

# BIOS 600: Principles of Statistical Inference

## Correlation

Fall 2012

# Reading

- ▶ Pagano and Gauvreau, Chapter 17

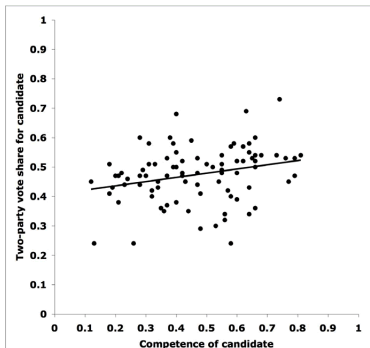
# Motivation

Types of comparisons of interest:

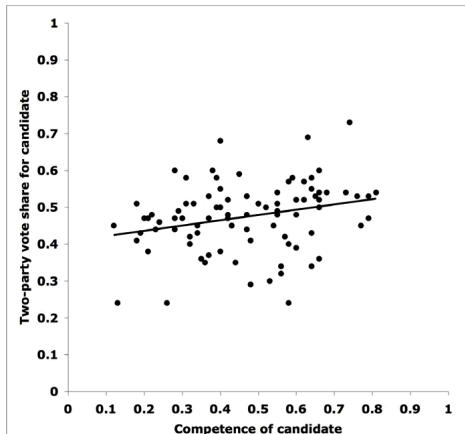
- ▶ Categorical predictor and categorical outcome (can use Fisher's exact test)
- ▶ Categorical predictor and continuous outcome (can use ANOVA, Wilcoxon test, t-test)
- ▶ Continuous predictor and continuous outcome (will begin discussion today!)
- ▶ Continuous predictor and categorical outcome (will discuss later)

# Inspecting a Two-Way Scatterplot

Ballew and Todorov (2007 PNAS) studied the association between snap judgments of the competence level of a gubernatorial candidate and the share of the vote that candidate received in an election.



# Inspecting a Two-Way Scatterplot



Questions of interest:

- ▶ Direction of relationship: are the variables positively or negatively related?
- ▶ Form: is any relationship linear or more complex?
- ▶ Strength of relationship: how accurately can one variable predict the other
- ▶ Outliers: are one or two points driving the relationship we see?

# Correlation

The *correlation coefficient*  $\rho$  quantifies the **linear relationship** between two random variables, which we can call  $X$  and  $Y$ .

IMPORTANT SEMANTICS: In English, we use the word *correlated* to mean *related*. In statistics, a *correlation coefficient* implies a very specific type of association – a linear association of some type (can be described with a straight line). A correlation coefficient of zero does NOT imply no relationship between two variables

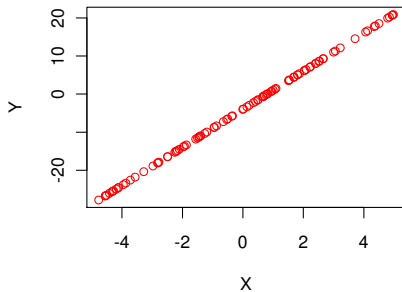
# Correlation

$\rho$  ranges from -1 to 1.

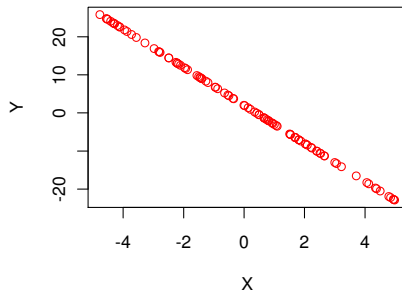
- ▶ When  $\rho > 0$ , larger values of  $X$  are associated with larger values of  $Y$  (scatterplot trends upward)
  - ▶  $\rho = 1$  implies perfect positive correlation, so that  $X$  is a linear function of  $Y$ .
  - ▶ Example: converting weight in pounds ( $X$ ) to kg ( $Y$ ), in which case  $Y = \frac{X}{2.2}$
- ▶ When  $\rho < 0$ , larger values of  $X$  are associated with smaller values of  $Y$  (scatterplot trends downwards)
  - ▶ Example: let  $X$  be minutes of vigorous physical activity per week and  $Y$  be weight
  - ▶  $\rho = -1$  implies perfect negative correlation
- ▶  $\rho = 0$  is consistent with no *linear* relationship between variables

