STAT011 Statistical Methods I

Lecture 7 Association and Causation

Lu Chen
Swarthmore College
2/12/2019

Review

- Assessing least squares regression line
  - Coefficient of determination $r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$
  - Residual plot
  - Transformation
- Relationship between two categorical variables
  - Two-way tables `table(Response, Explanatory)`
  - Bar plot `barplot()`
  - Interpreting two-way tables
    - Joint distribution `prop.table(two-way table)`
    - Marginal distribution `prop.table(table of each variable)`
    - Conditional distribution `prop.table(two-way table, margin = 2)`

Outline

- Relationship between a quantitative variable and a categorical variable
  - Summary statistics
  - Boxplot
- Association and causation
  - Examples of relationships
    - Simpson's paradox
- Lurking variable
- Types of associations

Relationship: Quantitative vs. Categorical

- Response: quantitative variable
- Explanatory: categorical variable

Is height of male students the same as height of female students?

- Response: Height
- Explanatory: Gender (male, female and other)

Do higher class year students drink more coffee?

- Response: Cups of coffee
- Explanatory: Class year (Fr, So, Jr and Sr)

Relationship: Quantitative vs. Categorical

```
# By default, aggregate() ignores NA values in the variables
aggregate(Height ~ Gender, data=Survey, FUN = summary)
```

```
##   Gender Height.Min. Height.1st Qu. Height.Median Height.Mean Height.3rd Qu.
## 1 Female       54.00          63.00         65.00       64.92          67.00
## 2   Male       63.00          68.38         70.00       70.16          73.00
## 3  Other       63.00          63.00         63.00       63.00          63.00
##   Height.Max.
## 1       72.00
## 2       78.50
## 3       63.00
```

```
aggregate(Height ~ Gender, data=Survey, FUN = sd)
```

```
##   Gender   Height
## 1 Female 3.401240
## 2   Male 3.622805
## 3  Other       NA
```

Relationship: Quantitative vs. Categorical

```r
aggregate(Coffee ~ ClassYear, data=Survey, FUN = summary)
```

```
##   ClassYear Coffee.Min. Coffee.1st Qu. Coffee.Median Coffee.Mean
## 1        Fr       0.000          0.000         0.000       1.841
## 2        So       0.000          0.000         0.000       4.000
## 3        Jr       0.000          0.000         1.000       2.769
## 4        Sr       0.000          0.000         2.000       3.143
##   Coffee.3rd Qu. Coffee.Max.
## 1          3.000      14.000
## 2          5.000      21.000
## 3          3.000      18.000
## 4          2.500      15.000
```

```r
aggregate(Coffee ~ ClassYear, data=Survey,
          FUN = function(x) c(SD = sd(x), IQR = IQR(x)))
```

```
##   ClassYear Coffee.SD Coffee.IQR
## 1        Fr  2.995730   3.000000
## 2        So  6.147009   5.000000
## 3        Jr  4.867474   3.000000
## 4        Sr  5.367450   2.500000
```

Relationship: Quantitative vs. Categorical

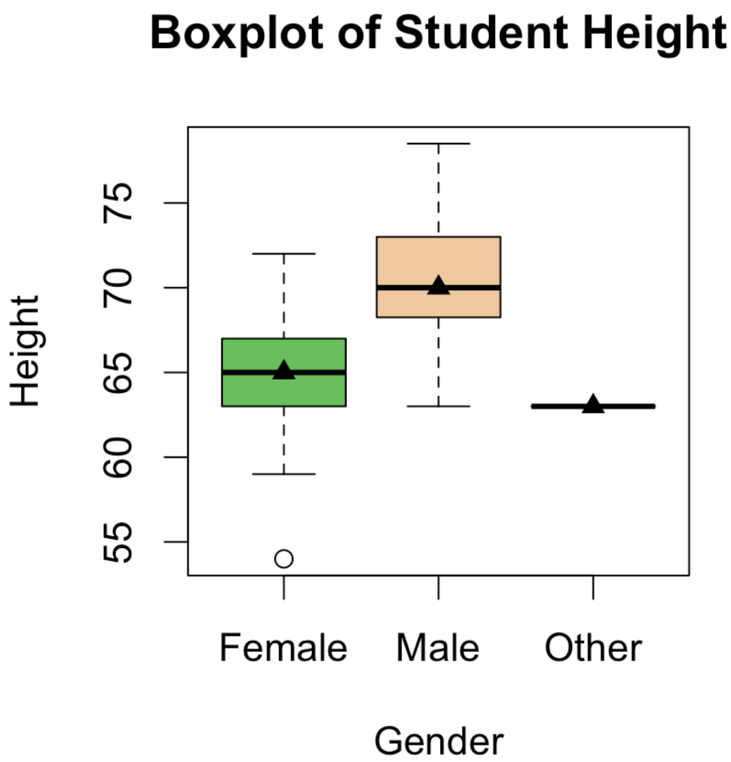
Height	Female	Male	Other
Mean	65.0	70.0	63.0
SD	3.4	3.6	NA

