STA 570 Spring 2011

Lecture 6 *Thursday, Feb 3*

- **Study Designs**
- **Summarizing Bivariate Data**
  - ❑ **Two categorical variables**
    Contingency Table
    Row/Column Relative Frequencies
    Relative Risk, Odds Ratio
  - ❑ **Two quantitative variables**
    Scatter Diagram
    Regression Line
    Correlation Coefficient, Coefficient of Determination, Slope and Intercept of Regression Line

Homework 4: Due next week in lab.

Example Revisited

- Data

5.5, 18.5, 6.0, 5.5, 5.3, 5.8, 11.0, 6.1, 7.0, 14.5,

10.4, 7.6, 4.3, 7.2, 10.5, 6.5, 3.3, 2.0, 4.1, 6.2

- Five-Number Summary

Min=2.0, Q1=5.4, Med=6.15,Q3=9.0, Max=18.5

IQR=Q3 – Q1=3.6

- Inner Fences

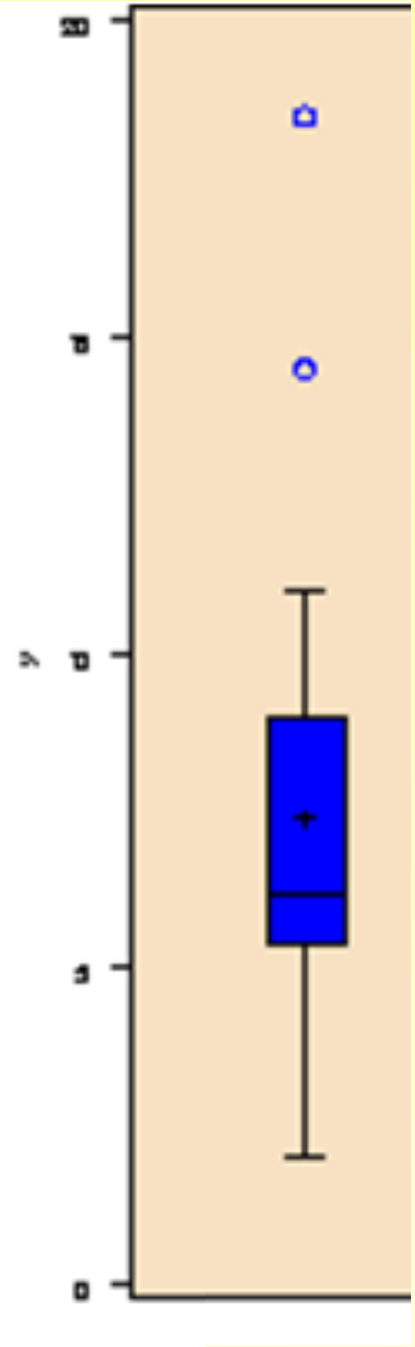
Q3+1.5*IQR=9.0+1.5*3.6=9.0+5.4=14.4

Q1–1.5*IQR=5.4–1.5*3.6=5.4–5.4=0.0

- Outliers?

Example Revisited

```
Stem Leaf                    #  Boxplot
  18 5                    1     0
  16
  14 5                    1     0
  12
  10 450                    3     |
   8                       +-----+
   6 0125026                 7  *--+--*
   4 133558                 6  +-----+
   2 03                   2     |
    ----+----+----+----+
```



Important Study Designs

- Prospective Study (Cohort Study)
  - Group of disease-free individuals (“cohort”) identified
  - Followed over time until some develop the disease
  - Relate development of disease (“incidence”) to variables measured at baseline (exposure variables)
- Retrospective Study (Case-Control Study)
  - Two groups of individuals identified: “cases” with disease, and “controls” without disease
  - Related the current disease status to prior health habits
- Cross-sectional Study (Prevalence Study)
  - Group of individuals is asked about current disease status and current or past exposure
  - Prevalence of disease compared between exposed and unexposed individuals

Describing the Relationship Between Two Nominal (or Ordinal) Variables

Contingency Table

- Number of subjects observed at all the combinations of possible outcomes for the two variables
- Contingency tables are identified by their number of rows and columns
- A table with 2 rows and 3 columns is called 2x3 table (“2 by 3”)

2x2 Table: Example

- 327 commercial motor vehicle drivers who had accidents in Kentucky from 1998 to 2002
- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

Contingency Table: Example, contd.

- How can we compare fatality rates for the two groups?
- Relative frequencies or percentages within each row
- Two sets of relative frequencies (for *seatbelt=yes* and for *seatbelt=no*), called ***row relative frequencies***
- If seat belt use and fatality of accident are related, then there will be differences in the row relative frequencies

Row relative frequencies

- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes			100
	No			100
				100

(No) Association

- If there is no association between two variables, then the row relative frequencies should be about the same

Fictitious Row Relative Frequency		VARIABLE Y		
		Yes	No	
VARIABLE X	Yes			100
	No			100
				100

Fictitious (Absolute) Frequency		VARIABLE Y		
		Yes	No	
VARIABLE X	Yes			242
	No			85
		63	264	327

- Note that the column relative frequencies should also be about the same when there is no association.

Association

- ***Association***: The distribution of the response variable changes in some way as the value of the explanatory variable changes
- Example: If all the Pattersonites love soccer, while the Communications students prefer basketball or baseball, then there is association between the two categorical variables “major” and “preferred sports”

Independence

- "*No association*" is called ***Independence***.
- For example, if 50% of ***every*** group (Patterson, Communications, …) prefer soccer, 30% prefer basketball, 20% prefer football, then the two categorical variables are independent
- In practice there is rarely data with perfect independence, and in samples, there is sampling variation

Measuring Association in 2x2 Tables: Example

- Case 1: Weak association

		Use Twitter	
		Yes	No
Gender	Male	15	25
	Female	25	15

- Case 2: Maximum association or strong association

		Use Twitter	
		Yes	No
Gender	Male	40	0
	Female	0	40

Measuring Association

- A measure of association is a statistic that summarizes the strength of the statistical dependence between two variables
- Common measures of association are
  - ***Difference between the group proportions*** for a given response level (Risk difference)
  - **Relative risk** (Risk ratio)
  - **Odds ratio**

Difference of Proportions (Risk Difference): Example

- Case 1:

Difference

37.5% – 62.5%
= – 25%

Proportions		Use Twitter		
		Yes	No	Total
Gender	Male	15/40=37.5%	25/40=62.5%	100%
	Female	25/40=62.5%	15/40=37.5%	100%

- Case 2:

Difference

100% – 0%
= 100%

Proportions		Use Twitter		
		Yes	No	Total
Gender	Male	40/40=100%	0/40=0%	100%
	Female	0/40=0%	40/40=100%	100%

Difference of Proportions (Risk Difference): Limitations

Proportions		Disease		
		Yes	No	Total
Exposure	A	37.5%	62.5%	100%
	B	34.5%	65.5%	100%

- Case 1:

Difference

37.5% – 34.5%
= 3%

Proportions		Disease		
		Yes	No	Total
Exposure	A	3.01%	96.99%	100%
	B	0.01%	99.99%	100%

- Case 2:

Difference

3.01% – 0.01%
= 3%

Solution: Ratio of Proportions (Relative Risk, Risk Ratio)

Proportions		Disease		
		Yes	No	Total
Exposure	A	37.5%	62.5%	100%
	B	34.5%	65.5%	100%

- Case 1:

Ratio

0.375 / 0.345
= 1.087

Proportions		Disease		
		Yes	No	Total
Exposure	A	3.01%	96.99%	100%
	B	0.01%	99.99%	100%

- Case 2:

Ratio

0.301 / 0.0001%
= 3010

Measures of Association

Proportions (not by row)		Disease		
		Yes	No	Total
Exposure	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	a+b+c+d

- Difference of proportions (Risk difference): $$\frac{a}{a+b}-\frac{c}{c+d}$$
- Relative Risk (RR): $$\frac{a}{a+b}/\frac{c}{c+d}$$
- Odds ratio (OR): $$\frac{a}{b}/\frac{c}{d}$$

Measures of Association: Example

Frequencies		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

Row Relative Frequencies		Accident Fatal		
		Yes	No	
Seat Belt	Yes	.124	.876	1
	No	.388	.612	1
		.193	.807	1

- Difference of Proportions: 0.124 – 0.388 = -0.264
- Relative Risk: 0.124 / 0.388 = 0.319 (=1/3.1)
- Odds Ratio: (30/212) / (33/52)

  = 0.1415 / 0.6346 = 0.223 (=1/4.5)

Measuring Association: Odds Ratio

- Within a row, the odds of success are defined to be

Odds = (Proportion of "event")/(Proportion of "no event")

- Example: Odds of fatal accident for seat-belt wearers were 30/212=0.1415, and for non-seat-belt wearers they were 33/52=0.6346
- Ratio of the odds from the two rows:***odds ratio***
- In this example, odds ratio = 0.1415 / 0.6346 = 0.223
- "For drivers wearing a seat belt, the odds of an accident being fatal were 0.223 times the odds for drivers not wearing a seat belt (or 'about 4.5 times smaller')."