## Different $\rho$ values

**Corr(X,Y)=1**



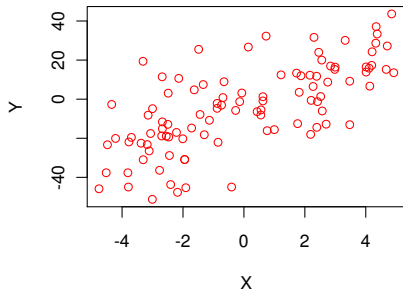
**Corr(X,Y)=-1**



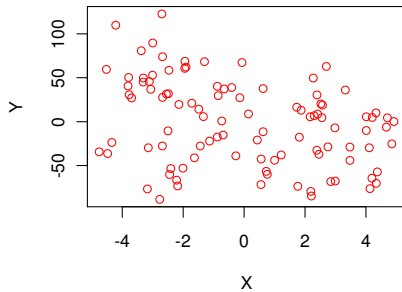


## Different $\rho$ values

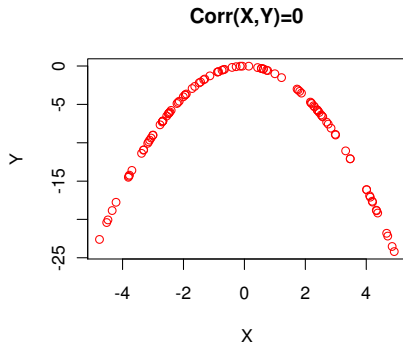
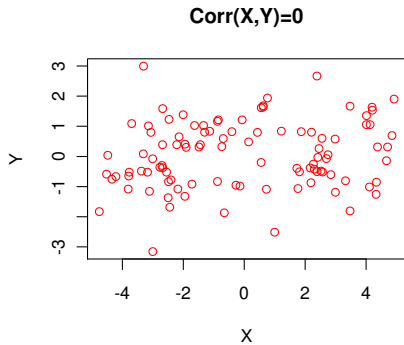
**Corr(X,Y)=0.7**



**Corr(X,Y)=-0.3**



## Different $\rho$ values



# Correlation Game!

Let's play **Guess the Correlation!**

# Now We Play for Real!!!

## Serious Correlation Game

## Formula

We can estimate  $\rho$  using Pearson's correlation coefficient  $r$ .

Assuming our data are the observations

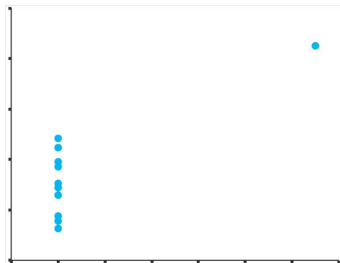
$(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$  we calculate  $r$  as

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

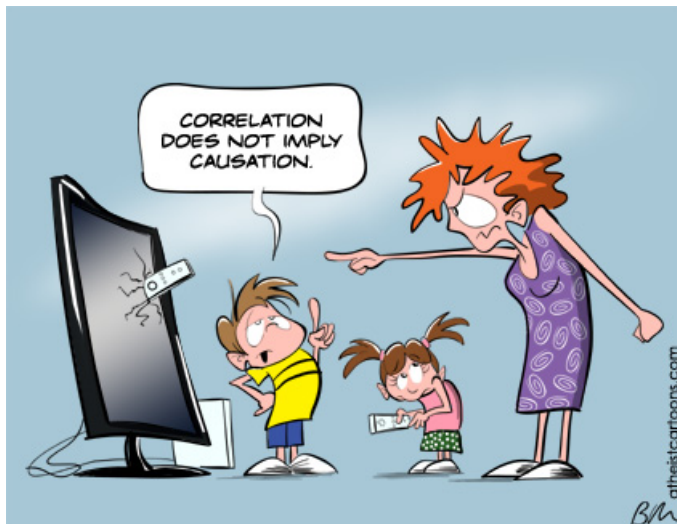
Stata **correlate**

## Correlation and Outliers

Pearson's correlation coefficient can be sensitive to outliers. The following plot is of data with  $r = 0.82$ , but without the one outlier you would obtain  $r = 0$ .



# Correlation and Causation



## Correlation and Causation

## THE FAMILY CIRCUS



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."



## Example: Hormone Replacement Therapy (HRT)

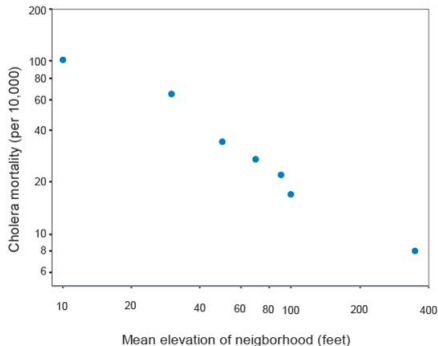
A famous example concerns HRT in the Nurse's Health Study, an observational study of over 100,000 nurses from the 1980's. One of its first major findings was that nurses who used HRT after menopause decreased their risk of heart disease by 50%. Because of the huge effect and the fact that heart disease was the biggest killer of women, millions of women were placed on hormone replacement after menopause.

Unfortunately, the group who took HRT turned out to be health conscious people who also engaged in many other health promoting behaviors - many of which were not fully controlled in the study's design.

A randomized clinical trial in the early 1990's, the Women's Health Initiative, showed that HRT actually increases the risk of heart disease, and many women rapidly discontinued the medication.

