Lecture 9

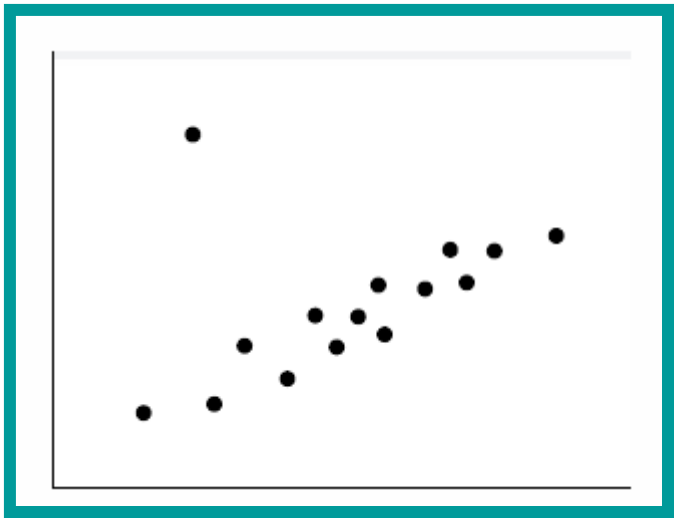
Relationships can be Deceiving

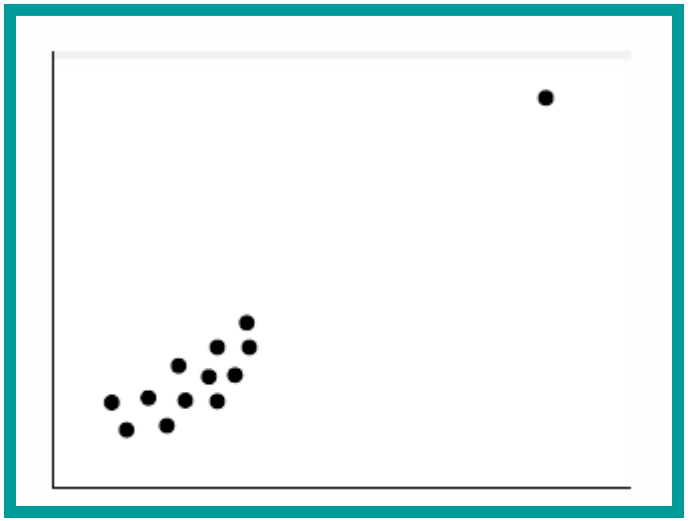
Thought Question 1:

Use following two pictures to speculate on what influence outliers have on correlation.

For each picture, do you think the ***correlation is higher or lower than it would be without the outlier***?

(*Hint:* Correlation measures how closely points fall to a straight line.)





Thought Question 2:

A strong correlation has been found in a certain city in the northeastern United States between weekly sales of hot chocolate and weekly sales of facial tissues.

Would you interpret that to mean that ***hot chocolate causes people to need facial tissues***? Explain.

Thought Question 3:

Researchers have shown that there is a positive correlation between the average fat intake and the breast cancer rate across countries. In other words, countries with higher fat intake tend to have higher breast cancer rates.

Does this ***correlation prove that dietary fat is a contributing cause of breast cancer***? Explain.

Thought Question 4:

If you were to draw a scatterplot of ***number of women in the work force*** versus ***number of Christmas trees sold*** in the United States for each year between 1930 and the present, you would find a very strong correlation.

Why do you think this would be true?

Does one cause the other?

11.1 Illegitimate Correlations

Problems with Correlations

- Outliers can substantially inflate or deflate correlations.
- Groups combined inappropriately may mask relationships.

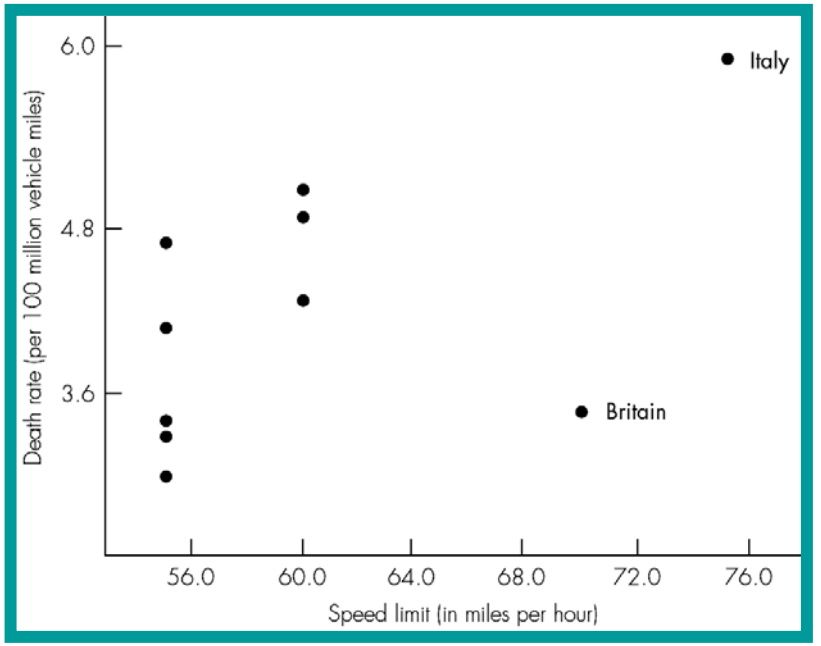
The Impact Outliers Have on Correlation

- An outlier that is **consistent** with the trend of the rest of the data will **inflate** the correlation.
- An outlier that is **not consistent** with the rest of the data can substantially **decrease** the correlation.

At least 5% of all data points are corrupted (when initially recorded or when entered into computer). Important to ***perform checks*** of the data.

Example 1: Highway Deaths and Speed Limits

COUNTRY	Death Rate (Per 100 Million Vehicle Miles)	Speed Limit (in Miles Per Hour)
Norway	3.0	55
United States	3.3	55
Finland	3.4	55
Britain	3.5	70
Denmark	4.1	55
Canada	4.3	60
Japan	4.7	55
Australia	4.9	60
Netherlands	5.1	60
Italy	6.1	75



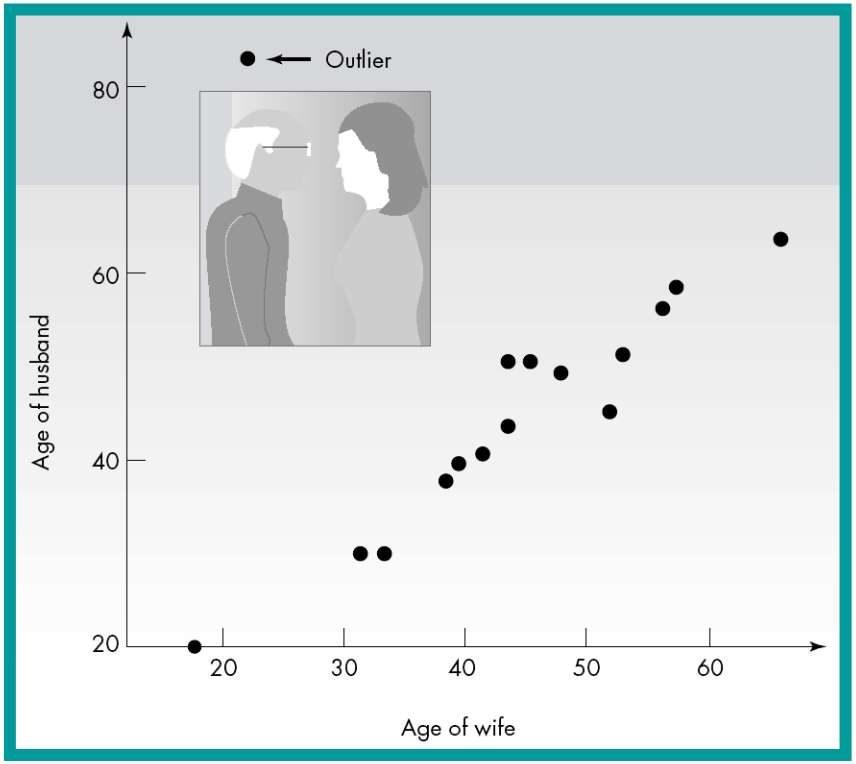
- Correlation between death rate and speed limit is 0.55.
- If Italy removed, correlation drops to 0.098.
- If then Britain removed, correlation jumps to 0.70.

