February 16

- Today
  - A couple of remaining items from chapter 1, then chapter 2
  - Dataframes in R
- Practical tomorrow
  - Practice w/ dataframes
    - TAs will verify that you've worked on this activity
- Homework
  - More dataframes practice; practice with dispersion, fit, and confidence intervals
    - due next Thursday before lecture

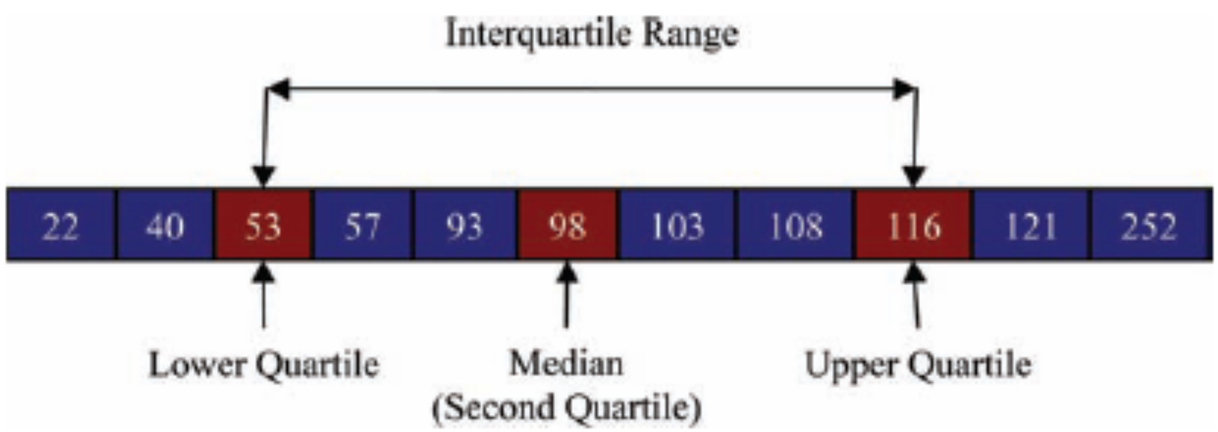
The Dispersion: Range

- The Range
  - The smallest score subtracted from the largest
- Example
  - Number of friends of 11 Facebook users.
  - 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252
  - Range = 252 – 22 = 230
  - Very biased by outliers

The Dispersion: The Interquartile range

- Quartiles (one type of quantile)
  - The three values that split the sorted data into four equal parts.
  - Second quartile = median.
  - Lower quartile = median of lower half of the data.
  - Upper quartile = median of upper half of the data.

FIGURE 1.7 Calculating quartiles and the interquartile range



Chapter 2: Everything You Ever Wanted to Know about Statistics

Aims and Objectives

- Know what a statistical model is and why we use them.
  - The mean
- Know what the 'fit' of a model is and why it is important.
  - The standard deviation
- Distinguish models for samples and populations

