STA291
Fall 2008

LECTURE 13

Thursday, 8th October

Administrative

Ch 6 Measures of Variation

Ch 7 Measures of Linear Relationship

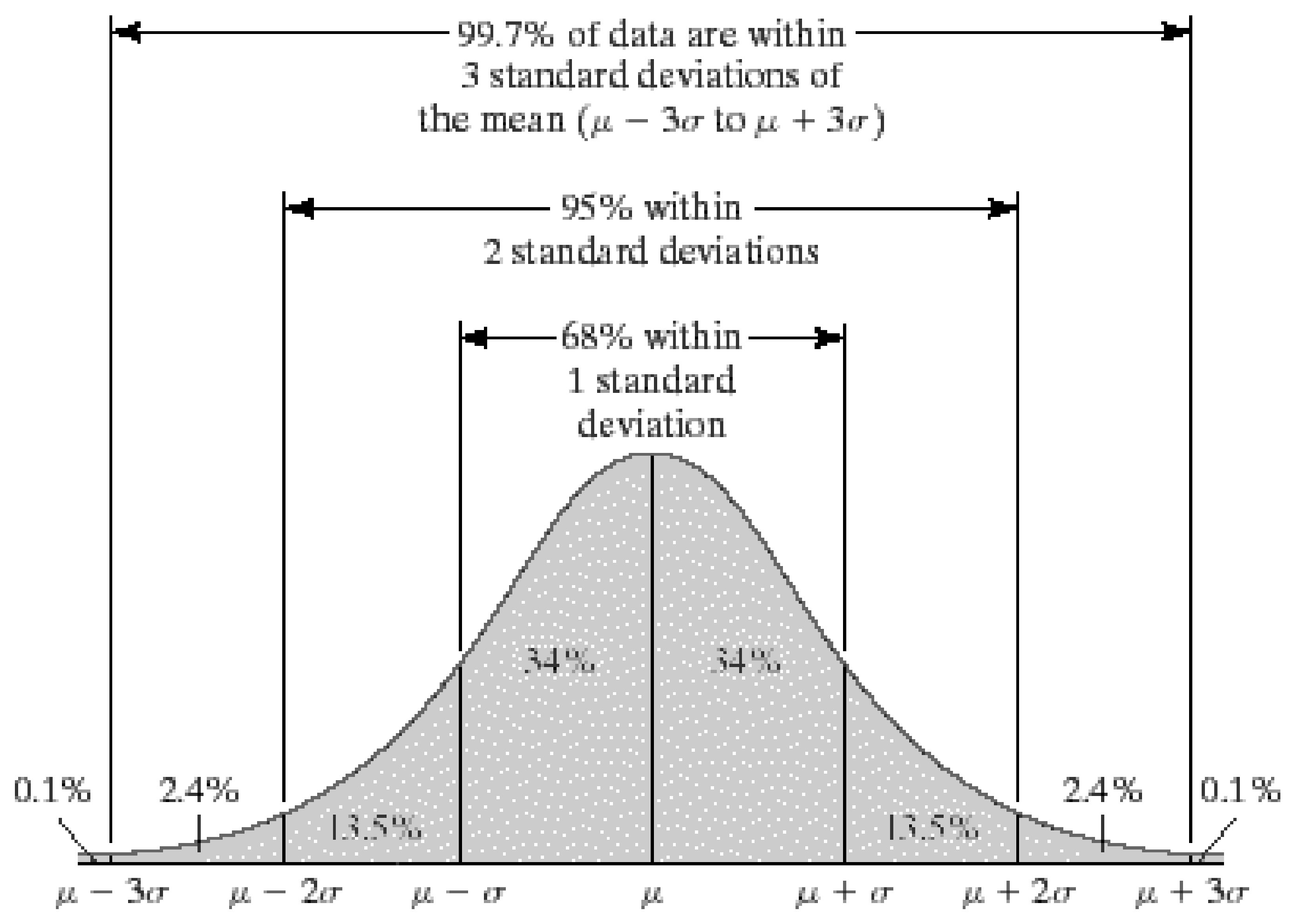
- Suggested Exercises: 7.2, 7.3, 7.4, 7.5 in the textbook

Empirical Rule Example

- Distribution of SAT score is scaled to be approximately bell-shaped with mean 500 and standard deviation 100

  - About 68% of the scores are between ___ ?
  - About 95% are between _____________ ?
  - If you have a score above 700, you are in the top _____________%?