Source: Rivkin, 1986.

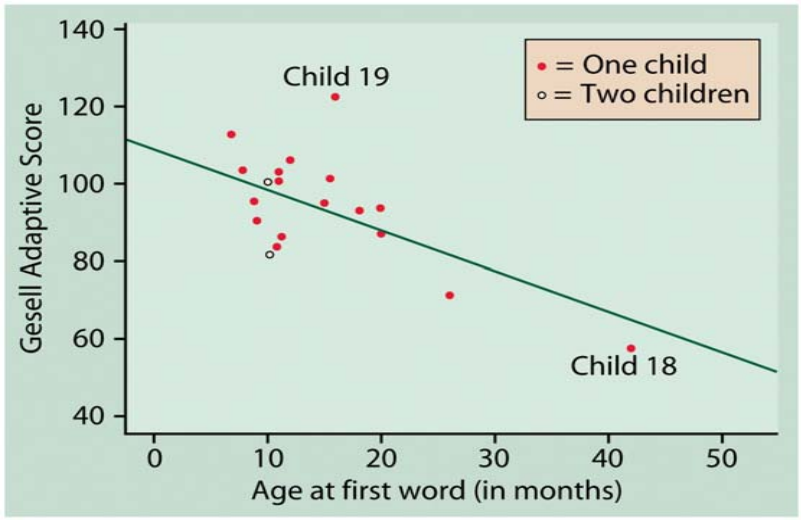
Example 2: Ages of Husbands and Wifes, Revisited

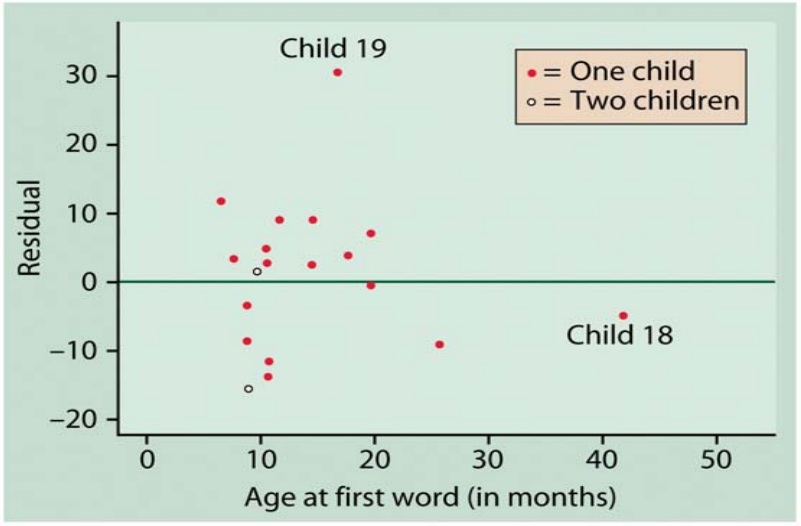
Subset of data on ages of husbands and wives, with one outlier added (entered 82 instead of 28 for husband's age).

- Correlation for data with outlier is **.39**.
- If outlier removed, correlation of remainder points is **.964** – a very strong linear relationship.



Unusual points in a scatter plot





Web site for practice

Legitimate Outliers, Illegitimate Correlation

Outliers can be legitimate data –
like Italy and Britain in Example 1.

