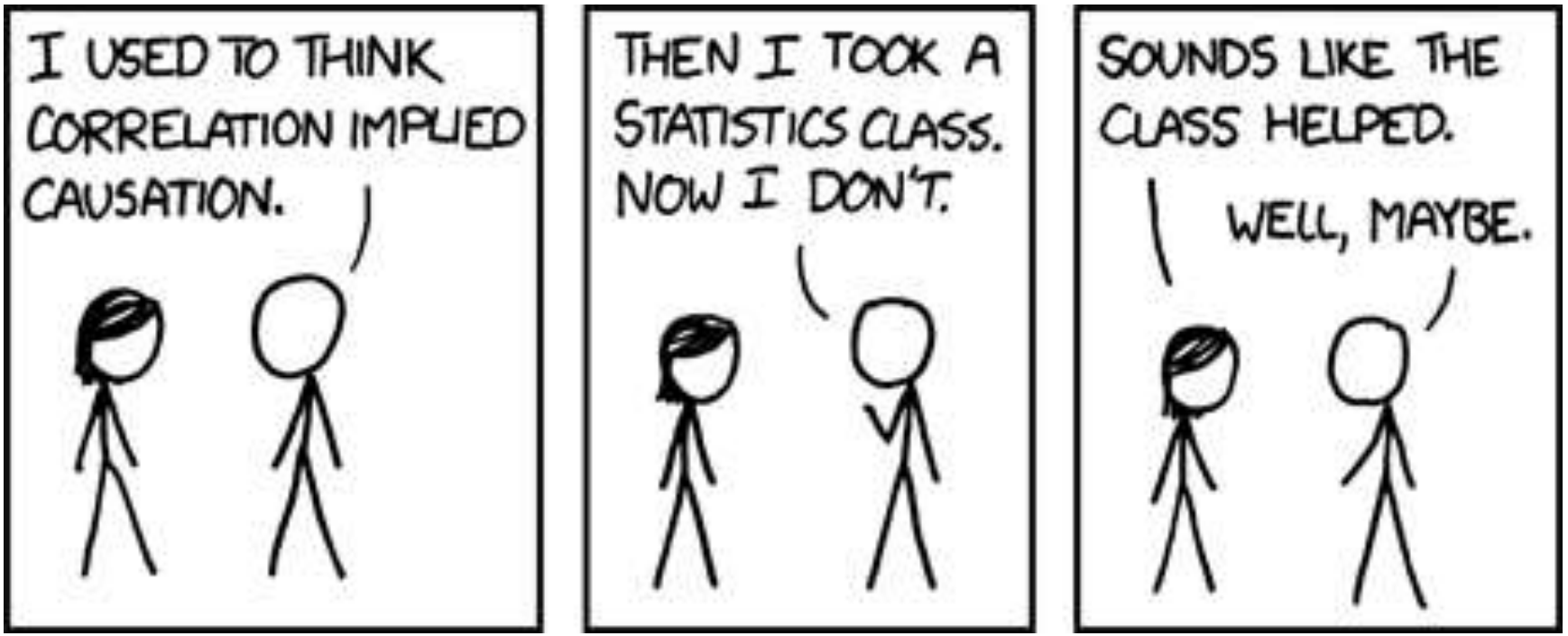
Correlation vs. Causation

- The idea here is that although some variables are correlated they one might not be the cause of the other.

- Let's revisit http://www.tylervigen.com/?categoria=%22dinero%22

Regressions – Problems

- Correlation does not imply causation
  - We go from saying there exists a correlation to saying that one variable's change causes the other to change.



Credit: XKCD