Standard Deviation Interpretation: Empirical Rule



Example Data Sets

- One Variable Statistical Calculator (link on web page)
- Modify the data sets and see how mean and median, as well as standard deviation and interquartile range change
- Look at the histograms and stem-and-leaf plots – does the empirical rule apply?
- Make yourself familiar with the standard deviation
- Interpreting the standard deviation takes experience

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:

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

- Sample Correlation Coefficient:

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

- 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), whereas the covariance does
- *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 linear association)*

- r = – 1 when all sample points fall exactly on a line with negative slope *(perfect negative linear association)*

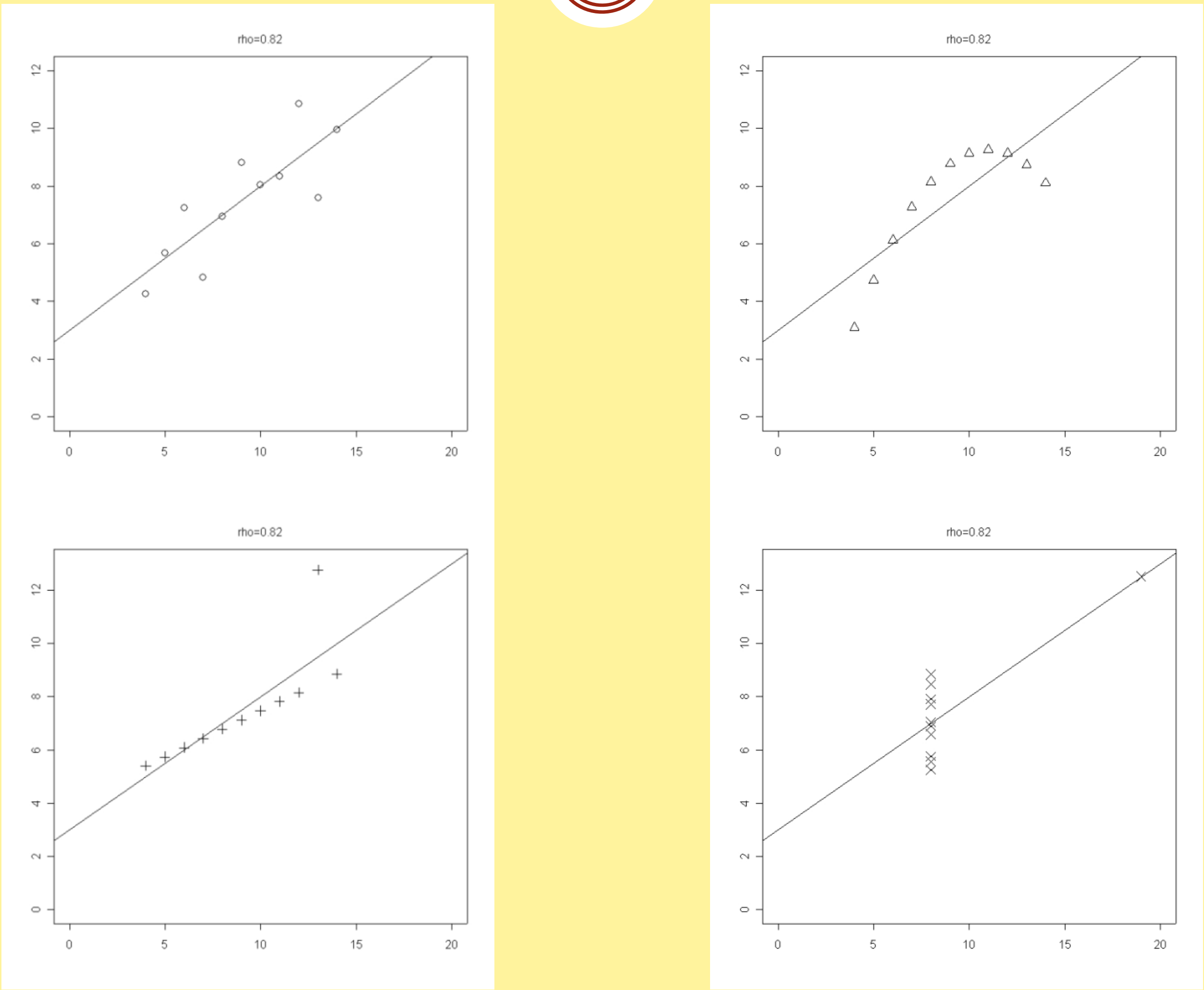
- The larger the absolute value of r, the stronger is the degree of linear association