Measuring Association: Odds Ratio

- There is a shortcut formula for the odds ratio:

  The odds ratio equals the ratio of the products of cell counts from diagonally opposite cells:
- (30 x 52) / (33 x 212) = 0.223
- When the odds ratio is greater than 1, the odds of “disease” are higher in row 1 than in row 2
- Values of the odds ratio farther from 1.0 in a given direction represent stronger association

Why the Odds Ratio?

- Isn’t the Relative Risk more intuitive, easier to interpret?
- Yes, but there are situations where the Relative Risk does not make sense: Case-Control Studies
- Example
  - *Case control study of prostate cancer risk and male pattern balding.*
  - *Are men with certain hair patterns at greater risk of prostate cancer?*
  - *Roughly equal numbers of prostate cancer patients and (healthy) controls were selected.*
  - *Among cancer patients, 72 out of 129 had either vertex or frontal baldness, compared to 82 out of 139 among controls*

	Cases	Controls	Total
Balding	72	82	154
Hairy	55	57	112
Total	129	139	268

Why the Odds Ratio?

- You can calculate the proportion of bald men among the cases and among the controls, but it makes no sense to estimate the population proportion of cancer patients among bald men using this data because the prevalence of cancer was artificially inflated to about 50% because of the study design.
- However, you can *always* calculate and interpret the odds ratio in a case control study (as long as the prevalence of the outcome is relatively rare)
- The inflated prevalence *cancels out* in the odds ratio formula, but not in the relative risk formula.

	Cases	Controls	Total
Balding	72	82	154
Hairy	55	57	112
Total	129	139	268

Another Advantage of the Odds Ratio

- For every problem, there are two ways to calculate the relative risk.
- Example
  - *How much does a treatment intervention increase the probability of success? OR*
  - *How much does a treatment intervention decrease the probability of failure?*

	Success	Failure	Total
Treatment	19 (37.3%)	32 (62.7%)	51
Control	5 (8.8%)	52 (91.2%)	57
Total	24	84	108

- Risk Ratio = 0.373/0.088 = 4.2 [times more success]   OR
- Risk Ratio = 0.627/0.912 = 0.7 [1.4 times less failure]
- Odds Ratio = (19x52)/(5x32)=6.2 [times higher odds for success] OR
- Odds Ratio = (32x5)/(19x52)=0.16 [6.2 times smaller odds for failure]

Summary: Measures of Association for Categorical Variables

- For events (diseases) with small prevalence, the difference of proportions is not informative.
- The relative risk is generally easier to interpret and more intuitive.
- However, sometimes it is not appropriate: In case-control studies, the odds ratio should be used. Also, there is potential ambiguity as to which relative risk should be compared.
- The odds ratio should only be used for events with small prevalence.
- When you read literature using any of these, be aware of the limitations.

