DISCOVERING STATISTICS USING R

Everything You Ever Wanted to Know about Statistics

Prof. Donald Davendra

ANDY FIELD

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

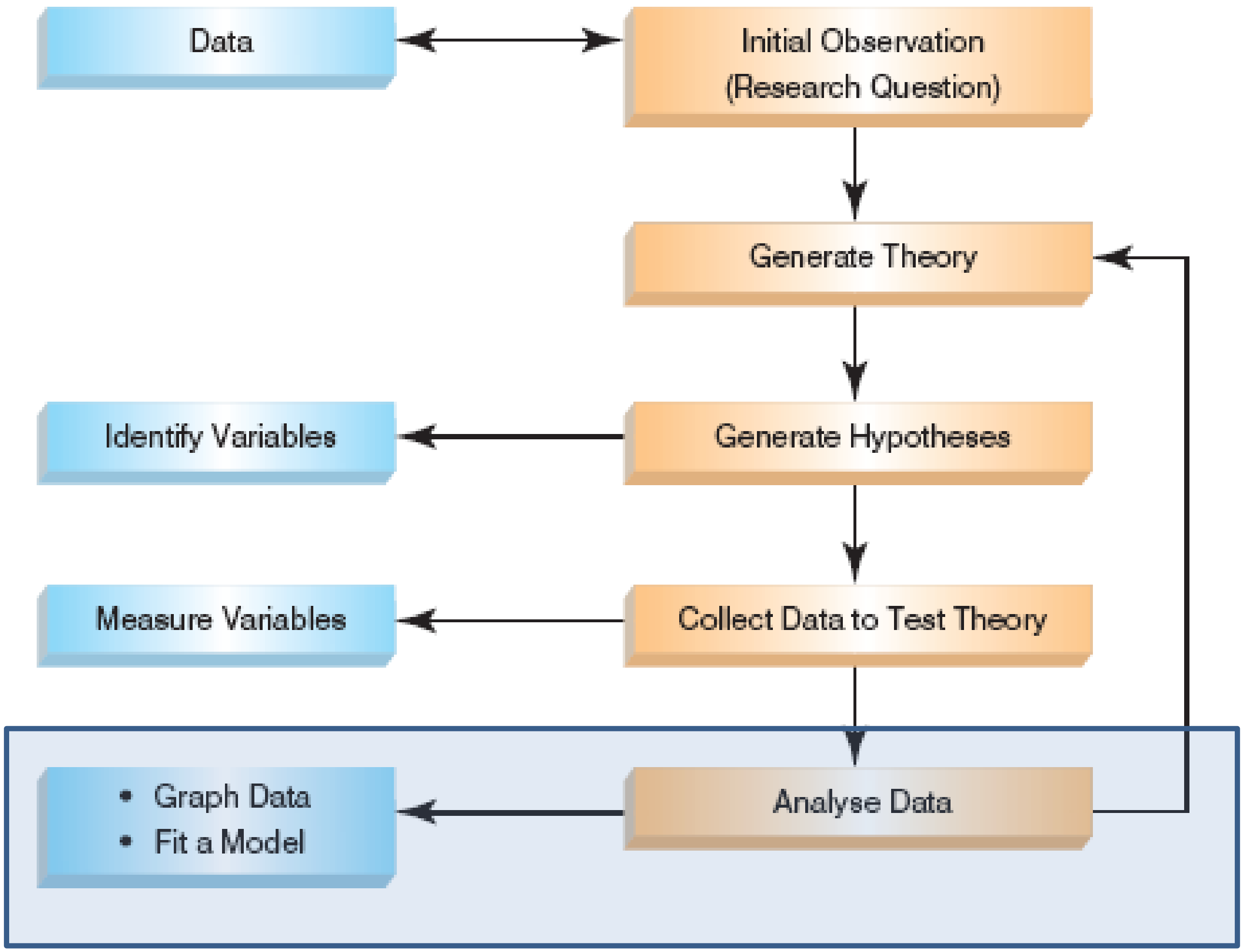


FIGURE 1.2
The research process

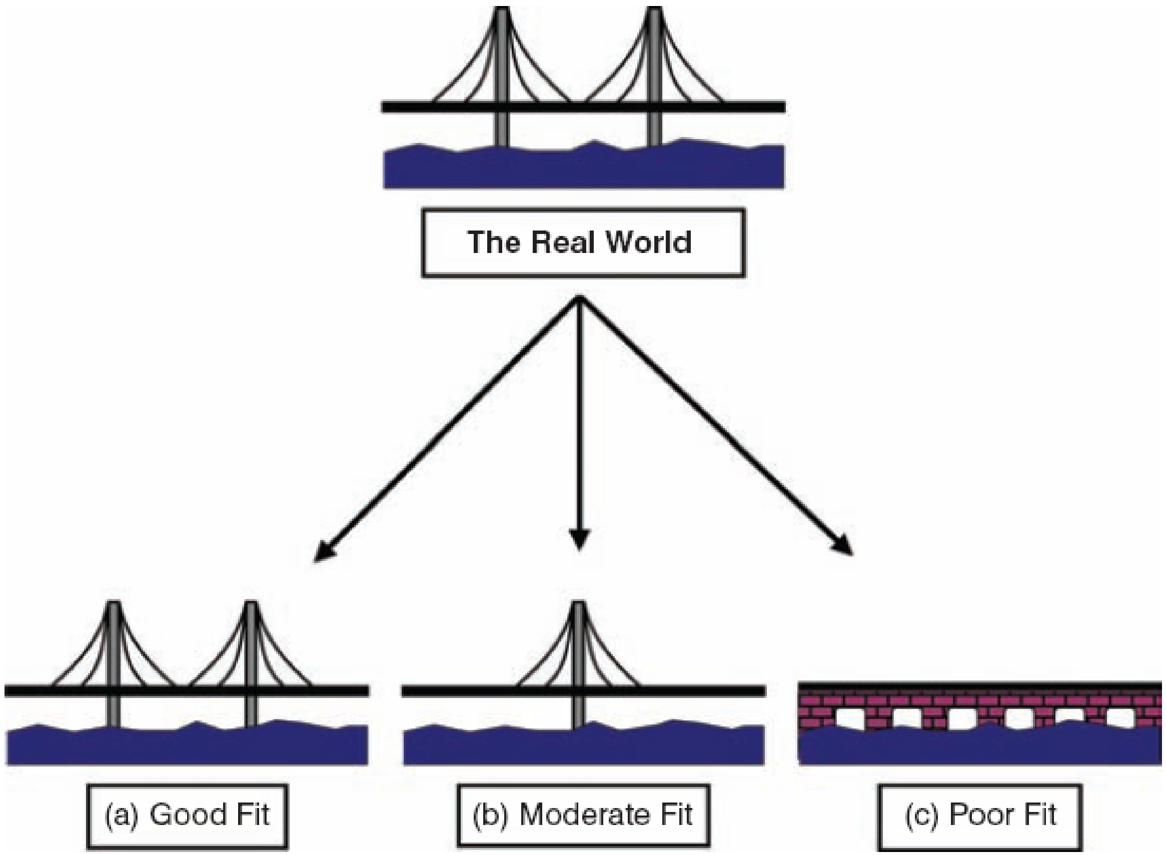
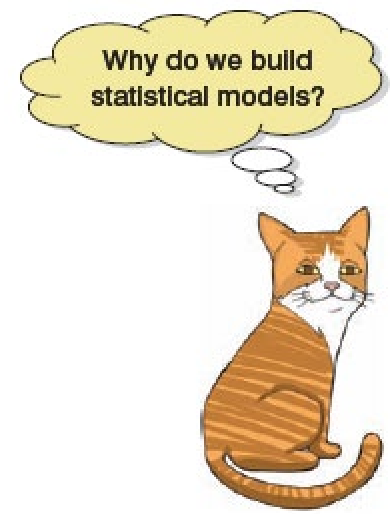