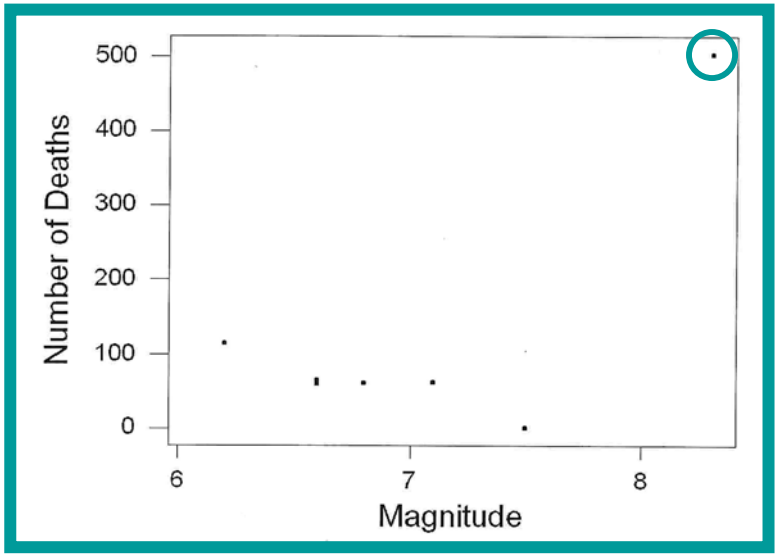
Be careful …

- when presented with data in which outliers are likely to occur
- When correlations presented for a small sample.

Example 3: Earthquakes in Continental U.S.

Date	Location	Deaths	Magnitude
August 31, 1886	Charleston, SC	60	6.6
April 18 – 19, 1906	San Francisco	503	8.3
March 10, 1933	Long Beach, CA	115	6.2
February 9, 1971	San Fernando Valley, CA	65	6.6
October 17, 1989	San Francisco area	62	6.9
June 28, 1992	Yucca Valley, CA	1	7.4
January 17, 1994	Northridge, CA	61	6.68

- Correlation is **.689**, relatively strong positive association.
- If SF earthquake of 1906 removed, correlation is **–.92**; higher magnitude quakes associated with fewer deaths.



The 1906 earthquake was before earthquake building codes enforced. Next largest quake in 1992 occurred in sparsely populated area.

Source: World Almanac and Book of Facts online, Nov 2003.

The Missing Link: A Third Variable

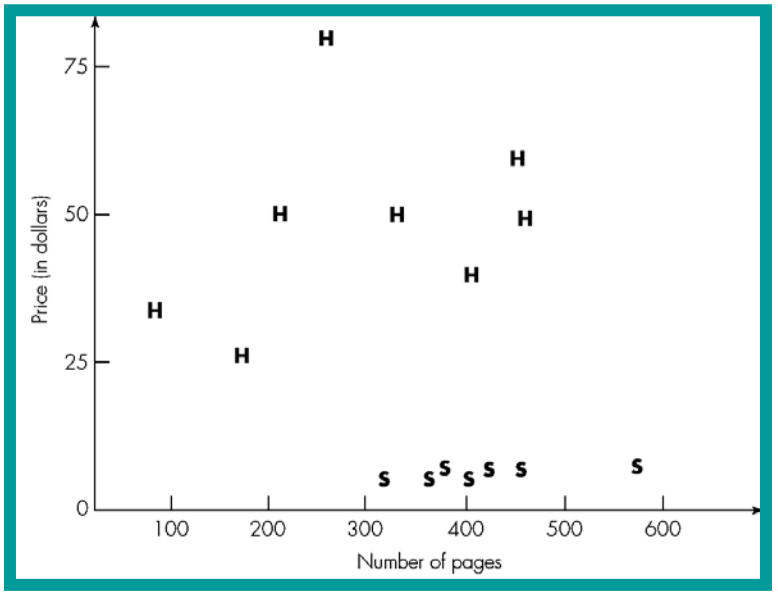
Simpson's Paradox:

- Two or more groups.
- Variables for each group may be strongly correlated.
- When groups combined into one, very little correlation between the two variables.