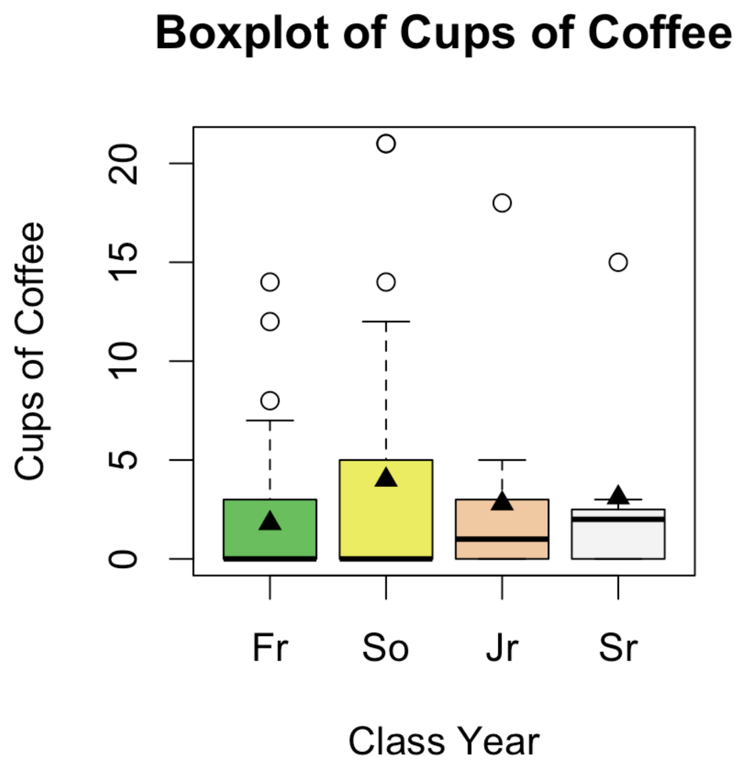
Coffee	Fr	So	Jr	Sr
Median	0.0	0.0	1.0	2.0
IQR	3.0	5.0	3.0	2.5

Why do we report mean and SD for *Height* but median and IQR for *Coffee*?

Relationship: Quantitative vs. Categorical

```
boxplot(Response ~ Explanatory, data= , col= , main= , xlab= , ylab= )
points(c(mean1, mean2, mean3, ...), pch= , col= )
```





Relationships between two variables

Exploratory Data Analysis		**No Explanatory**	**Explanatory**	
			Categorical	**Quantitative**
Response	**Categorical**	• Table of counts and proportions • Bar plot • Pie chart *(Lecture 2)*	• Two-way tables - Joint distribution - Marginal distribution - Conditional distribution • Bar plot *(Lecture 6)*	——
	Quantitative	• Mean, SD • Median, IQR • Histogram, density curve • Boxplot *(Lecture 2~4)*	• Table of summary statistics • Histogram, density curve • Boxplot *(Lecture 7)*	• Correlation • Regression • Scatterplot (Lecture 5~6)

