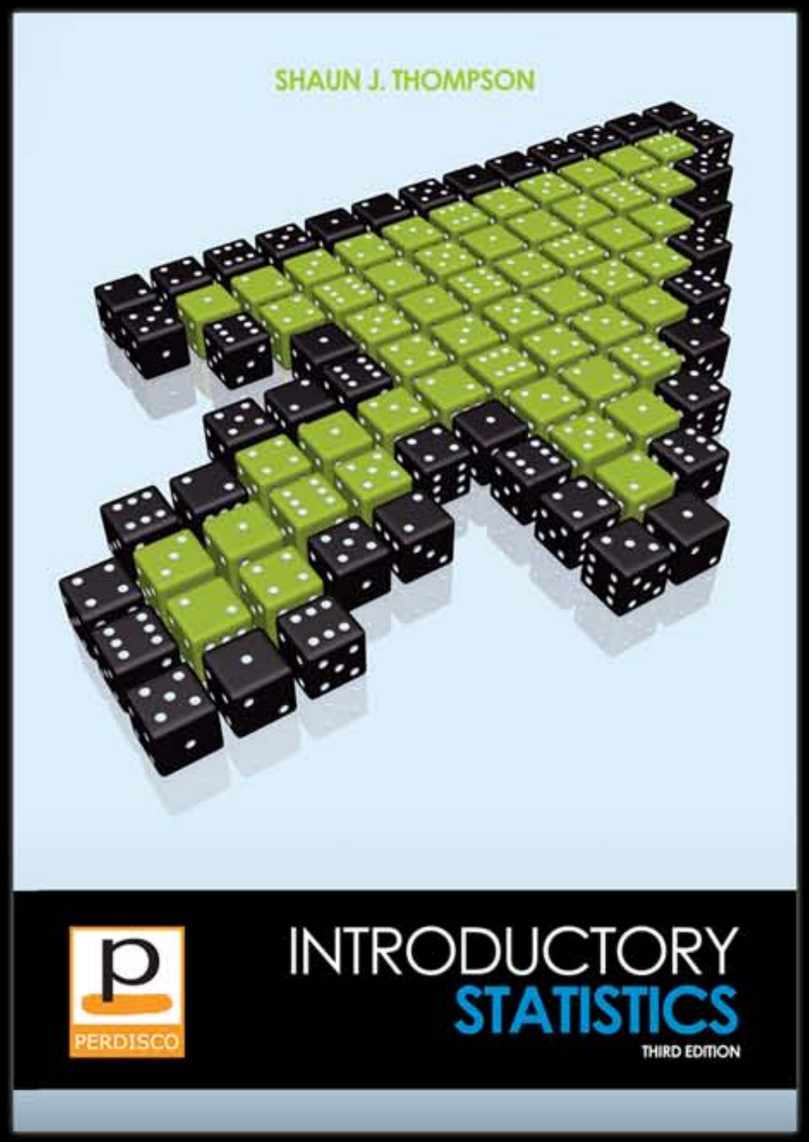


These lecture slides are designed to accompany:

Introductory Statistics, Third Edition

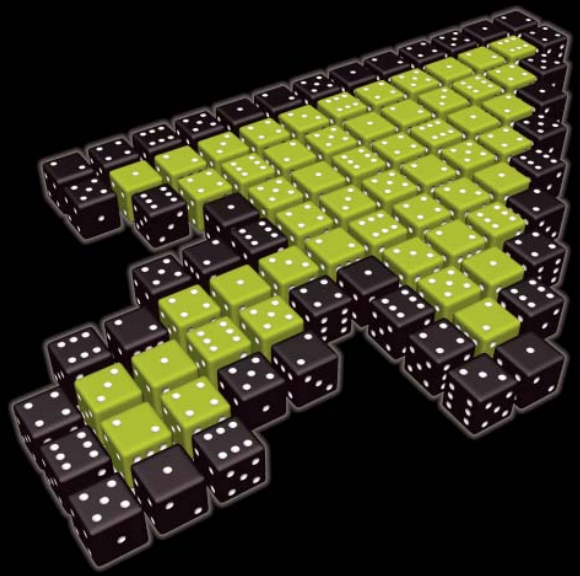
Other features include:

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

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

Chapter 3

Measuring data



Measuring data versus presenting data

- We present data to help us draw meaning from it

- But pictures of data are subjective
- They're also not susceptible to rigorous inference

- **Measuring** data is less subjective, more rigorous

- This is an option when we have numerical data

Measures of center

- ‘Where’ do the data values lie?