FIGURE 2.2 Fitting models to real-world data (see text for details)



Populations and Samples

- Population
  - The collection of units (be they people, plankton, plants, cities, suicidal authors, etc.) to which we want to generalize a set of findings or a statistical model
- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that 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 data (i.e. we use a statistical model to represent what is happening in the real world).
- 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 simple statistical model.

The Mean

- The mean is the sum of all scores divided by the number of scores.
- The mean is also the value from which the (squared) scores deviate least (it has the least error).

$$\text{mean}\ (\bar{X}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

DISCOVERING STATISTICS USING R

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 scores, *n*:

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

ANDY FIELD

DISCOVERING STATISTICS USING R

The mean as a model

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

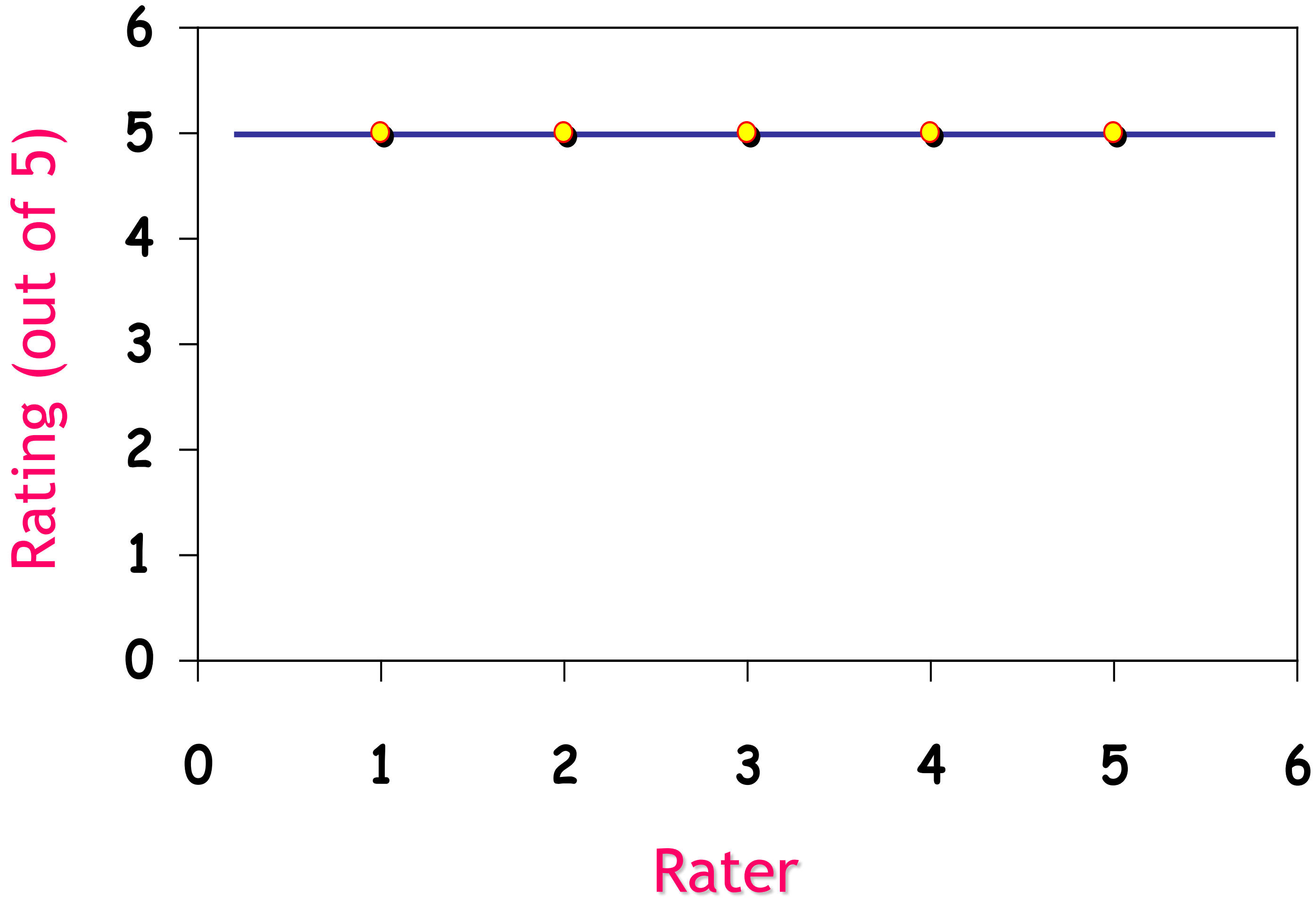
$$\text{outcome}_{\text{lecturer1}} = (\bar{X}) + \text{error}_{\text{lecturer1}}$$

$$1 = 2.6 + \text{error}_{\text{lecturer1}}$$

Measuring the 'Fit' of the Model

- The mean is a *model* of what happens in the real world: the *typical* score.
- It is not a perfect representation of the data.
- How can we assess how well the mean represents reality?

A Perfect Fit



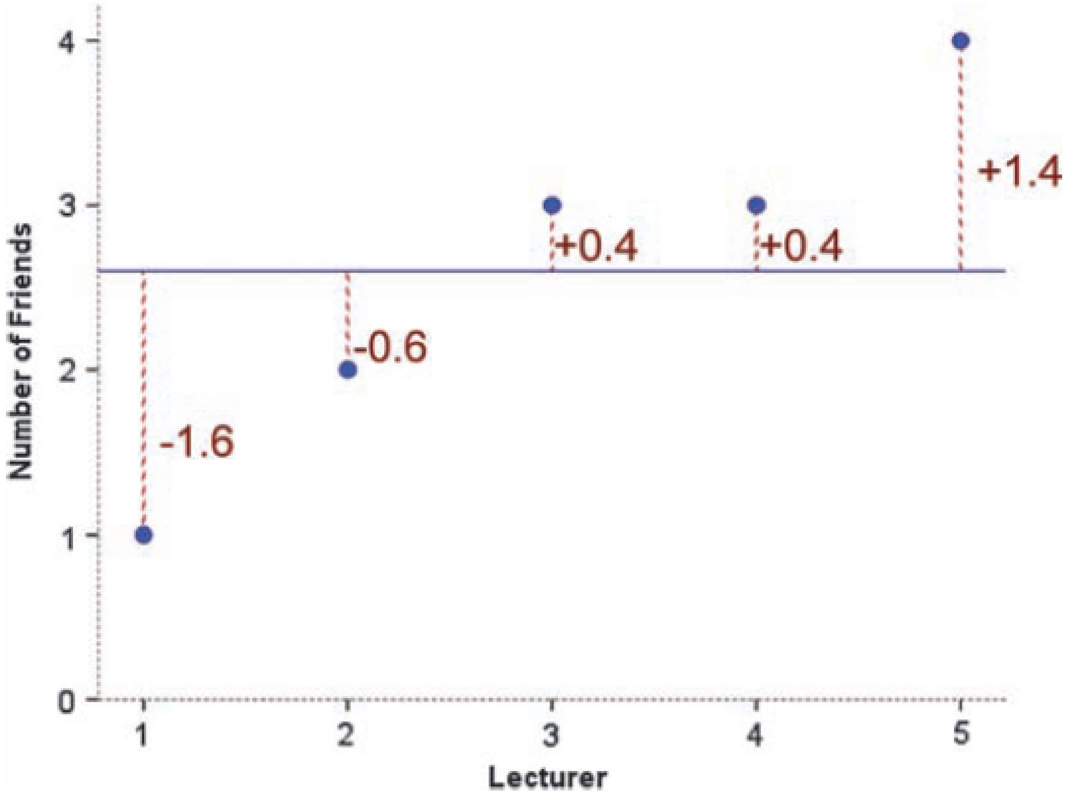
Calculating 'Error'

- A deviation 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:

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

DISCOVERING STATISTICS USING R

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

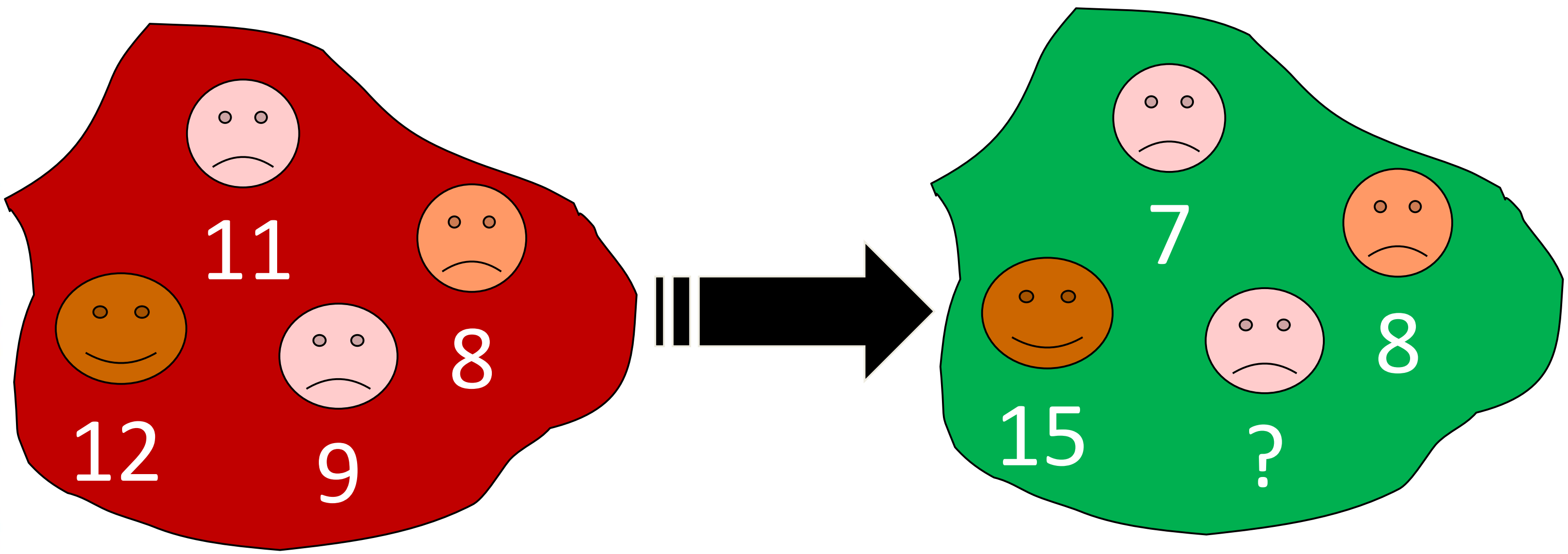
- The sum of squares is a good measure of overall variability, but is dependent on the number of scores.
- We calculate the average variability by dividing by the number of scores (*n*).
- This value is called the variance (s^2).

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

Degrees of Freedom

Sample

Population



$\bar{X} = 10$

$\mu = 10$

Standard Deviation

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

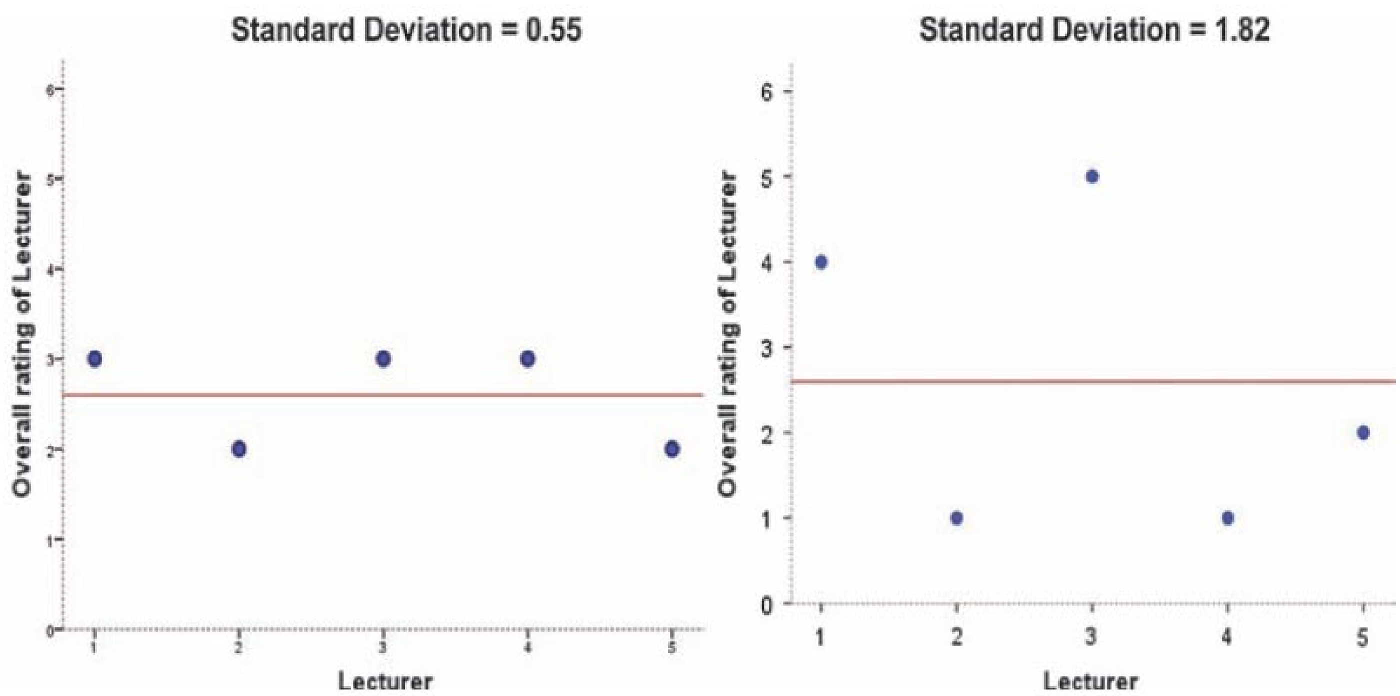
$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{5.20}{5}} = 1.02$$

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

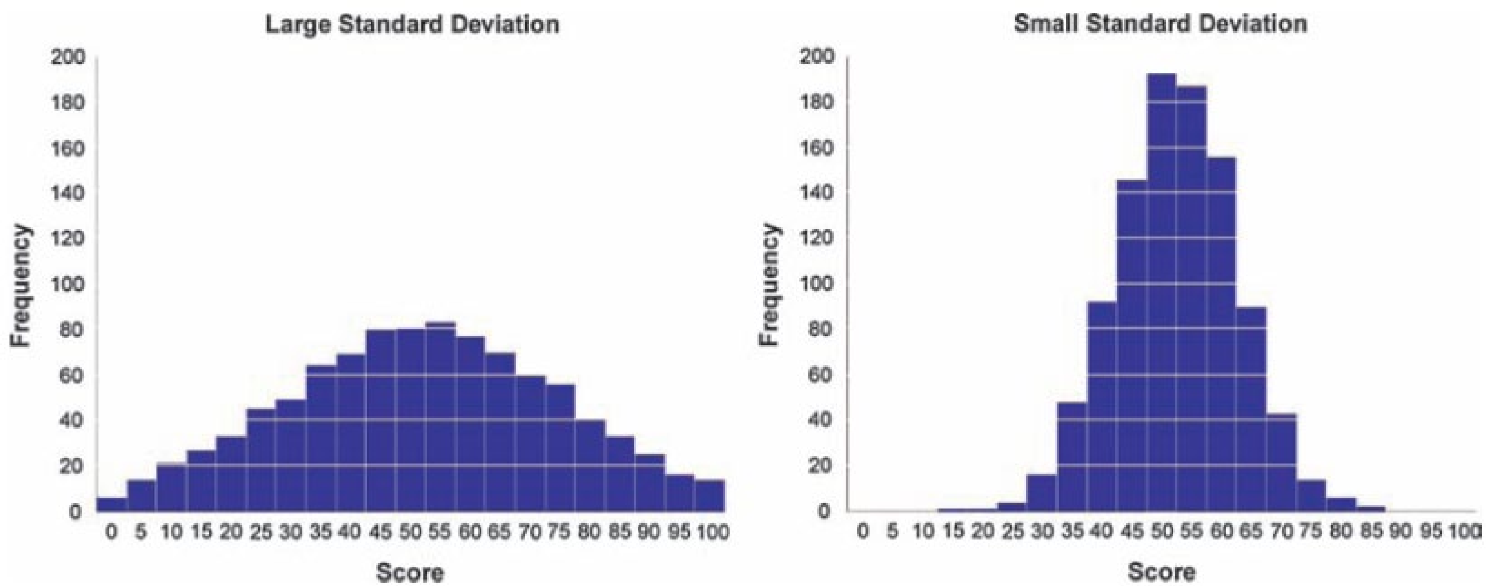
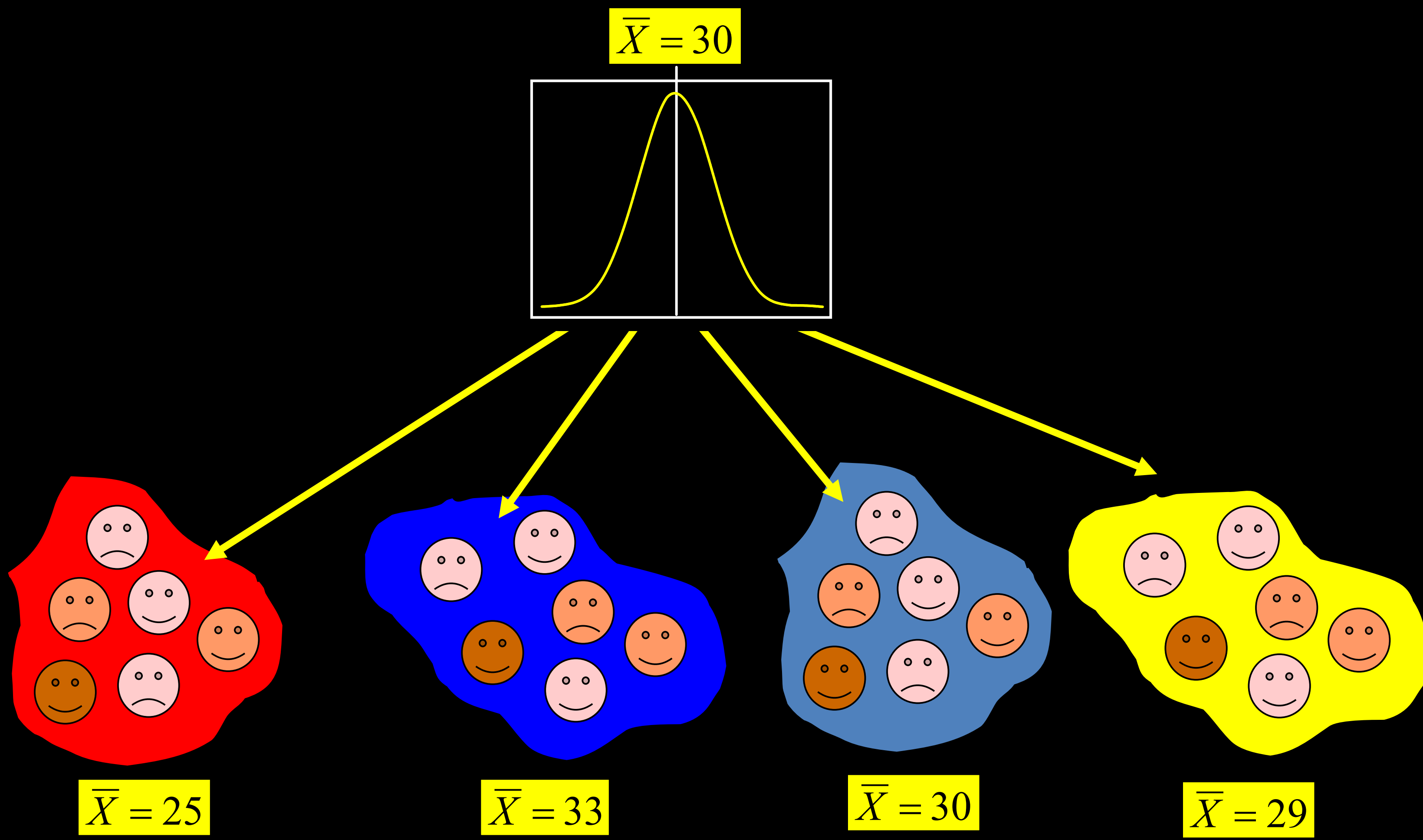


FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

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 (very rare in psychology).
- Sample to Population:
  - Mean and SD are obtained from a sample, but are used to estimate the mean and SD of the population (very common in psychology).



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

Standard Error

- Sampling Variation
  - When different samples are taken from a population, each sample has a unique mean value.
- Sampling Distribution
  - A plot of all the sampling means
- Standard Error of the Mean (SE)
  - The standard error of the sample means

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

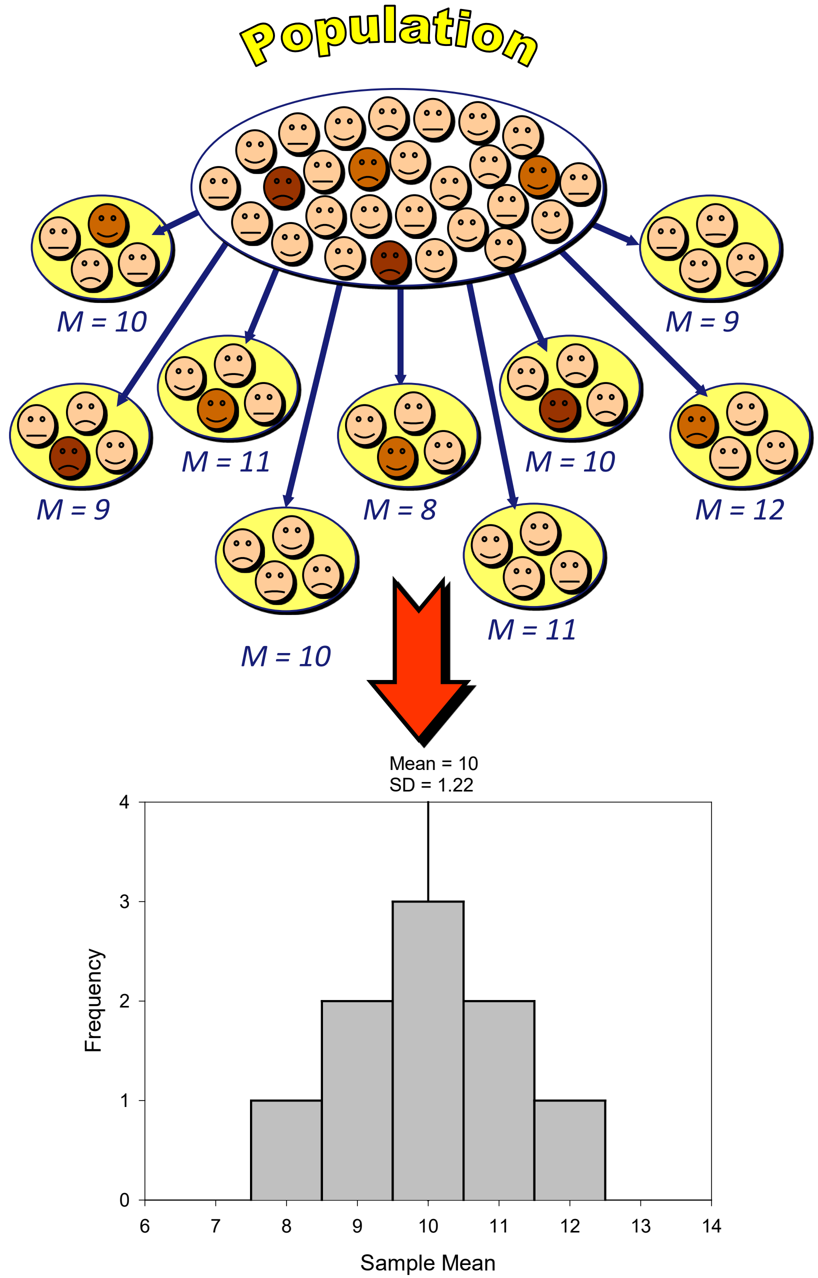
  - Known as the Central Limit Theorem

Standard Error

- Standard Error of the Mean (SE)
  – The standard error is the standard deviations of the sample means:

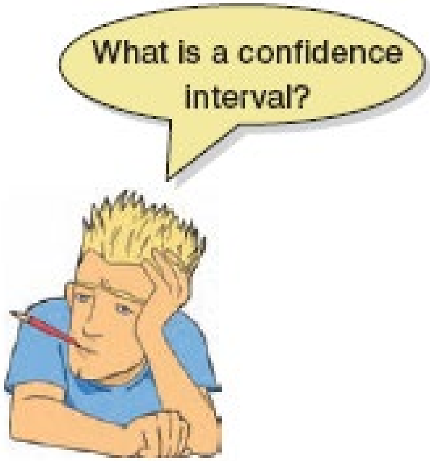
$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

- Central Limit Theorem
  – If the sample is large enough (> 30), the above equation can approximate the standard error.
  – If sample < 30, then the sampling distribution has a different shape, known as a ***t* – distribution.**



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

Confidence Intervals

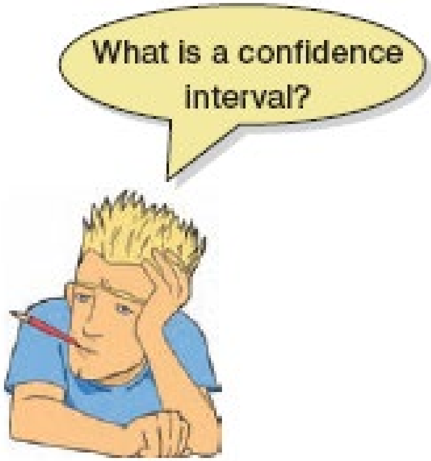


- Sample Mean
  - as an estimate of the value of the populati…
  - Different samples = different means
  - Standard error to see deviation in sample mean

- Confidence Interval
  - A different approach to set the boundaries within which we can assume the true mean to fall within.
  - This is called confidence interval (CI)

$$lower\ boundary = \bar{X} - (1.96 \times SE)$$
$$upper\ boundary = \bar{X} + (1.96 \times SE)$$

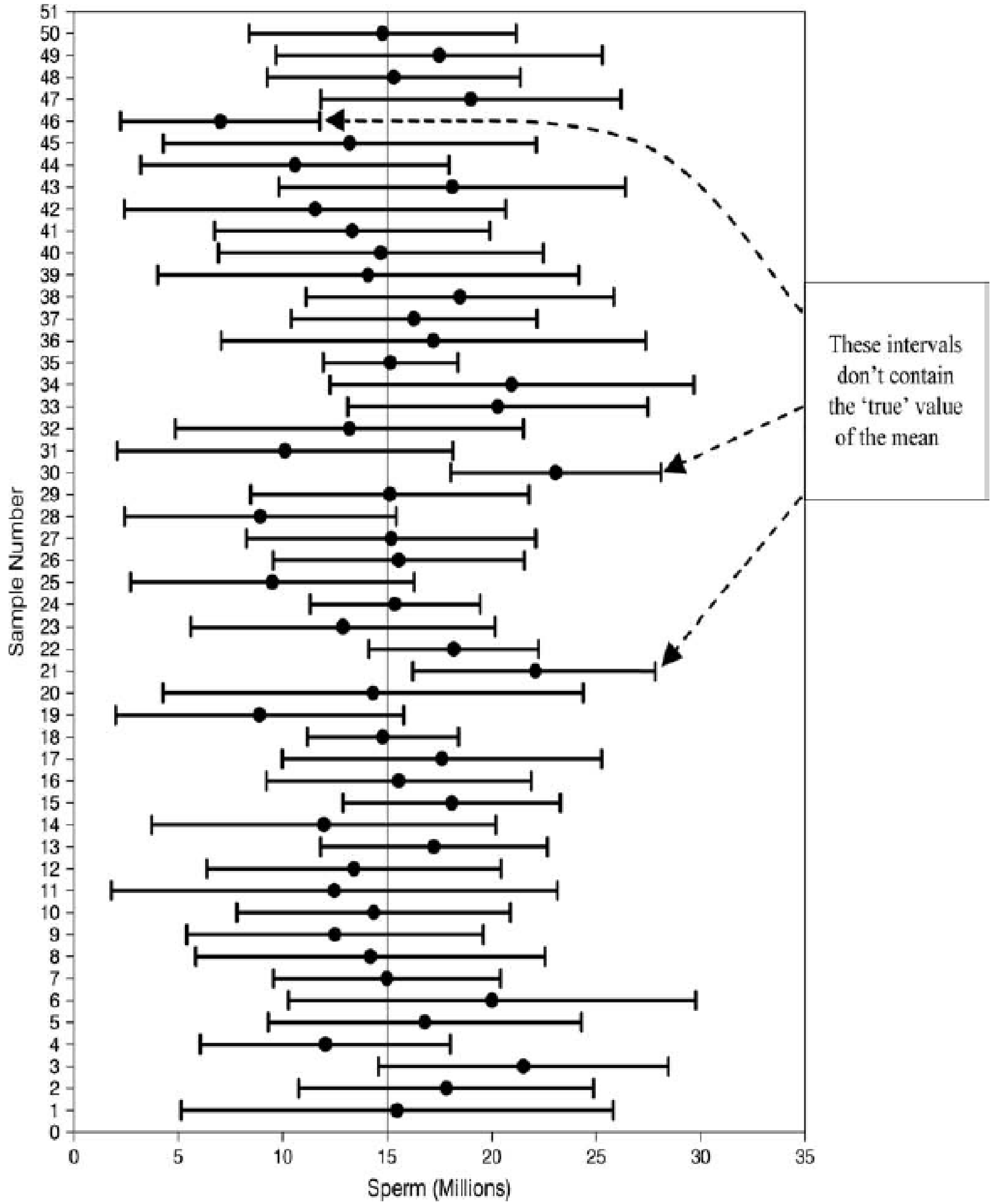
Confidence Intervals



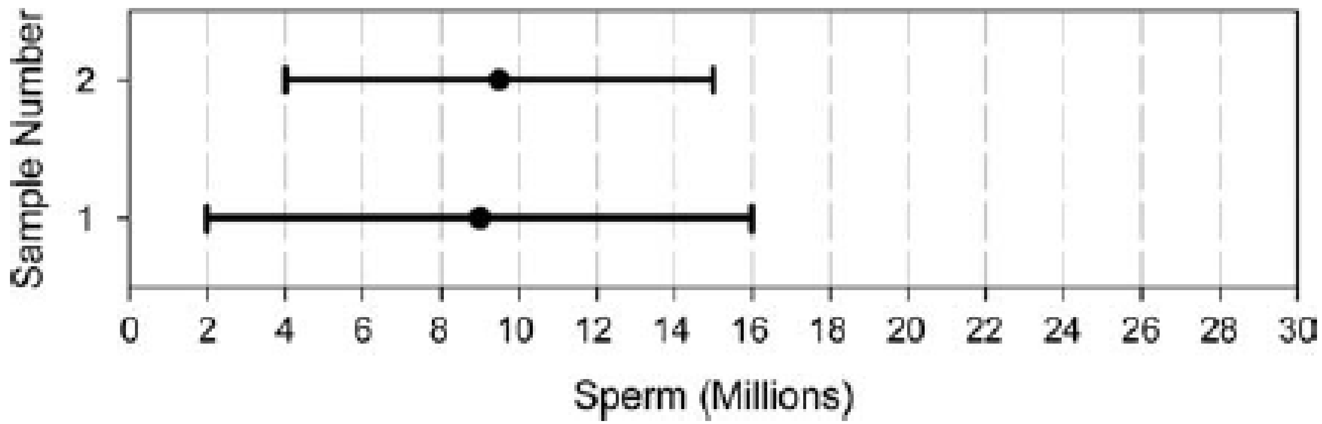
- Domjan et al. (1998)
  - ‘Conditioned’ sperm release in Japanese quail.
- True mean
  - 15 million sperm
- Sample mean
  - 17 million sperm
- Interval estimate
  - 12 to 22 million (contains true value)
  - 16 to 18 million (misses true value)
  - CIs constructed such that 95% contain the true value.

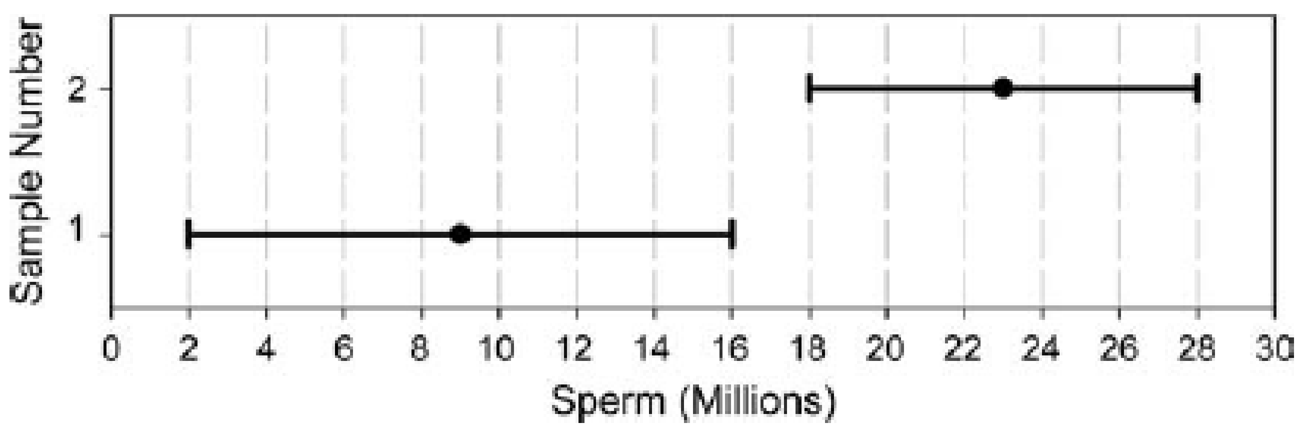
DISCOVERING STATISTICS USING R

FIGURE 2.8
The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)



ANDY FIELD


Sample Number
2
1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Sperm (Millions)


Sample Number
2
1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Sperm (Millions)

Test Statistics

- A statistic for which the frequency of particular values is known.
- Observed values can be used to test hypotheses.

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

One- and Two- Tailed Test

- One Tailed Test
  - A statistical model that tests the directional hypothesis
- Two Tailed Test
  - A statistical model that tests the non-directional hypothesis

One- and Two-Tailed Tests

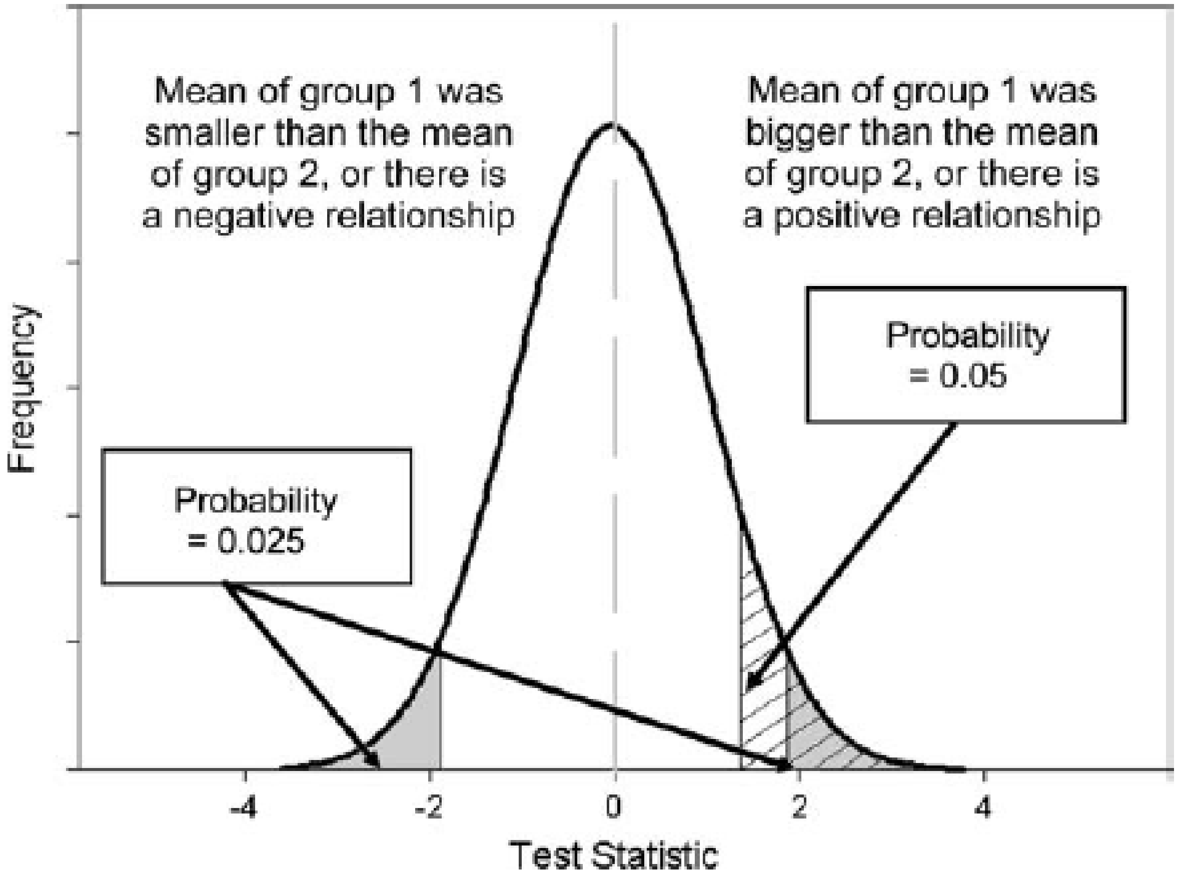


FIGURE 2.10 Diagram to show the difference between one- and two-tailed tests

Type I and Type II Errors

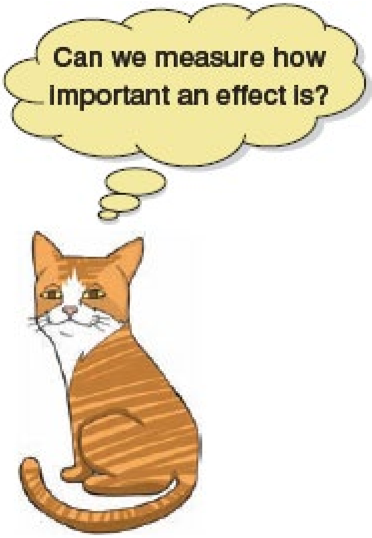
- Type I error
  - occurs when we believe that there is a genuine effect in our population when, in fact, there isn't.
  - The probability is the α-level (usually .05)
- Type II error
  - occurs when we believe that there is no effect in the population when, in reality, there is.
  - The probability is the β-level (often .2)

What Does Statistical Significance Tell Us?

- The importance of an effect?
  - No, significance depends on sample size.
- That the null hypothesis is false?
  - No, it is *always* false.
- That the null hypothesis is true?
  - No, it is *never* true.

Effect Sizes

- An effect size is a standardized measure of the size of an effect:
  - Standardized = comparable across studies
  - Not (as) reliant on the sample size
  - Allows people to objectively evaluate size of observed effect.



Effect Size Measures

- *r* = .1, *d* = .2 (small effect):
  - the effect explains 1% of the total variance.
- *r* = .3, *d* = .5 (medium effect):
  - the effect accounts for 9% of the total variance.
- *r* = .5, *d* = .8 (large effect):
  - the effect accounts for 25% of the variance.
- Beware of these 'canned' effect sizes though:
  - The size of effect should be placed within the research context.

Effect Size Measures

- There are several effect size measures that can be used:
  - Cohen's *d*
  - Pearson's *r*
  - Glass' Δ
  - Hedges' *g*
  - Odds ratio/risk rates
- Pearson's *r* is a good intuitive measure
  - Oh, apart from when group sizes are different …