THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Online Review Course of
Undergraduate Probability and Statistics

Review Lecture 15
Correlation

Chris A. Mack
Adjunct Associate Professor

Course Website: www.lithoguru.com/scientist/statistics/review.html

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Bivariate Data

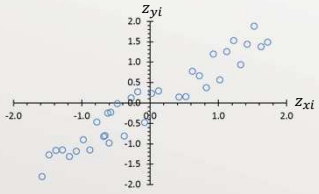
- Bivariate data: data for two variables, usually thought to be related
  - Generally a set of (x_i,y_i) pairs
- A frequent task is to understand what relationship (if any) exists between these two variables
- Two most important tools:
  - X-Y scatterplot
  - Correlation Coefficient (followed by regression)
    - Correlation = the strength of the linear relationship between two quantitative variables

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Converting Data to Z-scores

- Consider the set of (x_i,y_i) pairs, i = 1, 2, …, n
- Z-scores: $z_{xi} = \frac{x_i - \bar{x}}{s_x}$ $z_{yi} = \frac{y_i - \bar{y}}{s_y}$
- Now create a scatterplot of z-scores



© Chris Mack, 2014

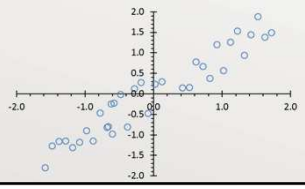
THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Correlation Coefficient

- The Pearson's Sample Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{xi} z_{yi} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Also called the product moment correlation



© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

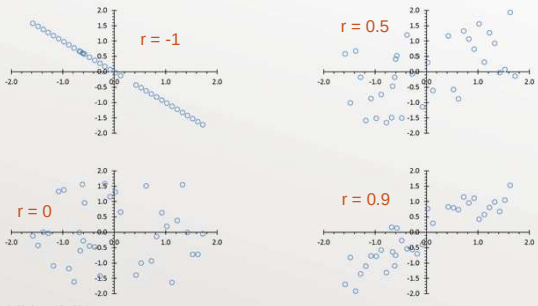
Notes on the Correlation Coefficient

- r is dimensionless (independent of units on x and y)
- r is independent of what is labeled x or y
- r is always between -1 and 1 $(-1 \le r \le 1)$
- If |r| = 1, the data forms a perfect straight line
- r is a measure of the linear relationship between x and y
- Correlation is not causation: the role of x and y is symmetric in calculating r

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Correlation Examples

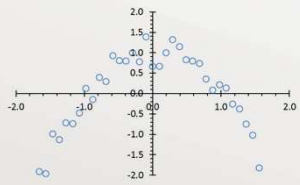


© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

When the Correlation Coefficient Goes Wrong

- r is a measure of the linear relationship between x and y



r = 0, but there is a strong relationship between x and y

The variables are highly related, but not linearly correlated

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Covariance

- The covariance of two random variables X and Y

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- The sample covariance of x and y is

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

- The correlation coefficient is the covariance normalized by the standard deviations

$$r = \frac{cov(x,y)}{s_x s_y}$$

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Covariance Properties

$$cov(X,Y) = E[XY] - E[X]E[Y]$$

$$cov(X,a) = 0$$

a, b = scalars

$$cov(X,X) = \sigma_X^2$$

$$cov(X,Y) = cov(Y,X)$$

$$cov(aX,bY) = abcov(X,Y)$$

$$cov(X+a,Y+b) = cov(X,Y)$$

If X and Y are independent, then $cov(X,Y) = 0$

© Chris Mack, 2014

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Review #15: What have we learned?

- Explain the significance of the correlation coefficient.
- What does the sign of the correlation coefficient tell you?
- How does the covariance relate to the correlation coefficient?
- How does covariance relate to variance?
- What is the covariance of two independent random variables?

© Chris Mack, 2014