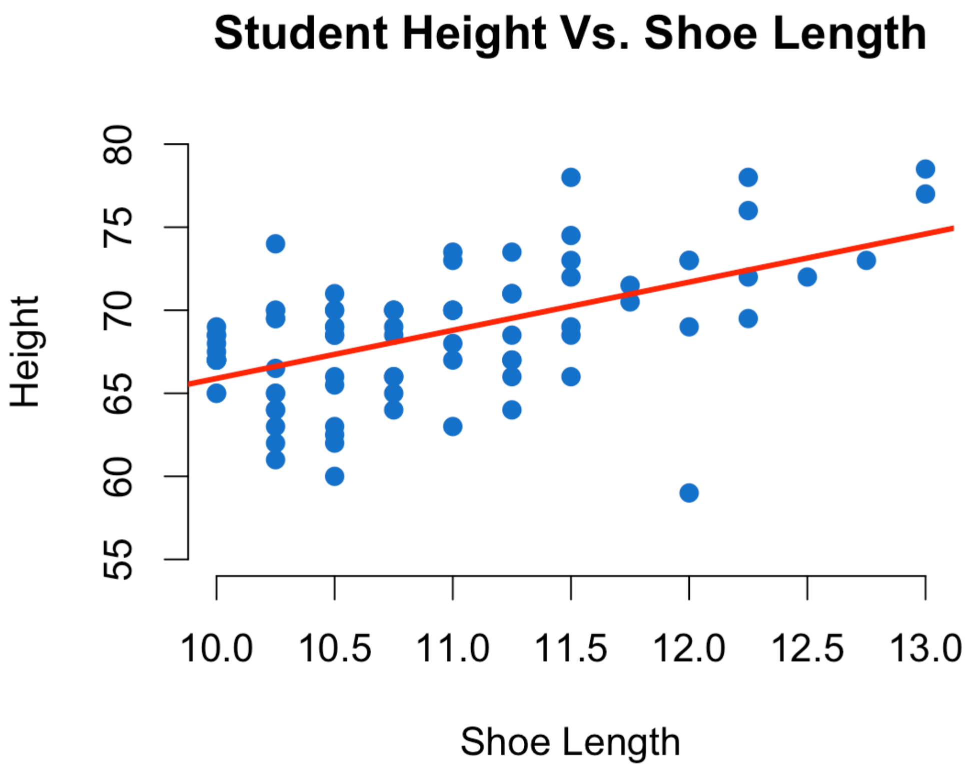
Association and causation

> Two variables measured on the same observations are **associated** if knowing the values of one of the variables tells you something about the values of the other variable.

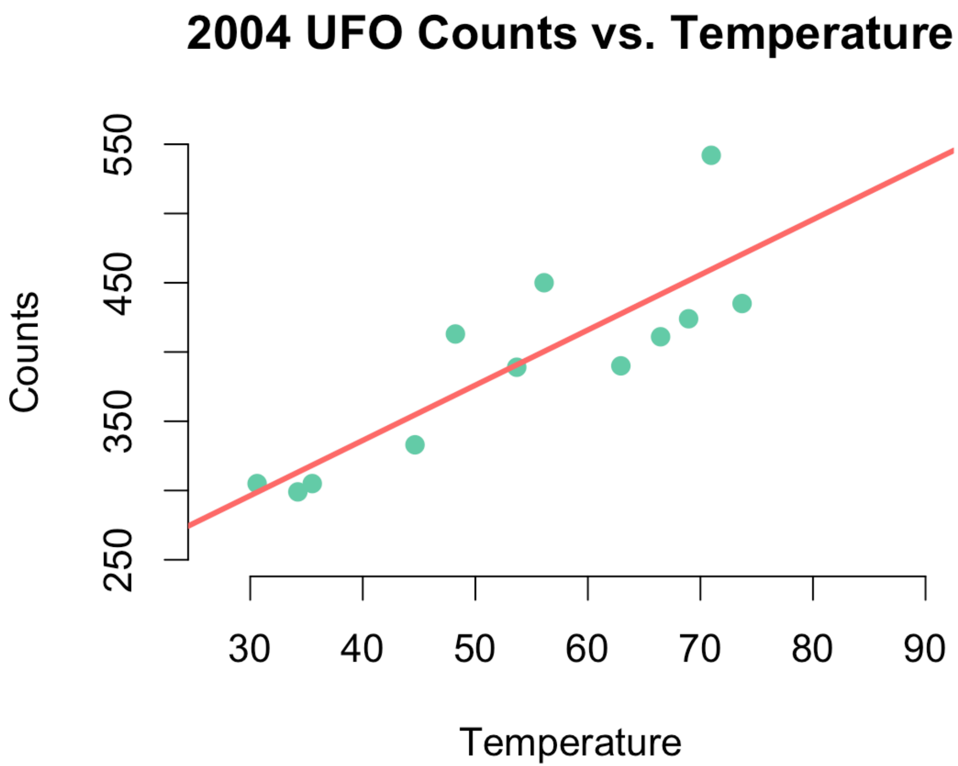
However,

> An association between an explanatory variable X and a response variable Y, even if it is very strong, is not by itself good evidence that changes in X actually **cause** changes in Y.

Example 1


Student Height Vs. Shoe Length
Height
55
60
65
70
75
80
10.0
10.5
11.0
11.5
12.0
12.5
13.0
Shoe Length

Example 2



2004 UFO Counts vs. Temperature
Counts
250
350
450
550
30
40
50
60
70
80
90
Temperature

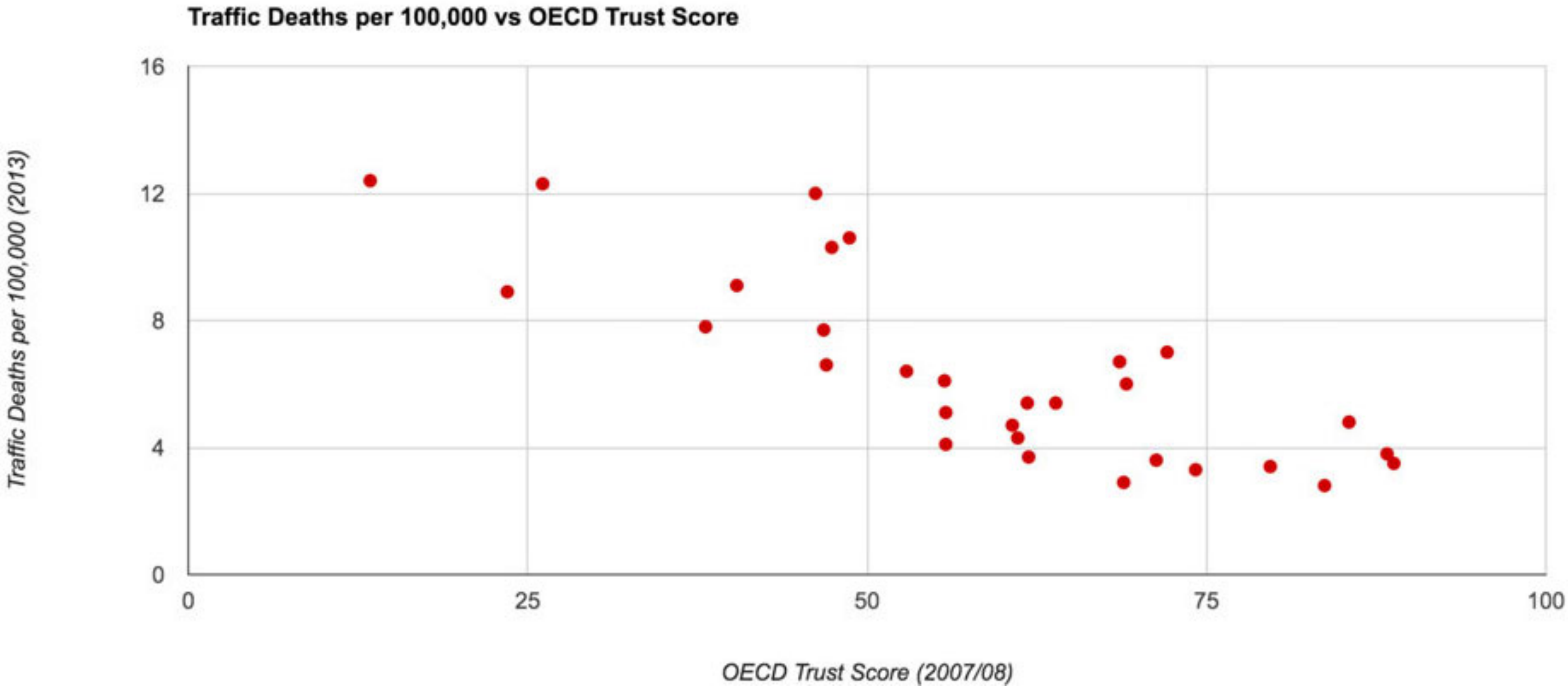
Example 3

The Less You Trust Your Government The More Likely You Are To Die In A Car Wreck