Example 4: The Fewer The Pages, The More Valuable The Book?

Pages	Price	Pages	Price	Pages	Price
104	32.95	342	49.95	436	5.95
188	24.95	378	4.95	458	60.00
220	49.95	385	5.99	466	49.95
264	79.95	417	4.95	469	5.99
336	4.50	417	39.75	585	5.95

Pages versus Price for the Books on a Professor's Shelf



- Correlation is **–.312**, more pages => less cost?
- Scatterplot includes book type: **H = hardcover, S = softcover**.
- Correlation for H books: **.64**
- Correlation for S books: **.35**

Combining two types masked the positive correlation and produced illogical negative association.

11.2 Legitimate Correlation Does Not Imply Causation

A Silly Correlation:

List of weekly tissue sales and weekly hot chocolate sales for a city with extreme seasons would probably exhibit a correlation because both tend to go up in the winter and down in the summer.

Data from an **observational study**,
in the absence of any other evidence, simply
cannot be used to establish causation.

Causation vs. Association

- Some studies want to find the existence of *causation*.
- Example of causation:
  - Increased drinking of alcohol causes a decrease in coordination.
  - Smoking and Lung Cancer.
- Example of association:
  - High SAT scores are associated with a high Freshman year GPA.
  - Smoking and Lung Cancer.

Correlation

Variable A could cause variable B.

(Eating ice cream could cause cramps, which could lead to drowning)

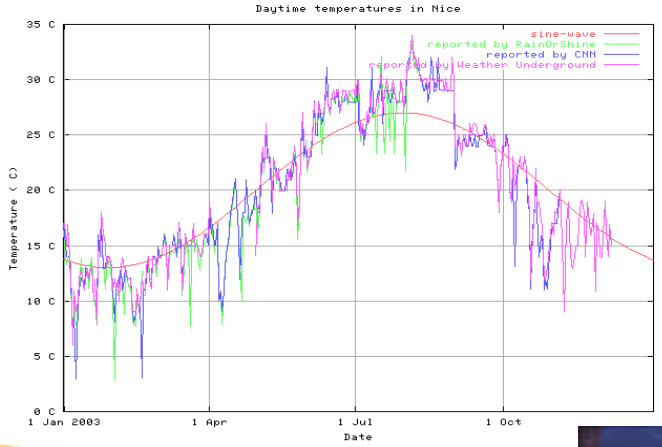
Correlation

Variable B could cause variable A.

Parents could buy kids more ice cream to console them after their friends drown

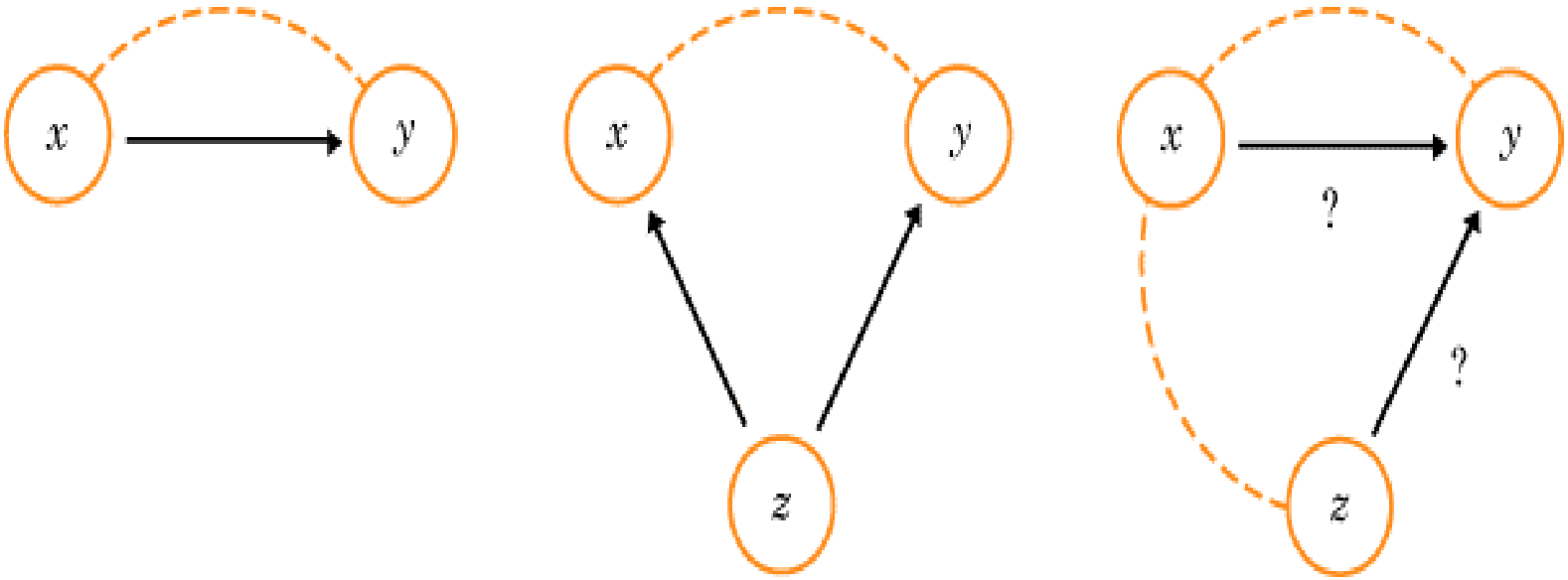
Correlation

Or a third variable could cause A and B.



Heat leads people to eat ice cream and to swim, but the two aren't directly

Explaining Association



(a) Causation (b) Common response (c) Confounding

Explaining Association: Causation

- Cause-and-effect
- Examples
  - Amount of fertilizer and yield of corn
  - Weight of a car and its MPG
  - Dosage of a drug and the survival rate of the mice

Explaining Association: Common Response

- Lurking variables
- Both x and y change in response to changes in z, the lurking variable
- There may not be direct causal link between x and y.
- Lurking variables can create nonsense correlations!

Lurking Variable

- A lurking (hidden) variable is a variable that has an important effect on the relationship among the variables in a study, but is not included among the variables being studied.
- Examples:
  - SAT scores and college grades
    - Lurking variable: IQ
  - Fireman and the total damage
    - Lurking variable:
  - Number of churches and number of bars
    - Lurking variable:

How to spot the presence of lurking variables?

- In general difficult.
- Many lurking variables change systematically over time.
  - Plot both the response variable and the residuals against the time order of the observations whenever possible.

Example 5: Happiness and Heart Disease

News Story #4: "Heart patients who are happy are much more likely to be alive 10 years down the road than unhappy heart patients."

An observational study – possible explanations for observed relationship between happiness and risk of death …

> *"The experience of joy seems to be a factor. It has physical consequences and also attracts other people, making it easier for the patient to receive emotional support. Unhappy people, besides suffering from the biochemical effects of their sour moods, are also less likely to take their medicines, eat healthy, or to exercise. (p. 9)"*

Taking one's medicine, exercising are **confounded** with the **explanatory variable**, ***mood***, in determining its relationship with the **response variable**, ***length of life***.

Example 6: Prostate Cancer and Red Meat

Study Details:

- Followed 48,000 men who filled out dietary questionnaires in 1986.
- By 1990, 300 men diagnosed with prostate cancer and 126 had advanced cases.
- For advanced cases: “men who ate the most red meat had a 164% higher risk than those with the lowest intake.”

Possible third variable that both leads men to consume more red meat and increases risk of prostate cancer … ***the hormone testosterone***.

11.3 Some Reasons for Relationships Between Variables

1. Explanatory variable is the direct cause of the response variable.
2. Response variable is causing a change in the explanatory variable.
3. Explanatory variable is a contributing but not sole cause of the response variable.
4. Confounding variables may exist.
5. Both variables may result from a common cause.
6. Both variables are changing over time.
7. Association may be nothing more than coincidence.

Reasons Two Variables Could Be Related:

1. **Explanatory variable is the direct cause of the response variable.**

e.g. Amount of food consumed in past hour and level of hunger.

2. **Response variable is causing a change in the explanatory variable.**

e.g. Explanatory = advertising expenditures and Response = occupancy rates for hotels.

3. **Explanatory variable is a contributing but not sole cause of the response variable.**

e.g. Carcinogen in diet is not sole cause of cancer, but rather a *necessary contributor* to it.

Example 7: Delivery Complications, Rejection, and Violent Crimes

Study Details:

- Observational study of males born in Copenhagen, Denmark, between 1956 and 1961.
- Delivery complications at birth associated with higher incidence of violent crimes later in life.
- Connection only held for men whose mothers rejected them (mothers did not want pregnancy, tried to abort fetus, and sent baby to institution).

Interaction of delivery complications and maternal rejection associated with higher levels of violent crime.

Source: Science (Mann, March 1994)

Reasons Two Variables Could Be Related:

4. **Confounding variables may exist.**

   A confounding variable is related to the explanatory variable and affects the response variable. So can't determine how much change is due to the explanatory and how much is due to the confounding variable(s).

e.g. Emotional support is a confounding variable for the relationship between happiness and length of life in Example 5.

5. **Both variables may result from a common cause.**

e.g. Verbal SAT and GPA: causes responsible for one variable being high (or low) are same as those responsible for the other being high (or low).

Example 8: Do Smarter Parents Have Heavier Babies?

Study Details:

- Observational study of about 3900 babies born in Britain in 1946.
- Relationship between birth weight and intelligence in childhood/early adulthood.
- Genetically smarter parents likely to have smarter offspring.
- Smarter parents more likely to have better diet and less alcohol consumption, which also contribute to birth weight.

Heavier birth weight and higher intelligence in child could both result from **common cause of parents' intelligence**.

Source: News Story #18

Reasons Two Variables Could Be Related:

6. Both variables are changing over time.

Nonsensical associations result from correlating two variables that have both changed over time.

Example 9: Divorce Rates and Drug Offenses

Correlation between divorce rate and % admitted for drug offenses is 0.67, quite strong. However, both simply reflect a trend across time.

Year	Divorce Rate (per 1000)	Percentage Admitted for Drug Offenses
1960	2.2	4.2
1964	2.4	4.1
1970	3.5	9.8
1974	4.6	12.0
1978	5.2	8.4
1982	5.1	8.1
1986	4.8	16.3

Sources: Information Please Almanac, 1991, p. 809.
World Almanac and Book of Fact, 1993, p. 950.

- Correlation between year and divorce rate is 0.92.
- Correlation between year and % admitted for drug offenses is 0.78.

Reasons Two Variables Could Be Related:

7. **Association may be nothing more than coincidence.**
   Association is a coincidence, even though odds of it happening appear to be very small.

Example:

New office building opened and within a year there was an unusually **high rate of brain cancer** among workers in the building. Suppose **odds** of having that many cases in one building were only **1 in 10,000**. **But there are thousands of new office buildings**, so we should expect to see this phenomenon just by chance in about 1 of every 10,000 buildings.

11.4 Confirming Causation

The only legitimate way to try to establish a causal connection statistically is ***through the use of randomized experiments.*** If a randomized experiment cannot be done, then **nonstatistical considerations** must be used to determine whether a causal link is reasonable.

Evidence of a possible causal connection:

- There is a reasonable explanation of cause and effect.
- The connection happens under varying conditions.
- Potential confounding variables are ruled out.