Alcohol and Cigarettes

- Alcohol vs. Cigarette use of senior high school students in Dayton, OH

		Cigarette Use		
		Yes	No	
Alcohol Use	Yes	1449	500	
	No	46	281	

Describing the Relationship Between Two Quantitative Variables

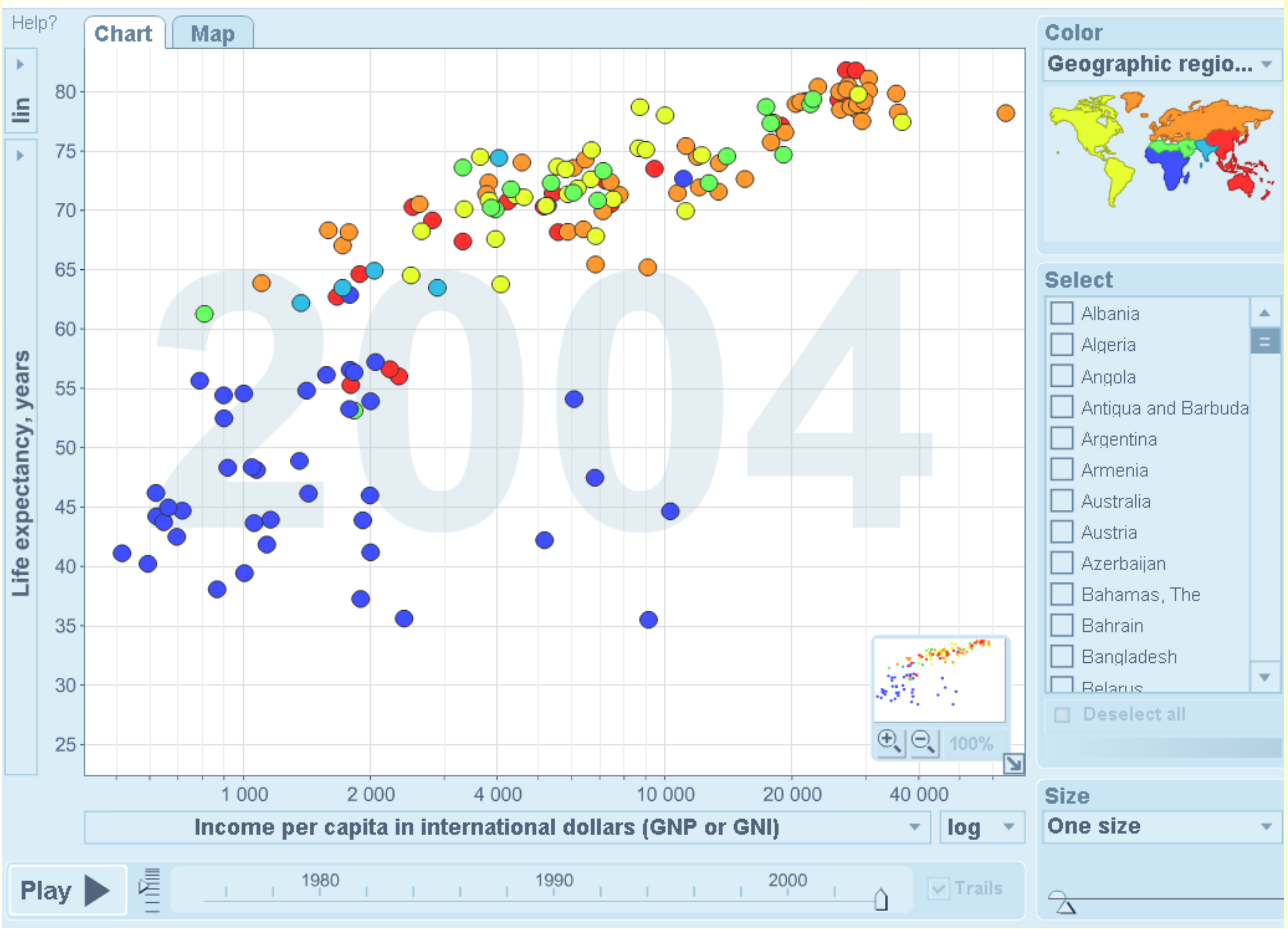
Scatter Diagram

- In applications where one variable depends to some degree on the other variables, we label the dependent variable Y and the independent variable X
- Example:

  Life expectancy = X

  Income = Y
- Each point in the scatter diagram corresponds to one observation

Scatter Diagram of Life Expectancy (Y) and Income (X) for Several Countries



Color: Geographic regio...

Size: One size

Source: gapminder.org

Analyzing Linear Relationships Between Two Quantitative Variables

- Is there an association between the two variables?
- Positive or negative?
- How strong is the association?
- Notation
  - Response variable: *Y*
  - Explanatory variable: *X*

Sample Measures of Linear Relationship

- Sample Covariance:

$$s_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{1}{n-1}\left(\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right)$$

- Sample Correlation Coefficient:

$$r = \frac{s_{xy}}{s_x s_y}$$

- Here, s_x and s_y are the standard deviations of the x and y variables
- Population measures: Divide by N instead of n-1

Properties of the Correlation I

- The value of *r* does not depend on the units (e.g., changing from inches to centimeters)
- *r* is standardized
- *r* is always between –1 and 1, whereas the covariance can take *any* number
- *r* measures the ***strength and direction of the linear association*** between *X* and *Y*
- r>0 positive linear association
- r<0 negative linear association

Properties of the Correlation II

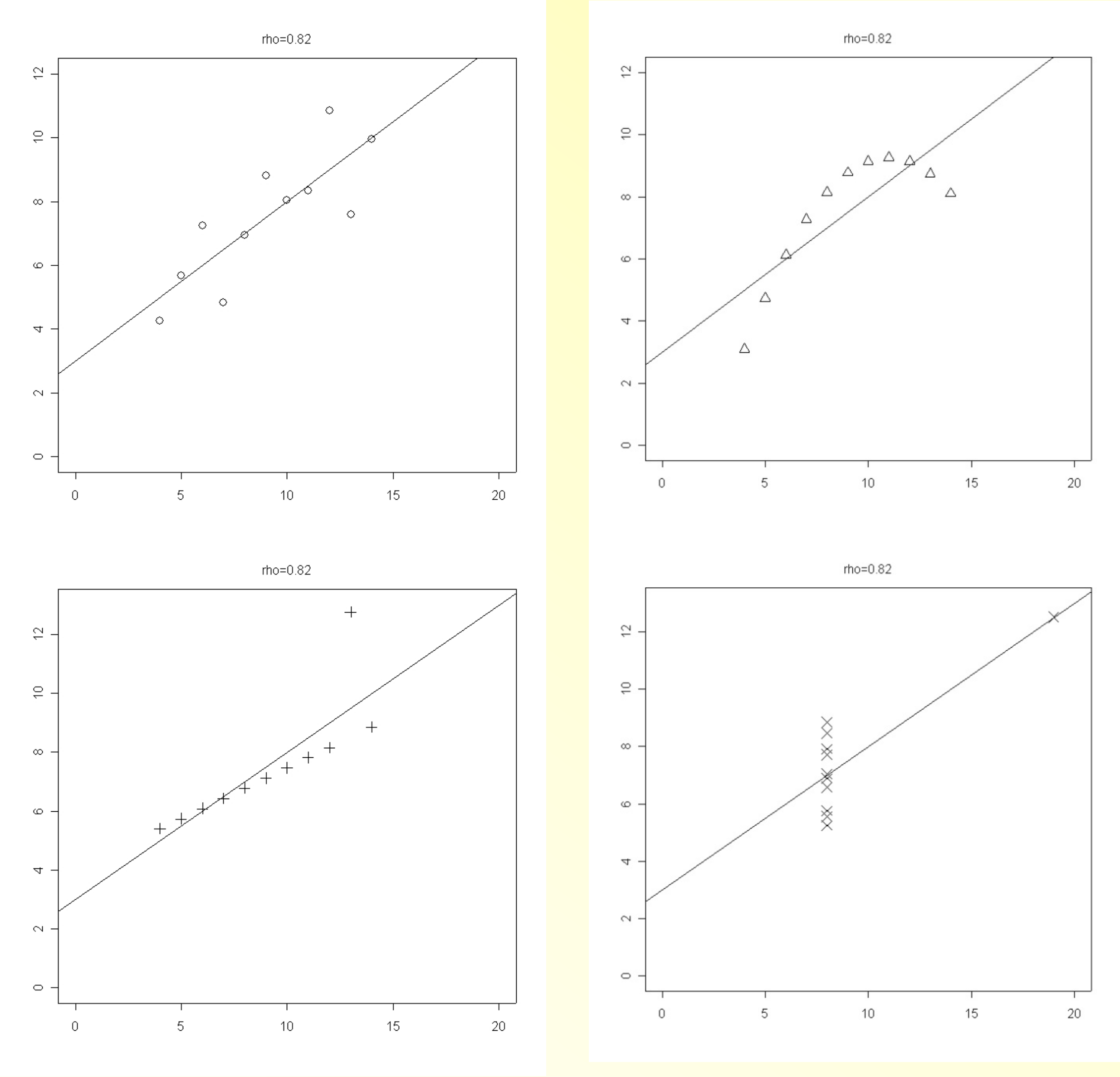
- $r = 1$ when all sample points fall exactly on a line with positive slope *(perfect positive association)*
- $r = -1$ when all sample points fall exactly on a line with negative slope *(perfect negative association)*
- The larger the absolute value of r, the stronger is the degree of linear association

Properties of the Correlation III

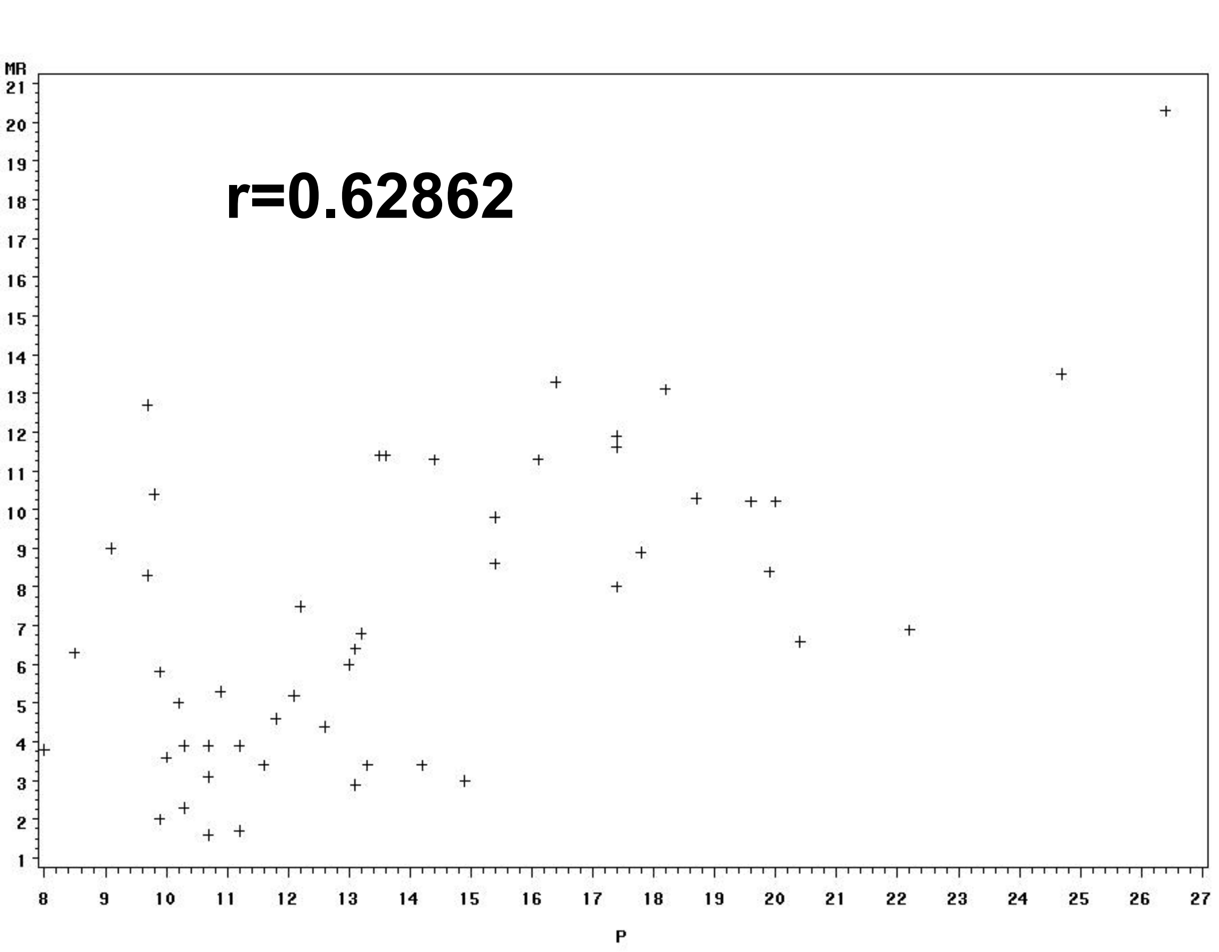
- If *r* is close to 0, this does not necessarily mean that the variables are not associated
- It only means that they are not *linearly associated*

- The correlation treats *X* and *Y* symmetrically
- That is, it does not matter which variable is explanatory (*X*) and which one is response (*Y*), the correlation remains the same

Example: Correlation = 0.82



Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



r=0.62862

Correlation and Scatterplot Applet

Correlation by Eye Applet

Simple Regression Analysis Tool

Model Assumptions and Violations

- **Factors Influencing the Correlation**
- The sample correlation depends on the range of *X*-value sampled
- When a sample has a much narrower range of variation in *X* than the population, the sample correlation tends to underestimate the population correlation
- The sample (*X*,*Y*) values should be a random sample of the population
- It should be representative of the *X* population values as well as the *Y* values

Correlation: Example

- For a sample of 100 people, the correlation coefficient between X = hours of statistics instruction and Y = annual income (in dollars) equals 0.70

a) Suppose instead that X refers to minutes instead of hours, and Y refers to monthly income converted into Euro. State the correlation.

b) Suppose that Y is treated as the explanatory variable and X is treated as the response variable. Will the correlation coefficient change in value?