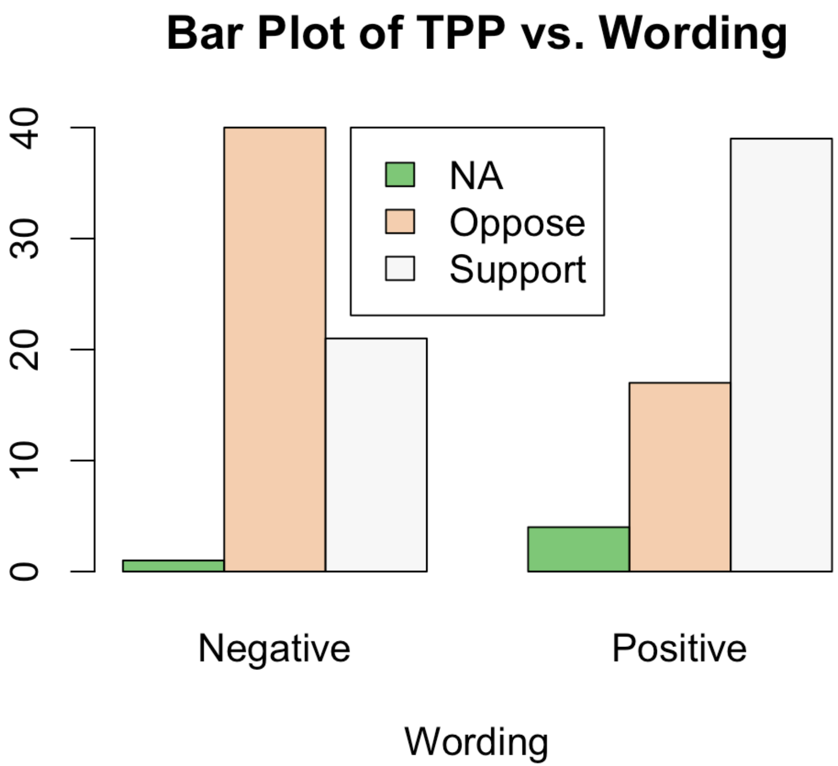
Jason Torchinsky
4/21/16 1:20pm · Filed to: CAR CULTURE

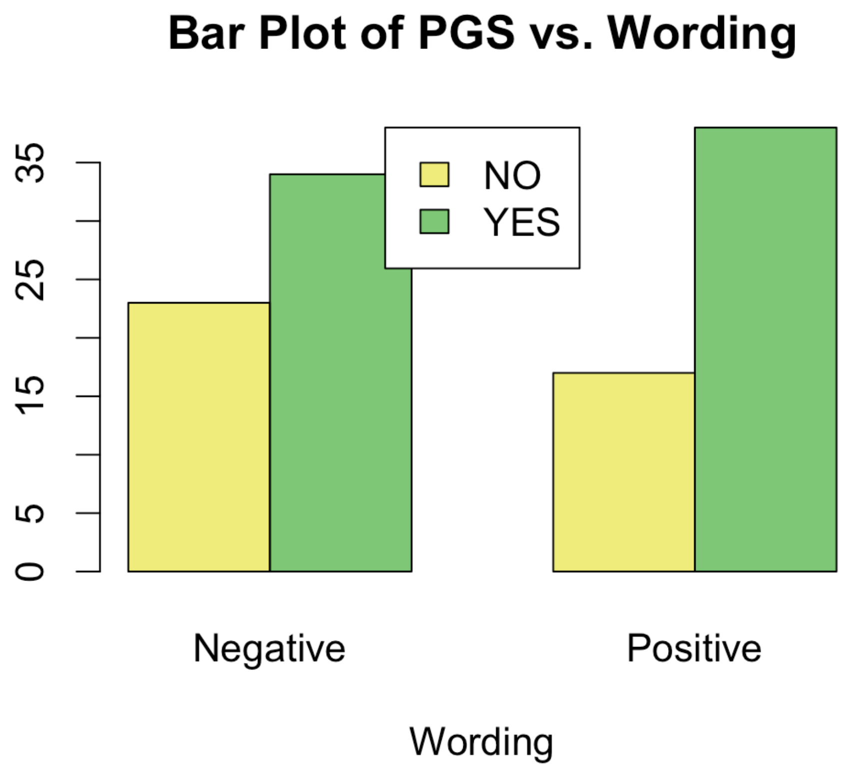
15.9K 121 6

Example 3


Traffic Deaths per 100,000 vs OECD Trust Score
16
12
8
4
0
0
25
50
75
100
Traffic Deaths per 100,000 (2013)
OECD Trust Score (2007/08)

Example 4


Bar Plot of TPP vs. Wording
0
10
20
30
40
NA
Oppose
Support
Negative
Positive
Wording


Bar Plot of PGS vs. Wording
0
5
15
25
35
NO
YES
Negative
Positive
Wording

Example 5 - Surgery vs. Treatment

Compare the success rates of two treatments for kidney stones

	Treatment A	Treatment B	Total
Succeed	273	289	562
Fail	77	61	138
Total	350	350	700

- The success rate of Treatment A is $\frac{273}{350} = 78\%$.
- The success rate of Treatment B is $\frac{289}{350} = 83\%$.
- Treatment B seems better.
- Note: the conclusion is based on the **conditional distribution** of the reponse variable given the explanatory variable levels.

Example 5 - Surgery vs. Treatment

Looking at the data for small and large kidney stones separately.

Small Stones	Treatment A	Treatment B
Succeed	81	234
Fail	6	36
Total	87	270

Large Stones	Treatment A	Treatment B
Succeed	192	55
Fail	71	25
Total	263	80

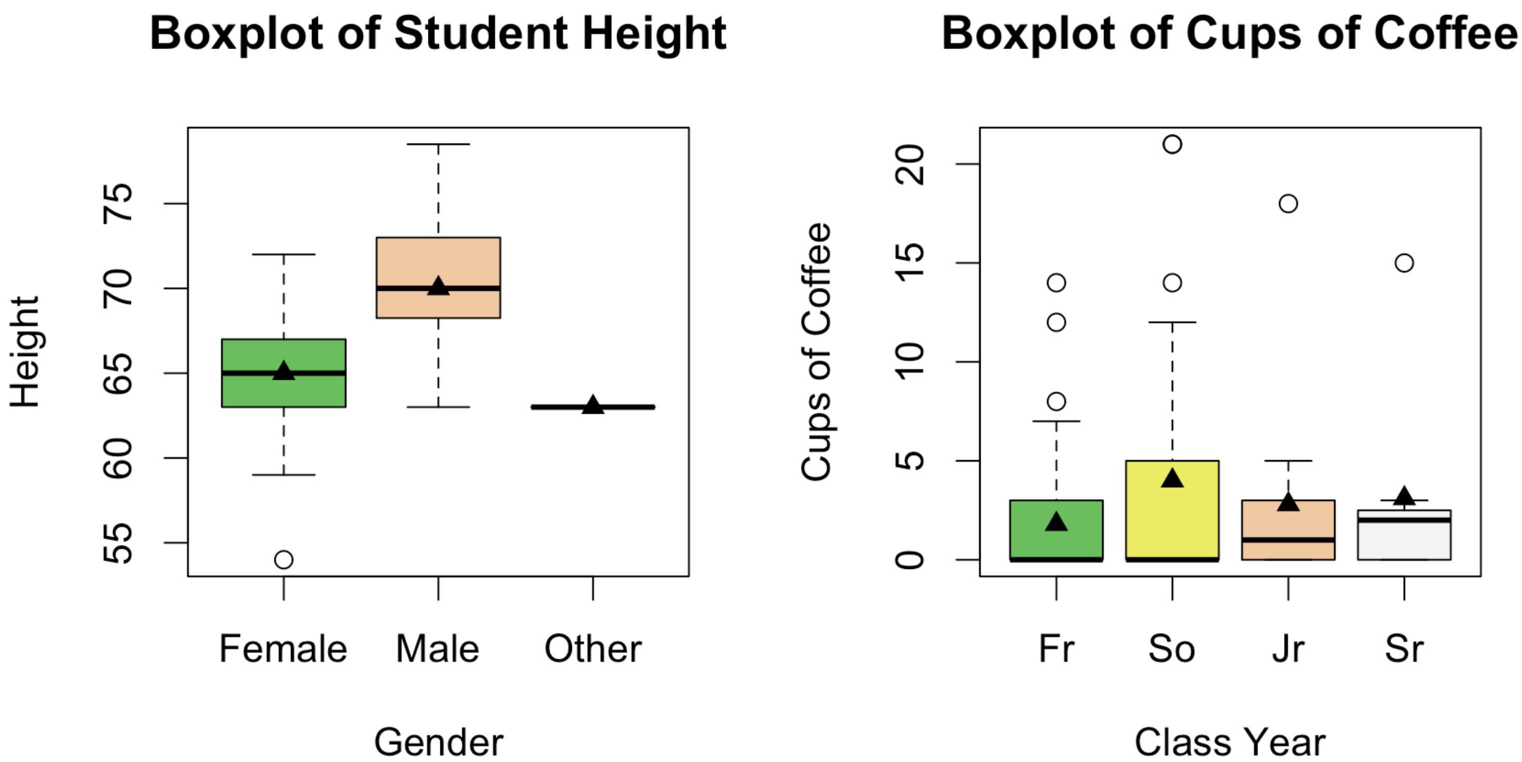
- For small kidney stones, the success rates of treatment A and B are $\frac{81}{87} = 93\%$ and $\frac{234}{270} = 87\%$.
- For large kidney stones, the success rates of treatment A and B are $\frac{192}{263} = 73\%$ and $\frac{55}{80} = 69\%$.
- **Which treatment is better now?**
- Treatment A.

Example 5 - Simpson's paradox

> An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

- How does this happen?
- Treatment A is applied more on large (severer) kidney stones; Treatment B is applied more on small (less severe) kidney stones.
- The overall success rate for small kidney stones is higher and for large kidney stones lower.
- When data are combined, Treatment B has higher success rate.
- But in fact, Treatment A is better for both small and large kidney stones.

Example 6


Boxplot of Student Height
Height
55
60
65
70
75
Female
Male
Other
Gender
Boxplot of Cups of Coffee
Cups of Coffee
0
5
10
15
20
Fr
So
Jr
Sr
Class Year

Relationships

1. Student height vs. shoe length
   - Both caused by genetics and nourishment
2. UFO count vs. temperature
   - Temperature → outdoor activities → UFO reports
3. Traffic death vs. government trust score
   - Both caused by government regulations
4. Respondents' answers vs. wording of survey question
   - Wording → respondents' answers
5. Kidney stones surgery result vs. treatment
   - Surgery result caused by both treatment and severity of disease
6. Coffee consumption vs. class year
   - Class year → workload → coffee consumption

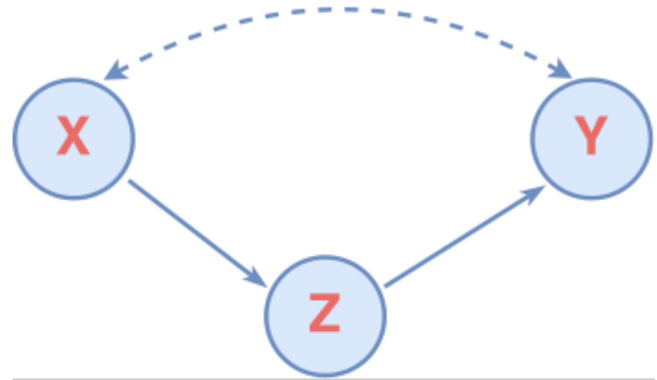
Lurking variable

A **lurking variable** is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

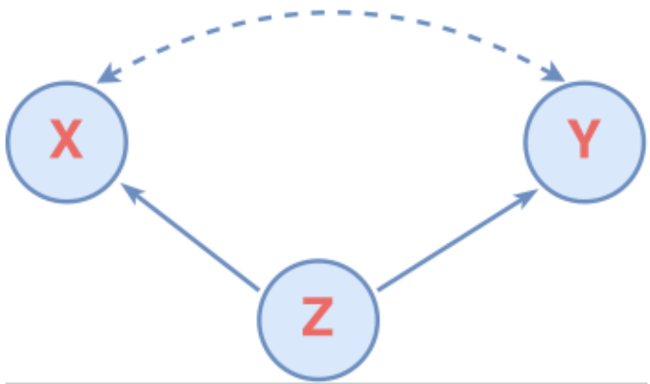
Types of associations

Direct Causation

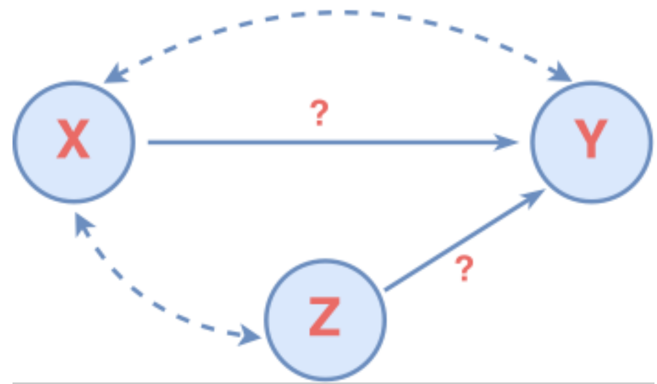
Mediation



Common Response



Confounding



Association and causation

- No matter how strong, association does not imply causation.
- Even direct causation is present, it may not be a complete explanation of the relationship - other causes
  - Risk factors of lung cancer: smoking, secondhand smoke, radon, family history, diet, …
- Studying associations is meaningful
  - Evaluate how response variable changes - prediction
- Finding causal relationships is often essential
  - Explain why response variable changes - e.g., cure diseases

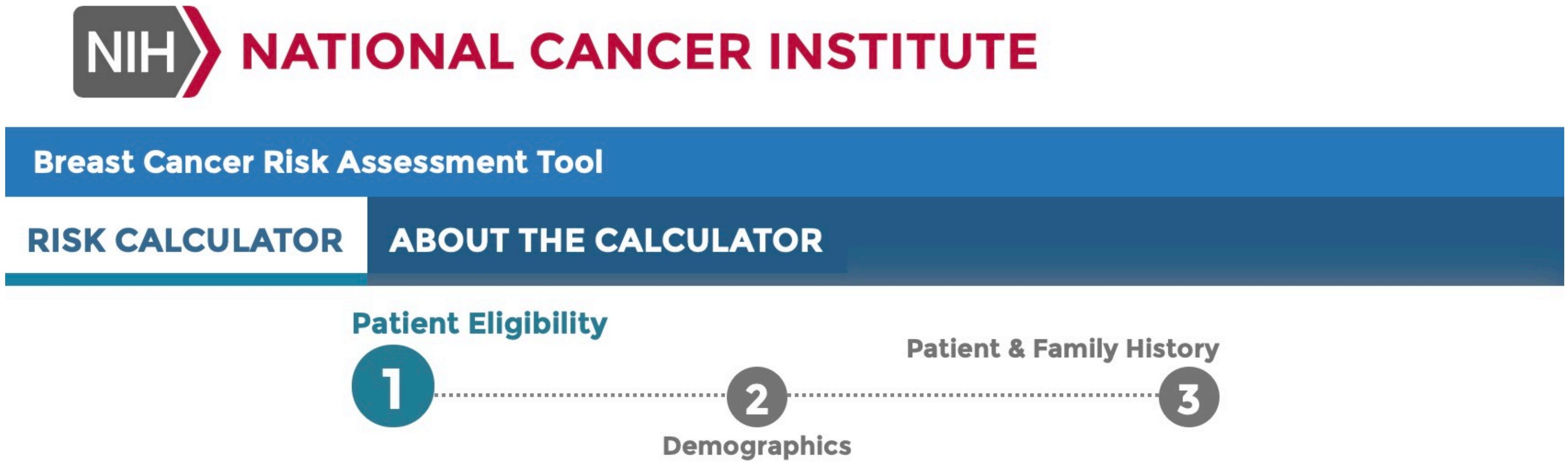
Breast cancer example

In 2013, actress, filmmaker, and human rights activist Angelina Jolie made headlines by announcing that she had undergone a preventative double mastectomy. The reason: A family history of breast cancer (her mother had died of it) and what she called a "faulty gene," referring to the BRCA gene (BRCA 1).

By Dr. Veronique Desaulniers

Breast Cancer Risk Assessment Tool

Breast Cancer Risk Assessment Tool (BCRAT)



- The risk of getting breast cancer for women with known mutations in either the *BRCA1* or *BRCA2* gene is estimated using the BOADICEA model rather than the BCRAT model. The former has much higher predictive accuracy than the latter mainly because the *BRCA1* or *BRCA2* mutation is a causal factor of breast cancer.

Summary

- Relationship between a quantitative variable and a categorical variable
  - Summary statistics `aggregate()`
  - Boxplot `boxplot(Response ~ Explanatory)`
- Association and causation
  - Examples of relationships
    - Simpson's paradox
- Lurking variable
- Types of associations: *direct causation, mediation, common response, confounding*.