## Example: Cholera Mortality

A near perfect negative correlation ( $r = -0.99$ ) was seen between cholera mortality and elevation above sea level during a 19th century epidemic.



The observed relationship between cholera and elevation was confounded by a lurking variable: proximity to polluted water.

# Correlation and Causation

*The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.*

Stephen J. Gould

# Testing Correlation

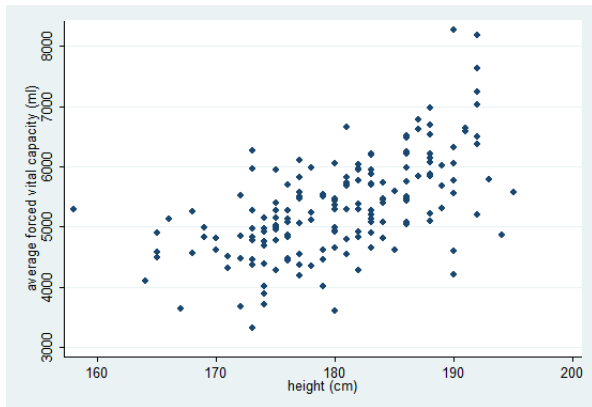
Suppose we wish to test  $H_0 : \rho = 0$  versus  $H_A : \rho \neq 0$ . If we can assume our two variables of interest,  $X$  and  $Y$ , are normally distributed, then we can use the t-statistic given by

$$t = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}},$$

which under  $H_0$  follows a  $t_{n-2}$  distribution.

## Example: EPA Study

The EPA's Health Effects Research Laboratory conducted a study of 172 adult males on campus. We examine the correlation between average forced vital capacity (ml) and height (cm).



What is a likely value of  $\rho$ ?

## Example: EPA Study

```
. pwcorr avgfvc heightcm, sig
```

	avgfvc heightcm	
avgfvc	1.0000	
heightcm	0.5775	1.0000
	0.0000	

The numbers on the top row of each cell are the correlations between row and column variables. (A variable is perfectly correlated with itself, which explains the 1's.) The number 0.5775 is the estimated correlation between height and FVC, and the number below it (0.000) is the p-value from testing the null hypothesis that the population correlation is zero.

# Coefficient of determination

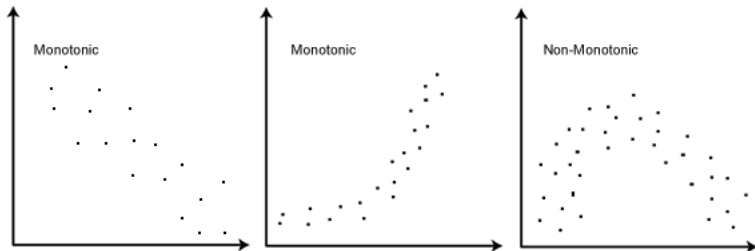
The square of the correlation coefficient,  $r^2$ , is called the coefficient of determination.

- ▶  $r^2$  ranges from 0 to 1
- ▶  $r^2$  quantifies the proportion of variance in  $Y$  that is explained by its linear relationship with  $X$
- ▶ For the EPA data, we have  $r = 0.5775$  (significantly different from 0) and  $r^2 = 0.33$ , so height explains about 33% of the variability of FVC.

# Spearman's correlation coefficient

What can you do if there are several influential outliers in your data, or if the relationship is not linear? Spearman's correlation coefficient is a nonparametric alternative.

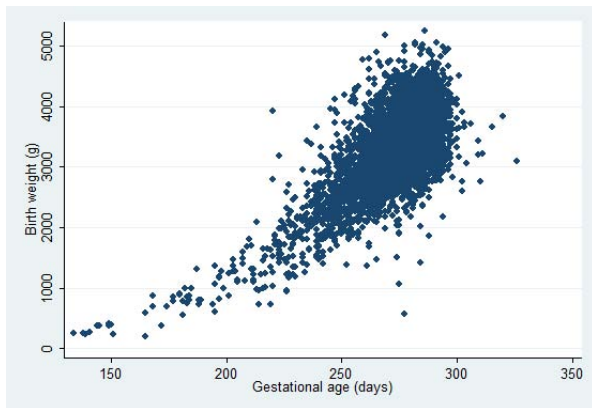
Spearman's correlation coefficient is just Pearson's correlation coefficient calculated on the *ranks* of the data. It can help with non-linear monotone relationships or with outliers.





## Example: Birth weight data

Consider the following data on gestational age in days and birth weight in g.



This looks monotone but not exactly linear.

## Example: Birth weight data

We can easily calculate Spearman's correlation coefficient for these data and test whether it is zero.

```
. spearman birthweight gestage
```

```
Number of obs =      5034  
Spearman's rho =      0.5481
```

```
Test of Ho: birthweight and gestage are independent  
Prob > |t| =      0.0000
```

## You Try It: Correlation Applet

This great **correlation applet** allows you to see how correlation is related to points you place on a scatterplot.