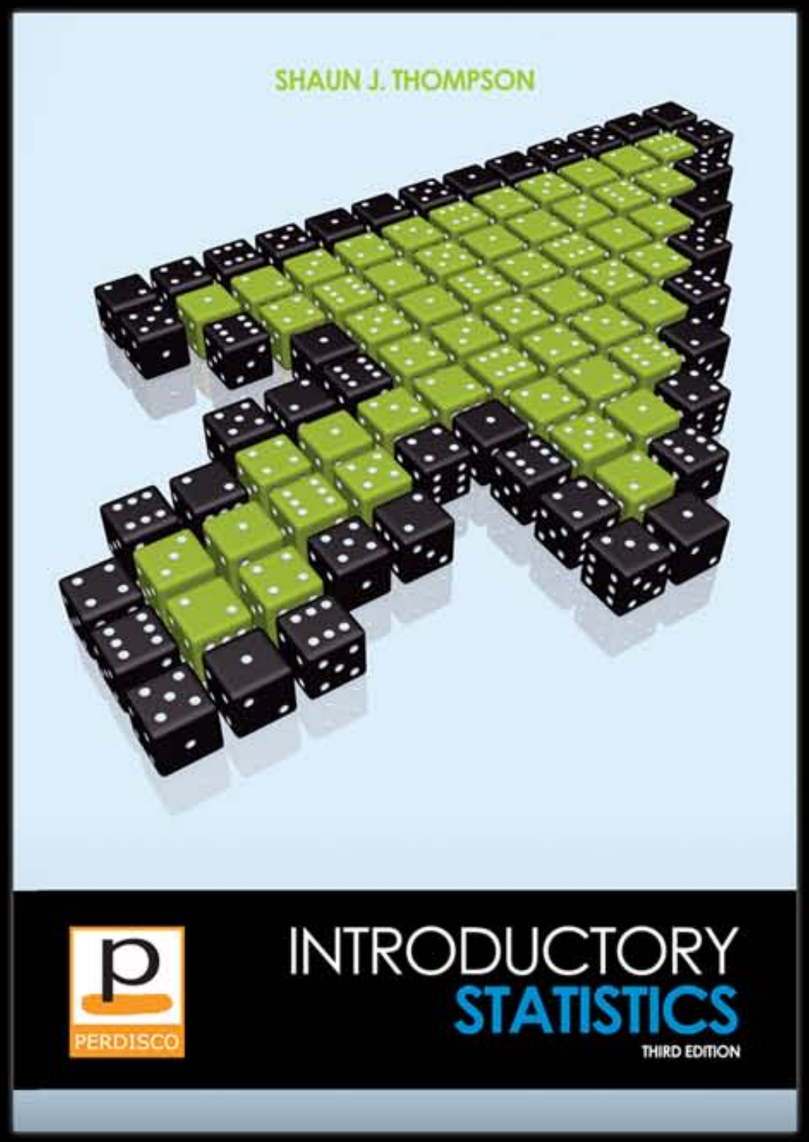


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

Other features include:

- Chapter summary videos
- MP3 audio podcasts
- VirtualTutor e-learning
- AntiCheat and AutoGrade homework
- Detailed instructor resources

To find out more, visit: perdisco.com/stats

- We usually measure samples, not populations
- But population measures are still important
- In fact, they are what we are trying to estimate when we measure a sample

Population mean

- The most common measure of the center of a population is the population mean, μ
- For a (finite) population of N data values $\{x_1,\ldots,x_N\}$

$$\mu = \frac{x_1 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- This is analogous to the sample mean
- But it is a measurement over the whole population

Population variance & standard deviation

- The two most common measures of variation in a population are:
  - variance, σ^2
  - standard deviation, σ
- For a finite population of N data values $\{x_1,\ldots,x_N\}$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \qquad \sigma = \sqrt{\sigma^2}$$

Measuring relationships

- What about the relationship between two variables?

- Scatterplots were used to present relationships
- Such presentations can be subjective!

- We can measure the strength and type of relationship between two variables
- This will be rigorous and less subjective

Correlation

- A measurement of the strength of the linear relationship between two numerical variables
- Always a number, r, between -1 and 1
- For two data sets $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$

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

- Quite complex – but calculating r is not as important as being able to interpret what r means

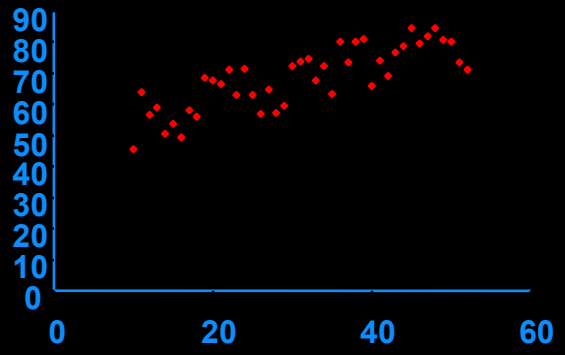
Interpreting correlation

- Sign of r → positive or negative association
  - Positive r (i.e. between 0 and 1) means positive association
  - Negative r means negative association

- Magnitude of r → strength of relationship
  - r close to -1 or 1 means strong linear relationship
  - r close to 0 means weak linear relationship

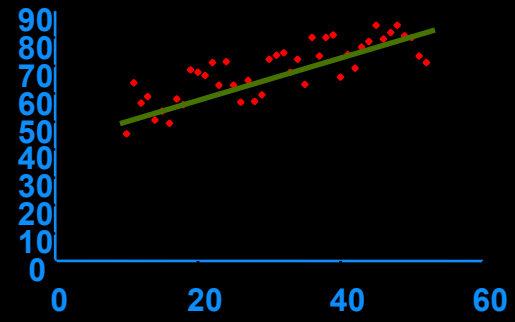
Interpreting correlation – example

- An experiment measures the number of times (out of 100) a subject throws a ball through a hoop relative to the time, in minutes, they practiced
- The correlation turns out to be $r = 0.7$
- This is positive, meaning more practice meant more accuracy
- It is also quite high, meaning that the relationship between practice and accuracy is quite strong



Least-squares regression line

- A description of the type of the linear relationship between two numerical variables
- That is, the line specifies which linear equation best approximates the relationship
- The equation for the line will depend on the sample means and sample standard deviations for the two variables



- It will also depend on the correlation in the data!

Least-squares regression line (cont'd)

- The formula for the approximating line is:

$$y = b_0 + b_1 x$$

where:

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1\overline{x}$$