Properties of the Correlation III

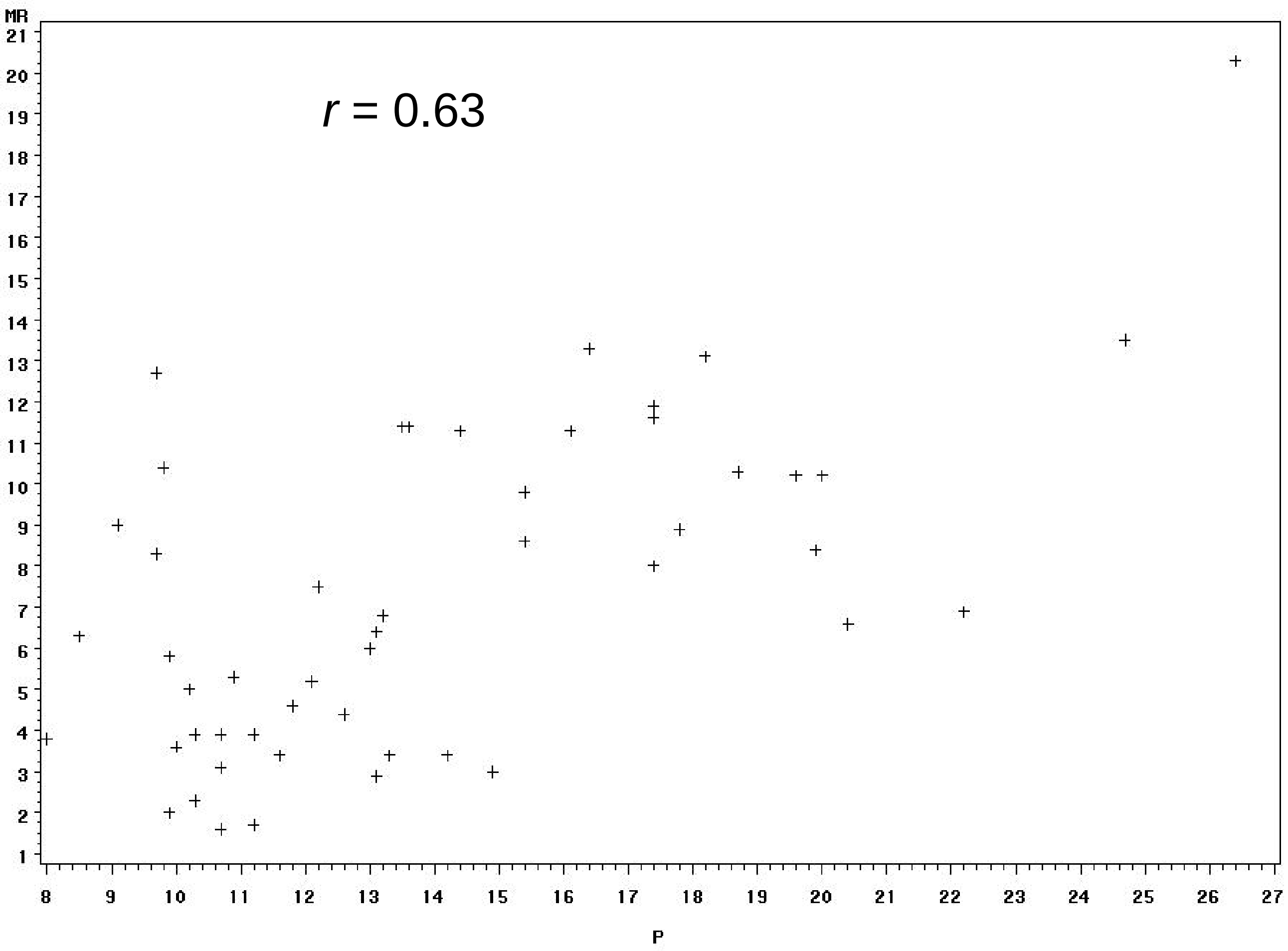
- If *r is close to 0, this does not necessarily mean* that the variables are not associated
- It only means that they are not *linearly* associated