- Summarize this by measuring the **center** of the data

- Three common measures of center:
  - mean
  - median
  - mode

Mean

- The mathematical average of a set of values
- Example: Data values are {13, 8, 10, 12, 12}
- Mean is:

$$\bar{x} = \frac{13+8+10+12+12}{5}$$

$$= \frac{55}{5}$$

$$= 11$$

- So the mean is 11
- Note that the mean is sensitive to all data values

Median

- The ‘middle’ value if values are ranked in order

- Example: Data values are {13, 8, 10, 12, 12}
- In order, these are: 8, 10, 12, 12, 13

- Median is third value: 12

- If number of data points is even, there are two middle values – median is average of these

Mode

- The data value with highest observed frequency
- Example: Data values are {13, 8, 10, 12, 12}
- The value 12 occurs twice – this is the mode
- There may be more than one mode

Comparing measures of center

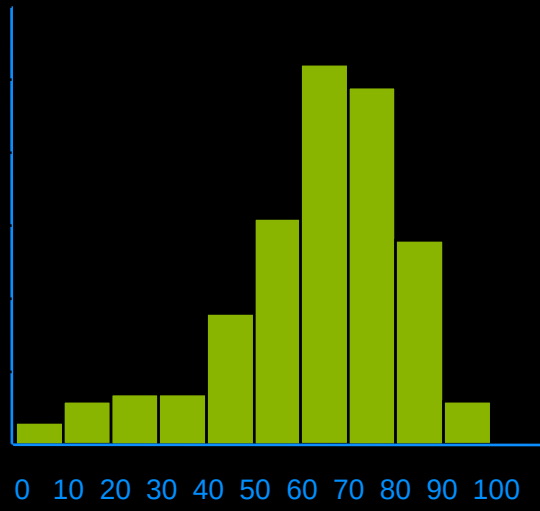
- Mean and median are very commonly used (the mode is not so common)
- Mean and median can be very different
- Example: Consider data set {0, 0, 0, 0, 50}
- Mean is 10, median is 0

Comparing measures of center (cont'd)

- Why do the mean and median differ so much?
- Because 50 is an outlier!
- The mean is sensitive to outliers, the median isn't
- This is important when considering which measure of center to use!

Measures and pictures of center

- How does measuring data compare with presenting data?
- Consider this histogram
- From this diagram, we can tell that the mean is less than the median



- Why? The mean is skewed by small values
- Measures of data make our ‘feeling’ of the data from the presentation more rigorous

Measures of variation

- Measuring center tells us ‘where’ the data are
- Variation tells us how ‘spread out’ they are
- The more spread out, the less consistent they are and the less we can conclude from them
- So variation tells us how ‘precise’ we can be

Range

- The simplest measure of variation

- *Range = maximum value – minimum value*

- Example: Test scores {87, 99, 65, 52, 41, 73, 67}
- Maximum value = 99, minimum value = 41
- Range = 99 – 41 = 58

Quartiles

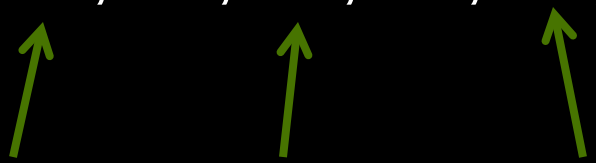
- There is a ‘better’ measure of the range that is less sensitive to outliers
- But to define it we need to define the three quartiles, three values that split data set into four quarters

- First quartile Q_1: middle of the lower half of values
- Second quartile Q_2: middle value, i.e. median
- Third quartile Q_3: middle of higher half of values

Quartiles – example

- More precisely, Q_1 is the 25^{th} percentile – that is, 25% of the data values fall below it, 75% above it
- Similarly, Q_2 is the 50^{th} percentile and Q_3 is the 75^{th}

- Example: Test scores {87, 99, 65, 52, 41, 73, 67}
- In order, this is {41, 52, 65, 67, 73, 87, 99}



$Q_1 = 52$, $Q_2 = 67$, $Q_3 = 87$

Inter-quartile range (IQR)

- A modified range: the range of the ‘middle’ values

- IQR = *third quartile – first quartile*

- Example: Test scores {87, 99, 65, 52, 41, 73, 67}
- $Q_1 = 52$ and $Q_3 = 87$
- IQR = 87 – 52 = 35
- IQR is less sensitive to outliers than the range is

Five-number summary

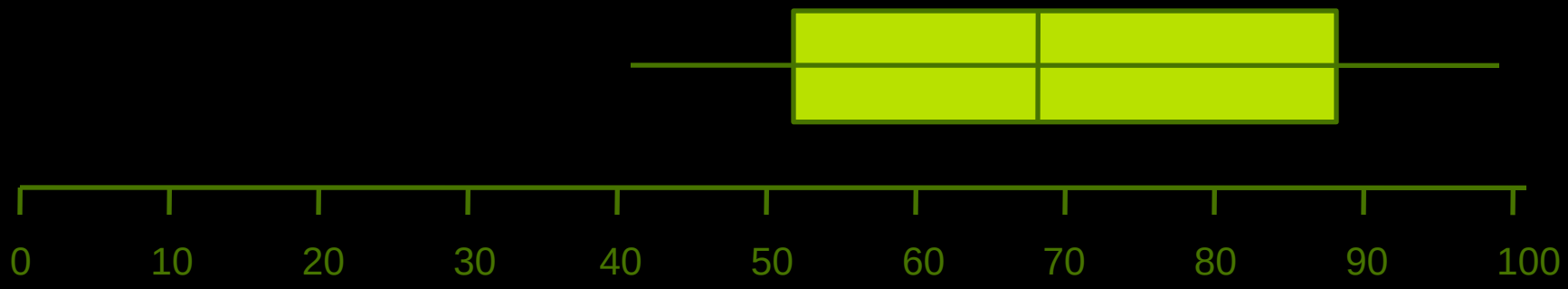
- Set of five key values in data set:
  - minimum value
  - first quartile
  - median (i.e. second quartile)
  - third quartile
  - maximum value

- These values give a concise description of both the variation in the data and the center of the data

Box plot

- A graphical presentation of five-number summary

- Example: Test scores {87, 99, 65, 52, 41, 73, 67}
- min = 41, Q_1 = 52, Q_2 = 67, Q_3 = 87, max = 99



Variance and standard deviation

- Range, IQR, five number summary and box plots all give some indication of variation in data
- But they are all fairly rudimentary measures

- Variance, s^2, and standard deviation, s, are the most common measures of variation in inference
- Both are sensitive to all data values, as the mean is
- Both measure the ‘average’ distance from the values in the data set to the middle of that data set

Variance and standard deviation (cont'd)

- We use the mean, $\overline{x}$, as our measure of the 'middle' of the data set
- For n sample data points $\{x_1,\ldots,x_n\}$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \qquad s = \sqrt{s^2}$$

Measures of a population

- 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

- For example, we may find that the average I.Q. in a survey of 50 people is 103
- This is a sample mean – but we probably want to know the population mean!

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

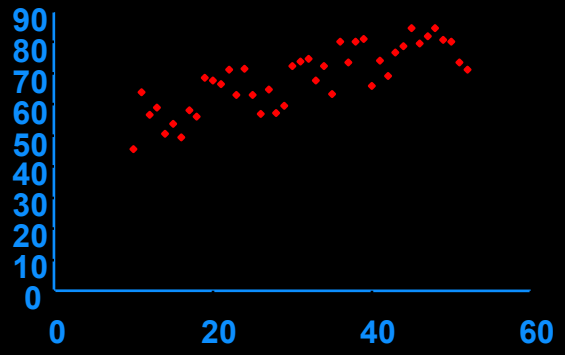
Interpreting correlation

- 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

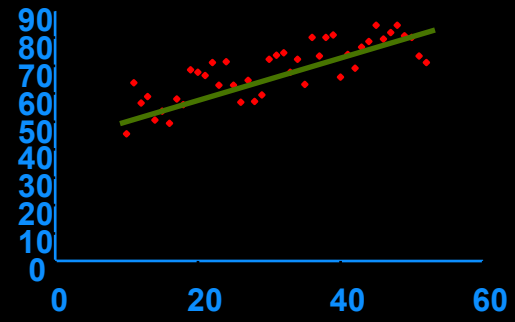
Interpreting correlation – example

- 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:

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