The Research Process

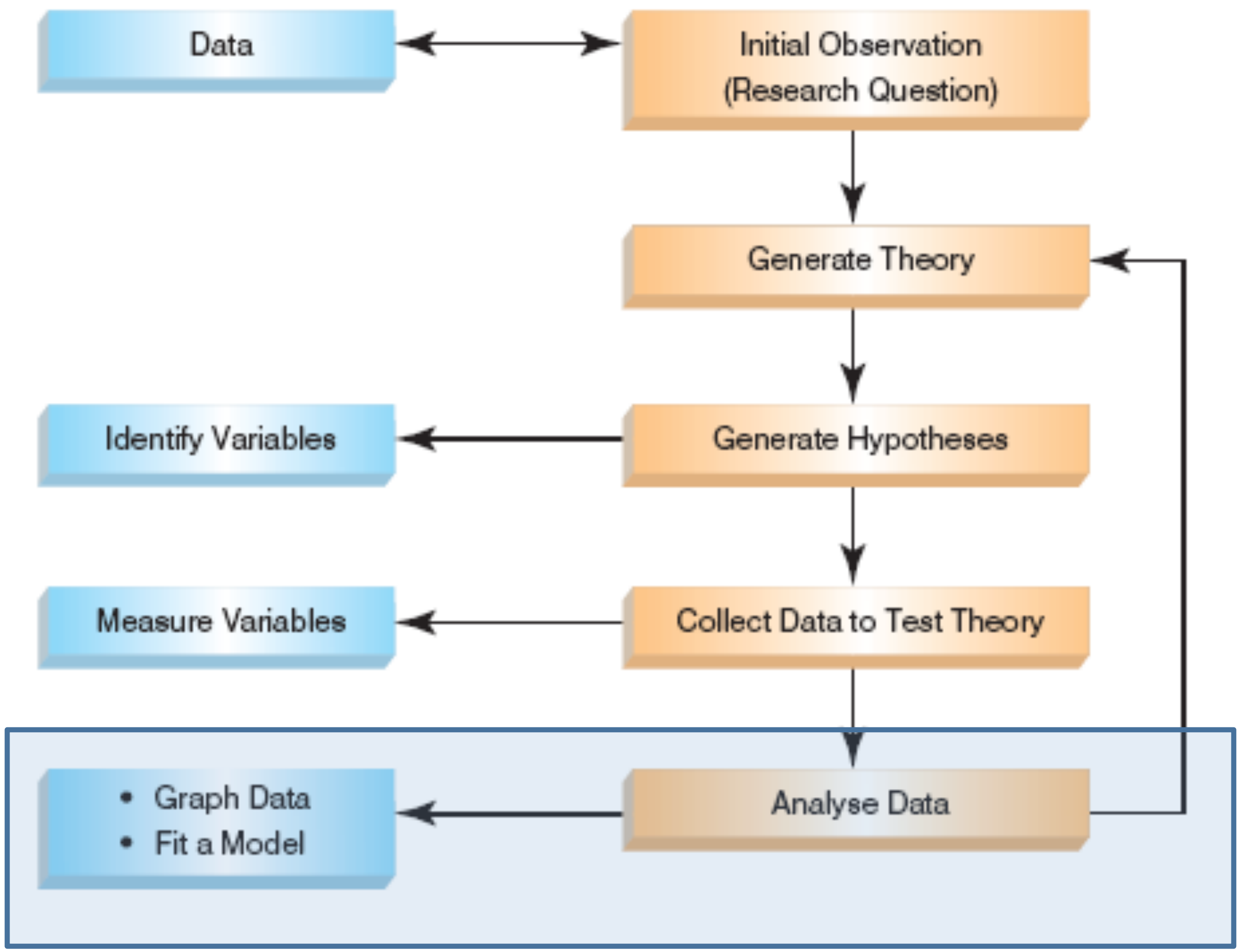


FIGURE 1.2 The research process

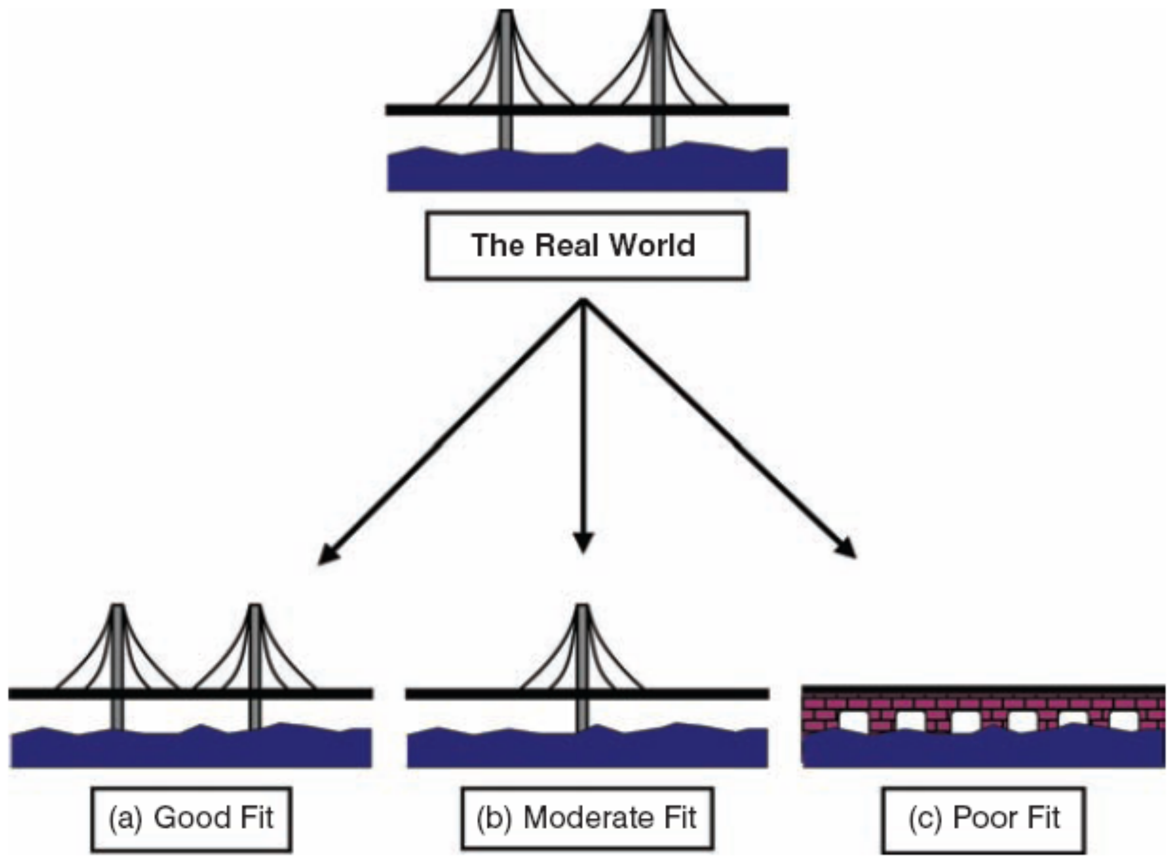
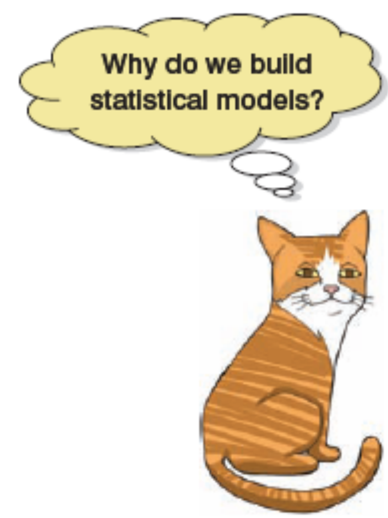


FIGURE 2.2
Fitting models to real-world data (see text for details)



Populations and Samples

- Population
  - The collection of units (be they people, plankton, plants, cities, etc.) to which we want to generalize a set of findings or a statistical model
    - usually unmeasurable
- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population

Populations and Samples: Example

- 1% of general population is made up of narcissists
  - Test everyone in the population?
    - No!
    - Found a representative sample of people, tested them, and then generalized to the whole population

The Only Equation You Will Ever Need

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

A Simple Statistical Model

- In statistics we fit models to our sample data (i.e. we use a statistical model to represent what is happening in the real world, [the population]).
- The mean is a hypothetical value (i.e. it doesn't have to be a value that actually exists in the data set).
- As such, the mean is a simple statistical model.

The Mean: Example

- Collect some data:

1, 3, 4, 3, 2

- Add them up:

$$\sum_{i=1}^{n} x_i = 1 + 3 + 4 + 3 + 2 = 13$$

- Divide by the number of observations, *n*:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{13}{5} = 2.6$$

The mean as a model

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

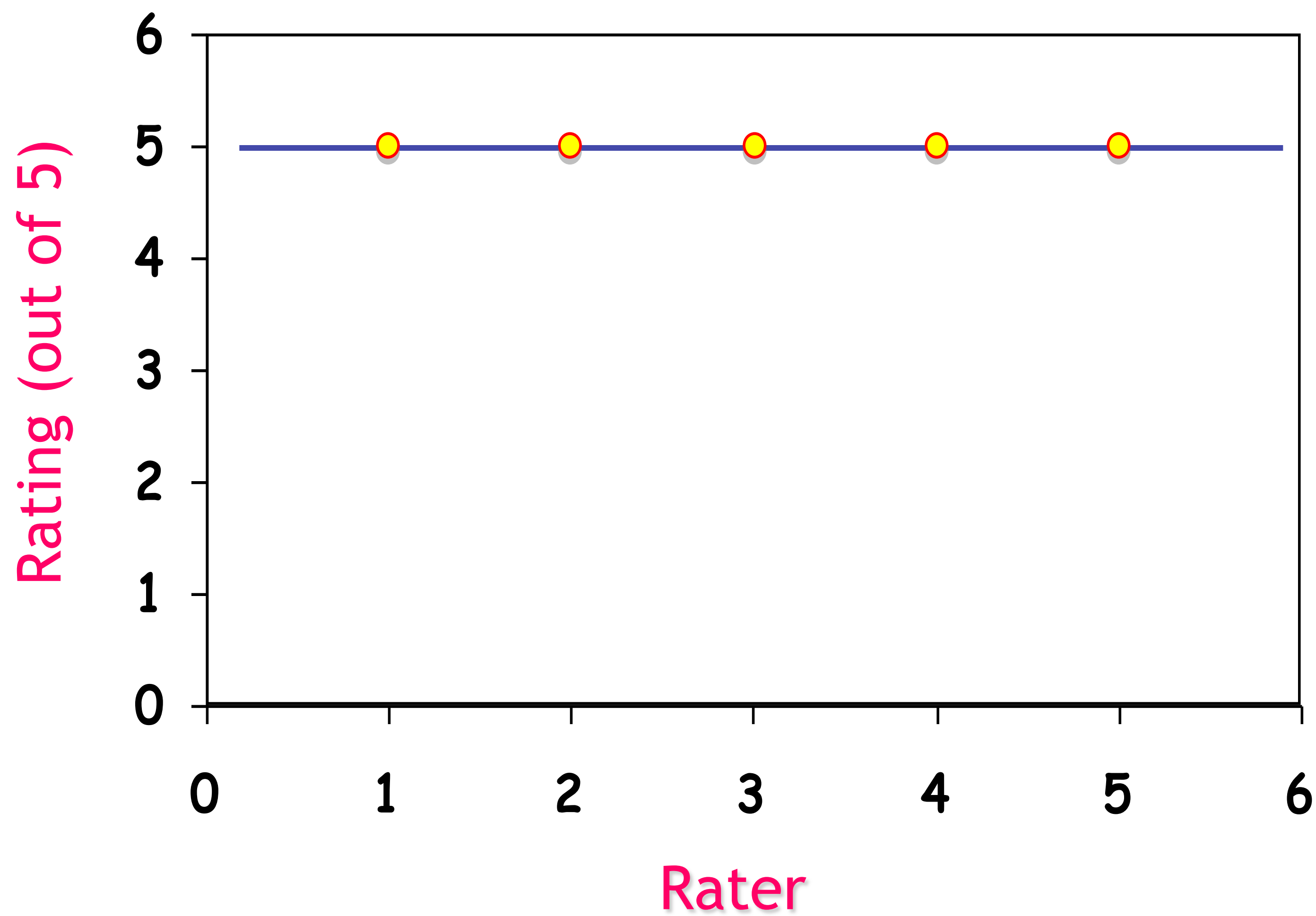
$$\text{outcome}_{\text{lecturer}1} = (\bar{X}) + \text{error}_{\text{lecturer}1}$$

$$1 = 2.6 + \text{error}_{\text{lecturer}1}$$

Measuring the ‘Fit’ of the Model

- How can we assess how well the mean represents reality?

A Perfect Fit

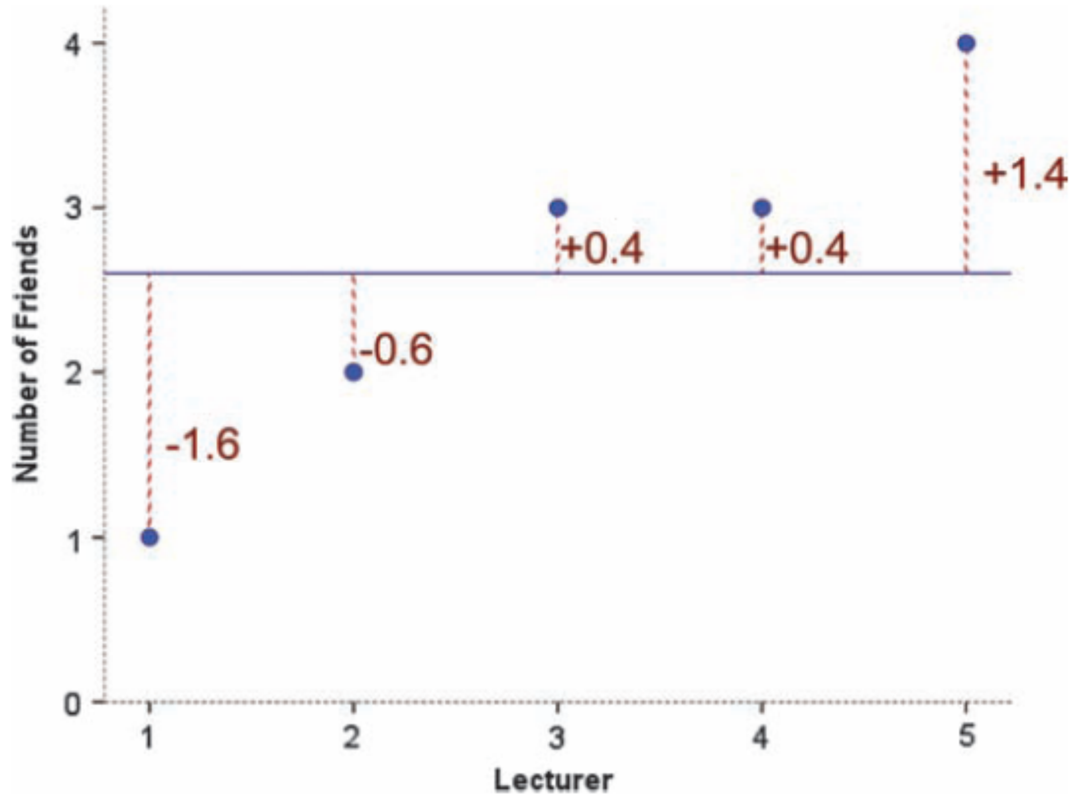


Calculating 'Error'

- A deviation (or error) is the difference between the mean and an actual data point.
- Deviations can be calculated by taking each score and subtracting the mean from it:

$$\text{deviation} = x_i - \bar{x}$$

FIGURE 2.4
Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends



Use the Total Error?

- We could just take the error between the mean and the data and add them.

Score	Mean	Deviation
1	2.6	-1.6
2	2.6	-0.6
3	2.6	0.4
3	2.6	0.4
4	2.6	1.4
	Total =	0

$$\boxed{\sum(X-\bar{X})=0}$$

Sum of Squared Errors

- We could add the deviations to find out the total error.
- Deviations cancel out because some are positive and others negative.
- Therefore, we square each deviation.
- If we add these squared deviations we get the sum of squared errors (*SS*).

Score	Mean	Deviation	Squared Deviation
1	2.6	-1.6	2.56
2	2.6	-0.6	0.36
3	2.6	0.4	0.16
3	2.6	0.4	0.16
4	2.6	1.4	1.96
		Total	5.20

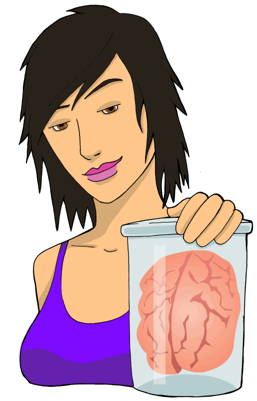
$$\boxed{SS = \sum (X - \bar{X})^2 = 5.20}$$

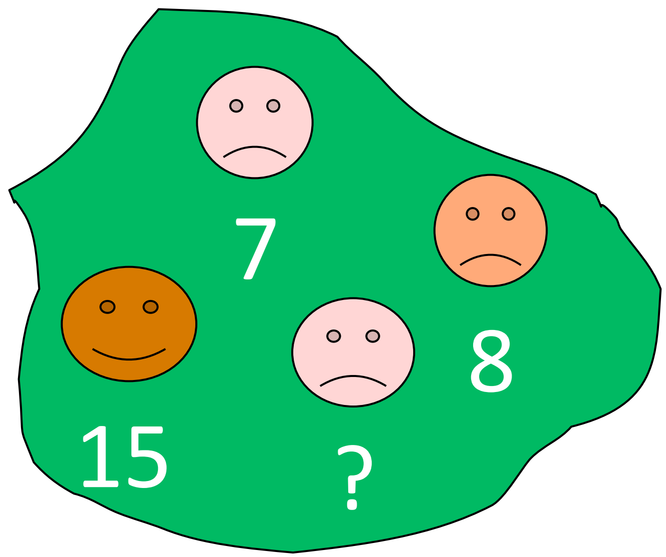
Variance

- The sum of squares is a good measure of overall variability, but is dependent on the number of scores.
- We calculate the average variability by dividing by the number of scores (*n*).
- This value is called the variance (s^2).

$$\text{variance } (s^2) = \frac{\text{SS}}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.3$$

Degrees of Freedom





$$\bar{X} = 10$$

Standard Deviation

- The variance has one problem: it is measured in units squared. (1.3 friends squared???)
- This isn't a very meaningful metric so we take the square root value.
- This is the standard deviation (s).

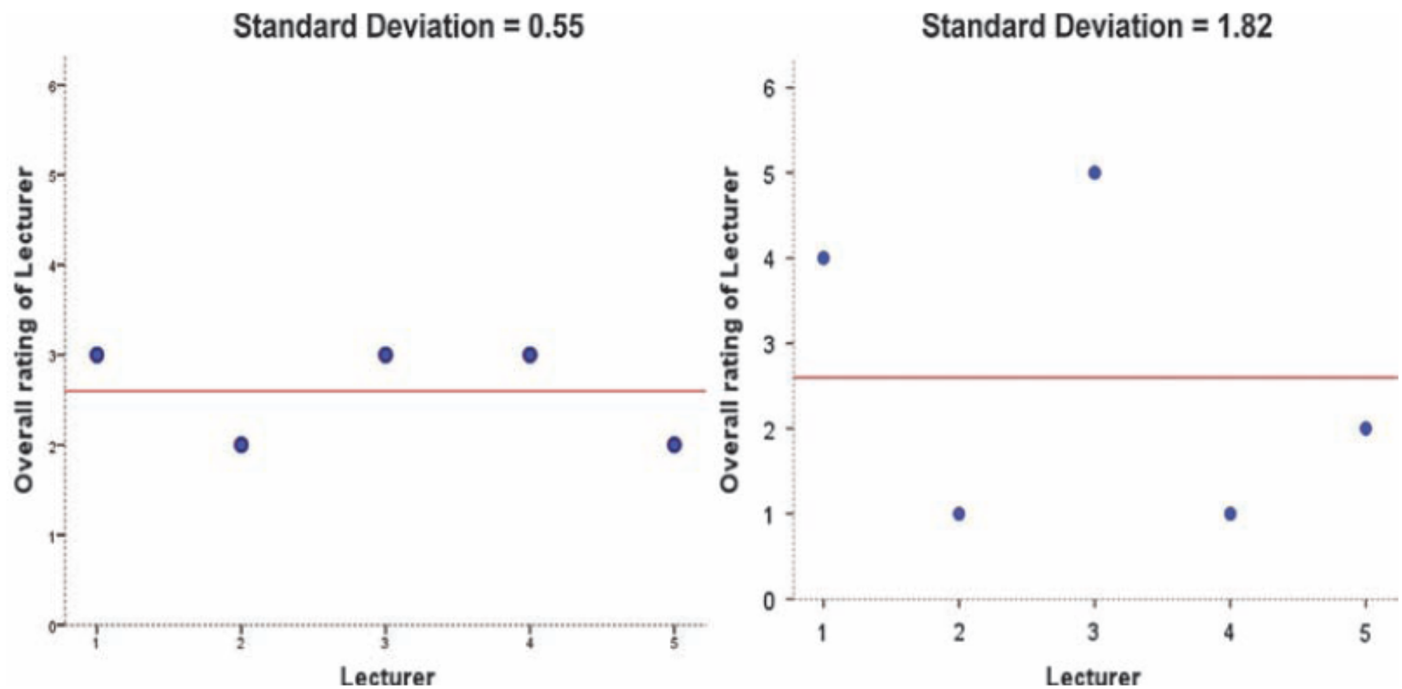
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5.20}{4}} = 1.14$$

Important Things to Remember

- The sum of squares, variance, and standard deviation represent the same thing:
  - The 'fit' of the mean to the data
  - The variability in the data
  - How well the mean represents the observed data
  - Error

Same Mean, Different SD

FIGURE 2.5
Graphs illustrating data that have the same mean but different standard deviations



The SD and the Shape of a Distribution

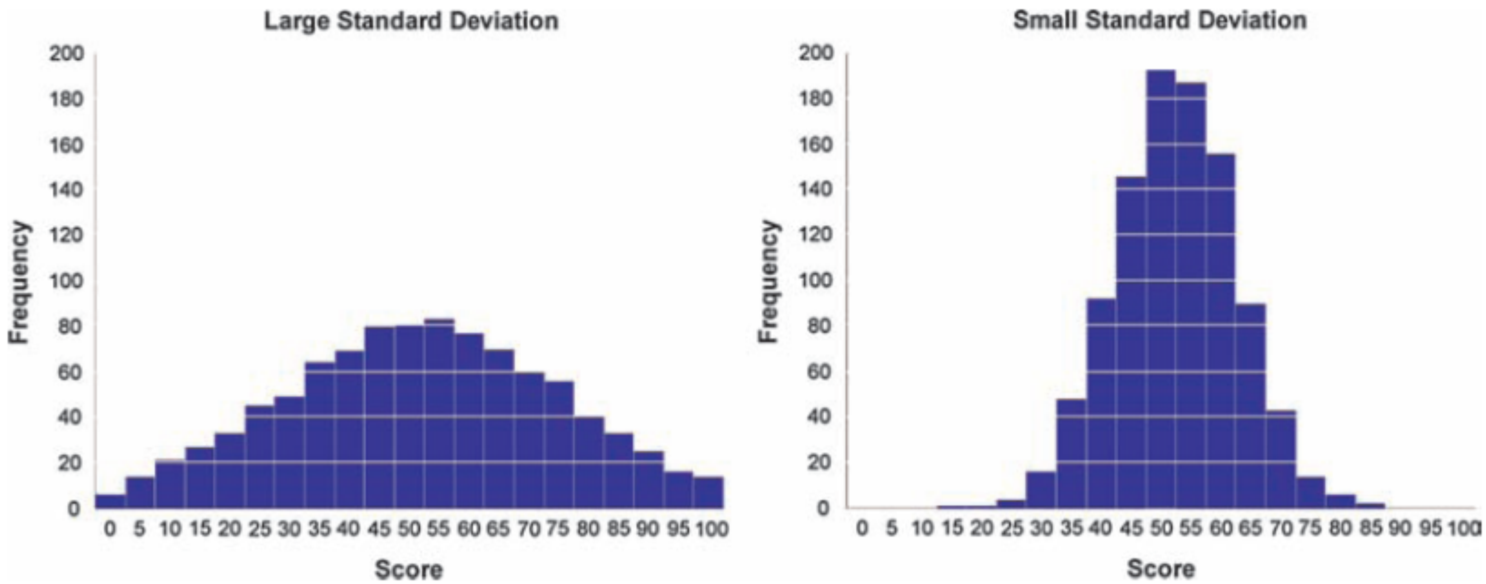
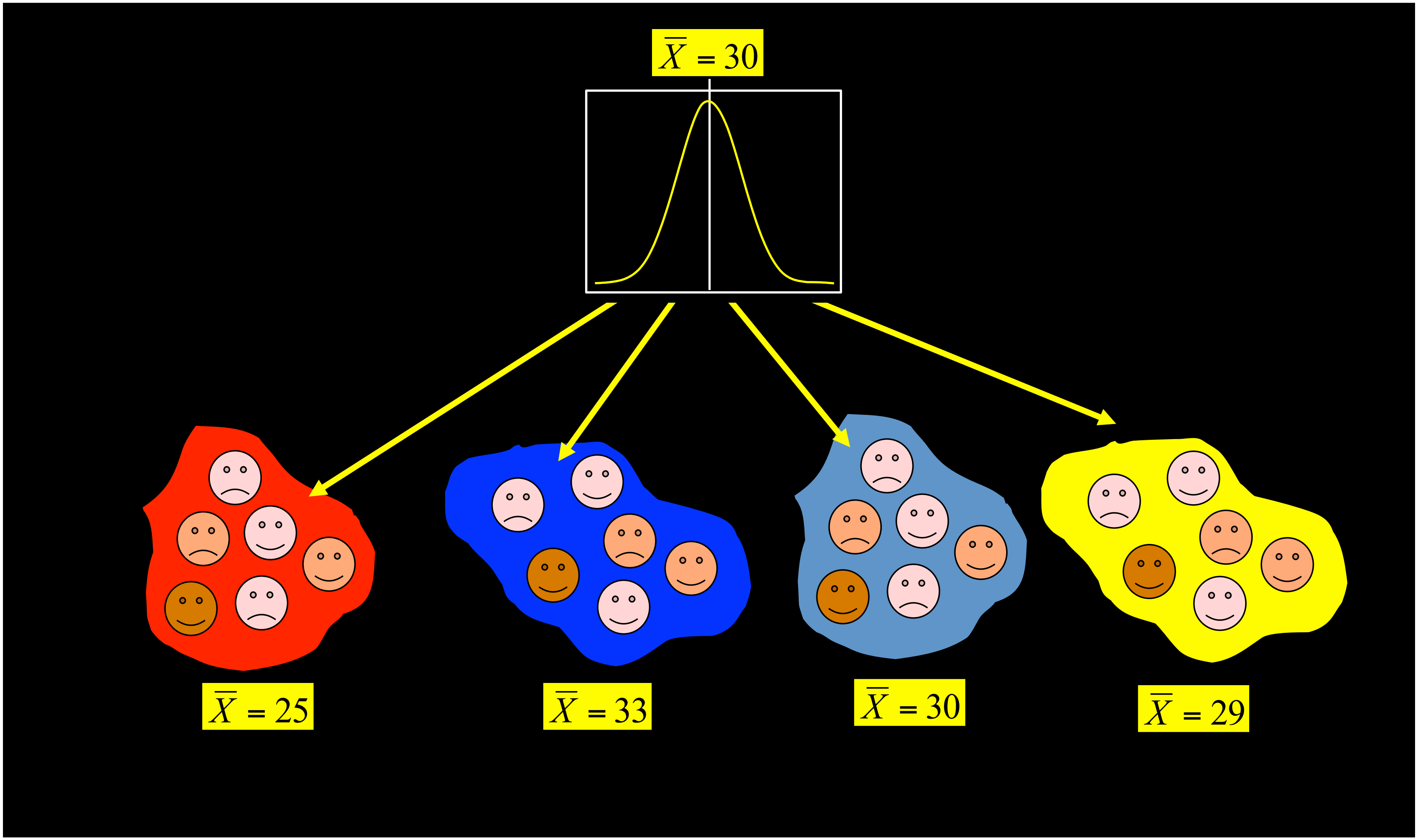


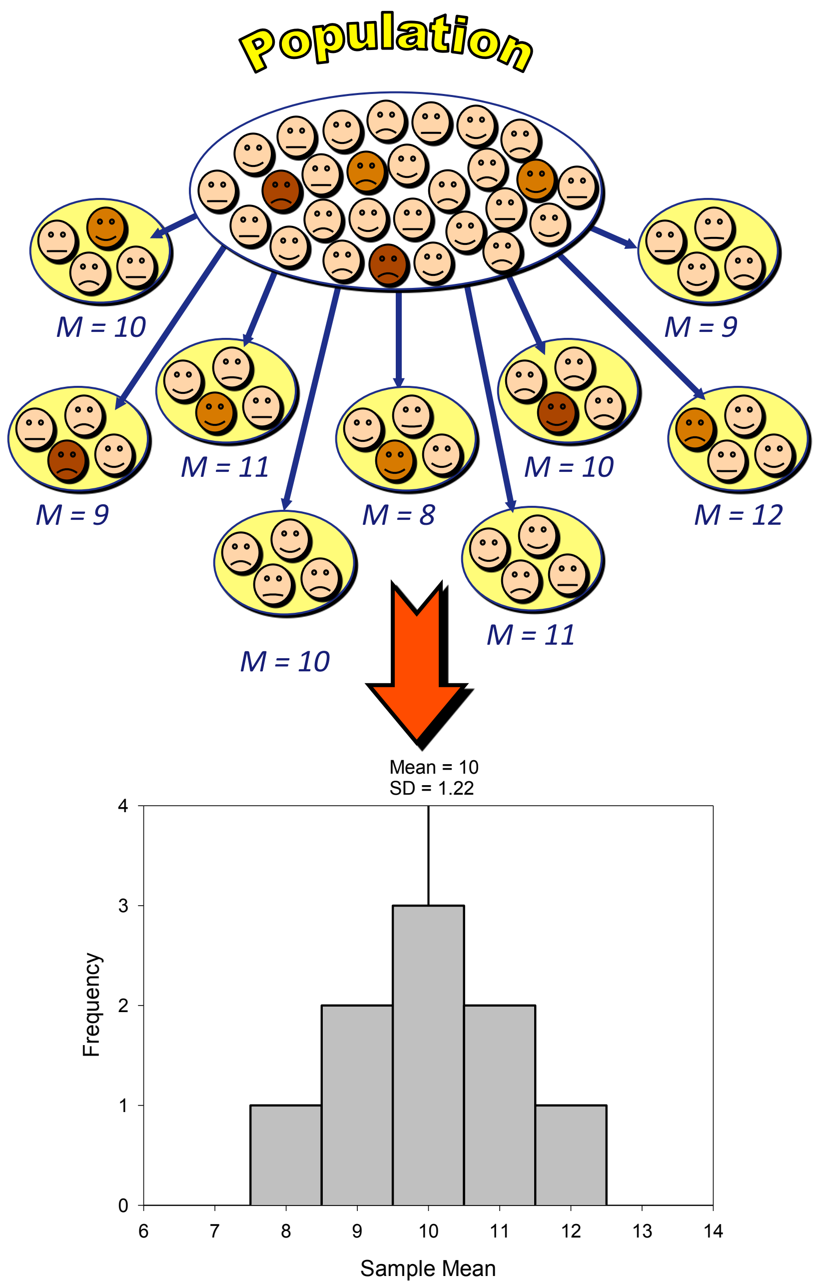
FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

Samples vs. Populations

- Sample
  - Mean and SD describe only the sample from which they were calculated.
- Population
  - Mean and SD are intended to describe the entire population.
- Sample to Population:
  - Mean and SD are obtained from a sample, but are used to estimate the mean and SD of the population.



$\overline{X} = 30$
$\overline{X} = 25$
$\overline{X} = 33$
$\overline{X} = 30$
$\overline{X} = 29$



$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$