- 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

All Correlation $r = 0.82$



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



Correlation and Scatterplot Applet

Correlation by Eye Applet

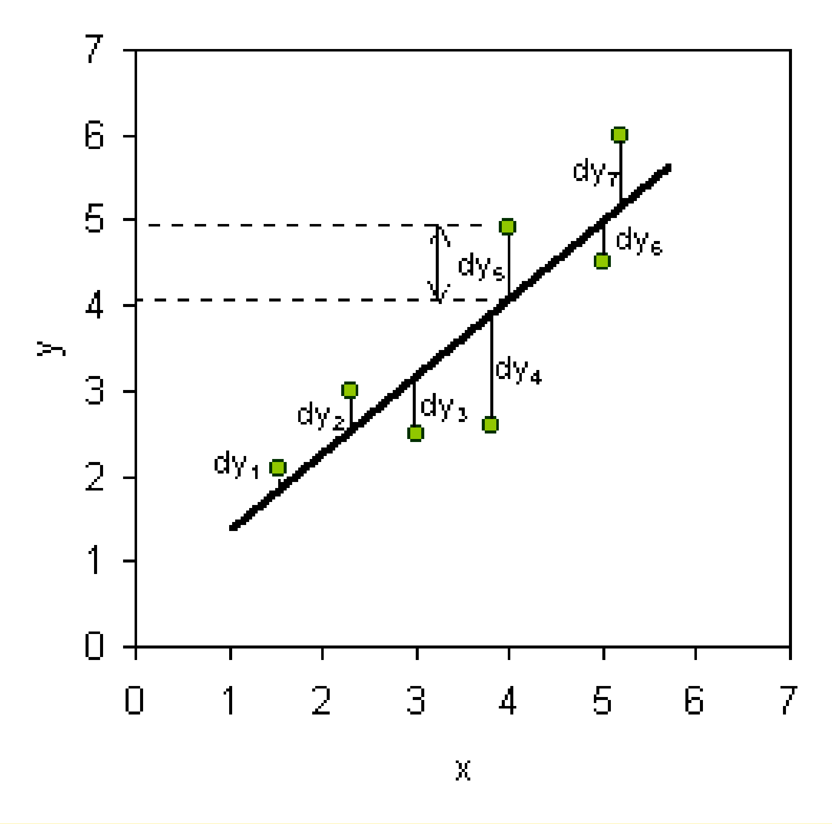
Simple Regression Analysis Tool

r Measures Fit Around *Which* Line?

- As you'll see in the applets, putting the "best" line in is, uh, challenging—at least by eye.
- Mathematically, we choose the line that minimizes error as measured by vertical distance to the data
- Called the "least squares method"
- Resulting line: $\hat{y} = b_0 + b_1 x$
- where the slope, $b_1 = \dfrac{s_{xy}}{s_x^2}$
- and the intercept, $b_0 = \bar{y} - b_1 \bar{x}$

What line?

- *r* measures "closeness" of data to the "best" line. How best? In terms of least squared error:



“Best” line: least-squares, or regression line

- Observed point: (x_i, y_i)

- Predicted value for given x_i: $\hat{y}_i = b_0 + b_1 x_i$
  (How? Interpretation?)

- “Best” line minimizes $\sum (y_i - \hat{y}_i)^2$, the *sum of the squared errors*.

Interpretation of the b_0, b_1

$$\hat{y}_i = b_0 + b_1 x_i$$

- b_0 **Intercept**: *predicted* value of y when $x = 0$.
- b_1 **Slope**: *predicted* change in y when x increases by 1.

Interpretation of the b_0, b_1, $\hat{y}_i$

In a fixed and variable costs model:

$$\hat{y}_i = 9.95 + 2.25x_i$$

- b_0 =9.95? **Intercept**: *predicted* value of *y* when *x* = 0.

- b_1 =2.25? **Slope**: *predicted* change in *y* when *x* increases by 1.

Properties of the Least Squares Line

- b_1, slope, always has the same sign as r, the correlation coefficient—but they measure different things!

- The sum of the errors (or *residuals*), $\left(y_i - \hat{y}_i\right)$, is always 0 (zero).

- The line always passes through the point $\left(\bar{x}, \bar{y}\right)$.

Attendance Survey Question 13

- ***On a your index card:***
  - Please write down your name and section number
  - Today's Question: