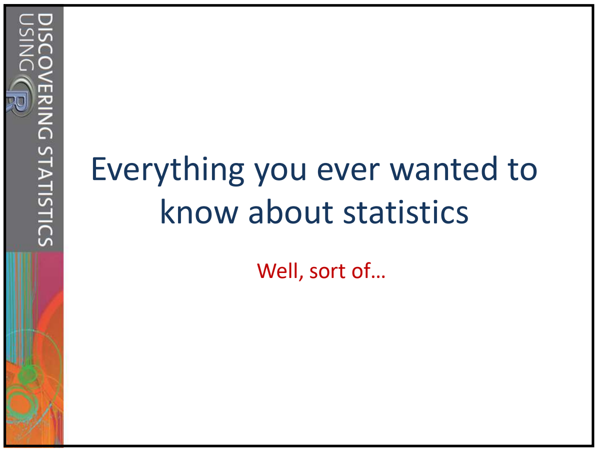
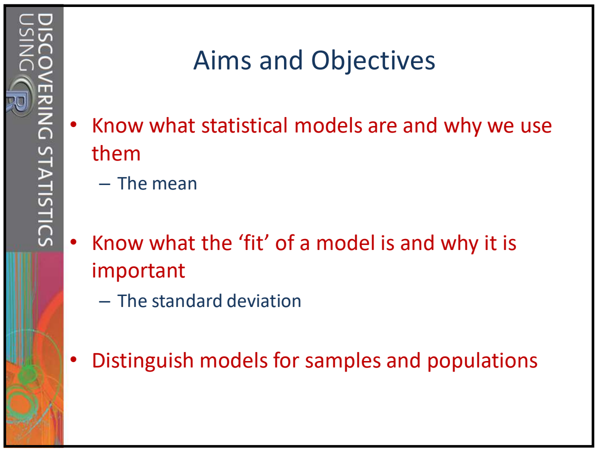


DISCOVERING STATISTICS USING R

Everything you ever wanted to know about statistics

Well, sort of…

1

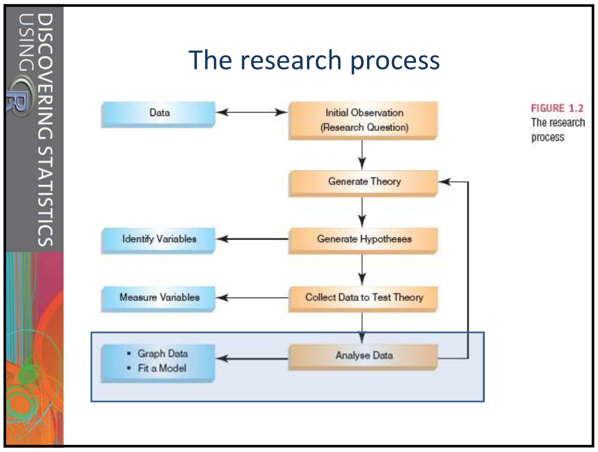


DISCOVERING STATISTICS USING R

Aims and Objectives

- Know what statistical models are 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

2

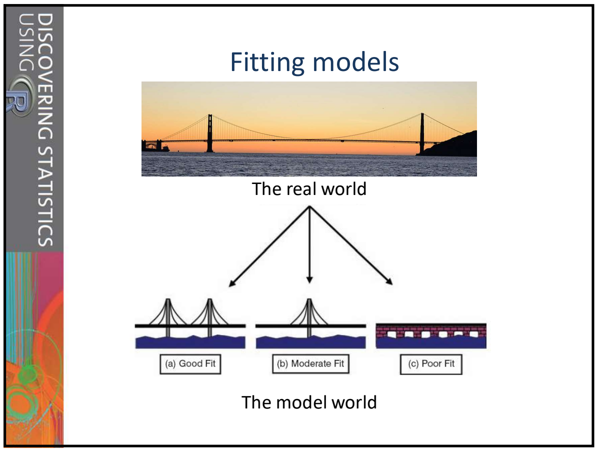


DISCOVERING STATISTICS USING R

The research process

FIGURE 1.2 The research process

3



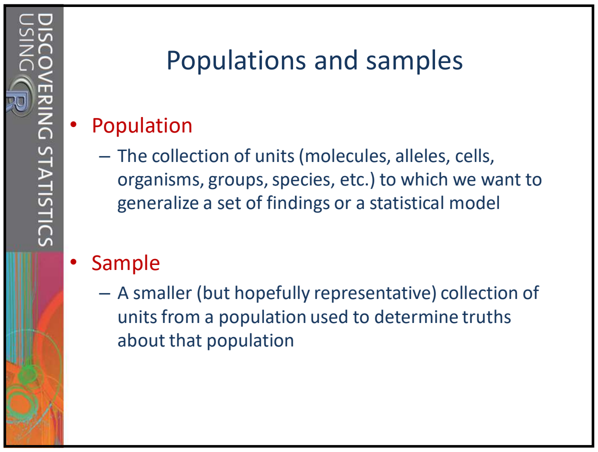
Fitting models

The real world

(a) Good Fit
(b) Moderate Fit
(c) Poor Fit

The model world

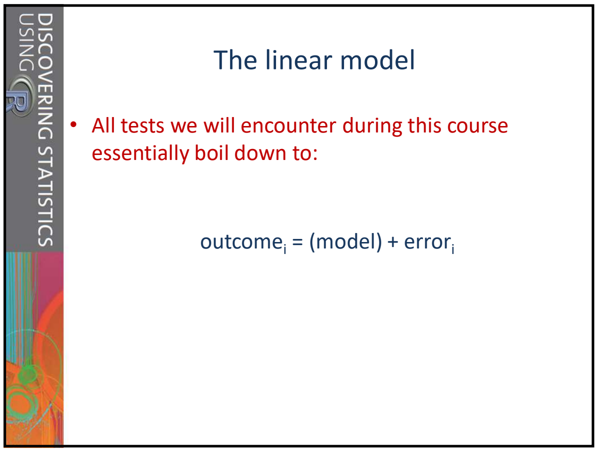
4



Populations and samples

- Population
  - The collection of units (molecules, alleles, cells, organisms, groups, species, 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

5

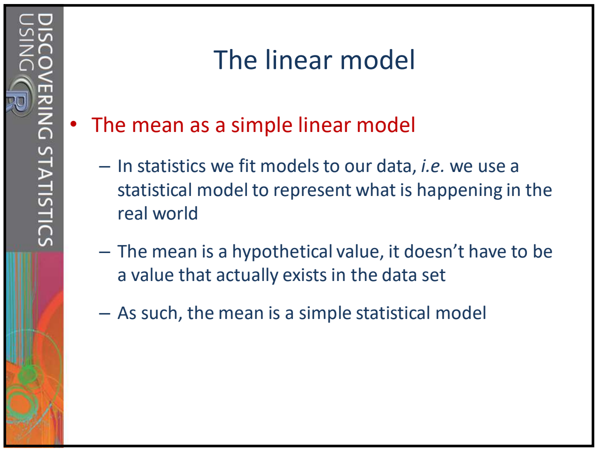


The linear model

- All tests we will encounter during this course essentially boil down to:

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

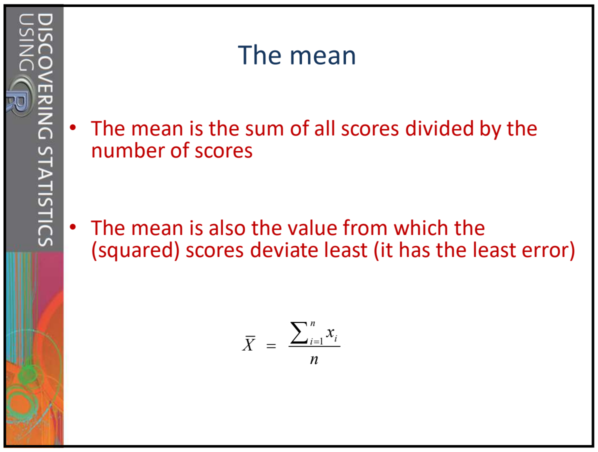
6


DISCOVERING STATISTICS USING R

The linear model

- The mean as a simple linear 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, it doesn't have to be a value that actually exists in the data set
  - As such, the mean is a simple statistical model

7

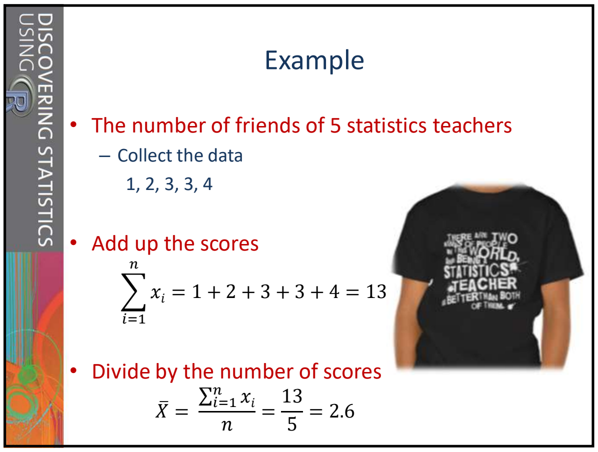

DISCOVERING STATISTICS USING R

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)

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

8


DISCOVERING STATISTICS USING R

Example

- The number of friends of 5 statistics teachers
  - Collect the data

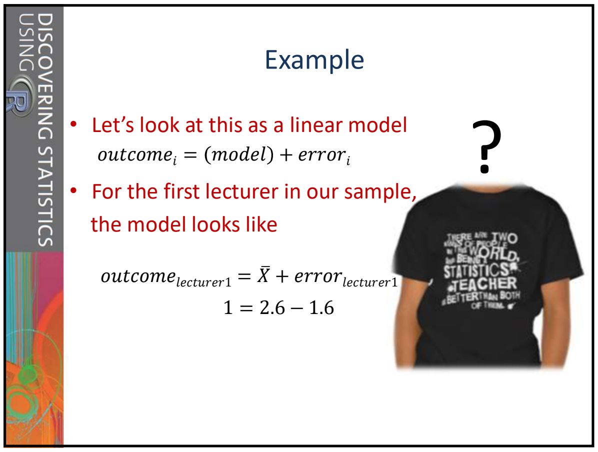
    1, 2, 3, 3, 4
- Add up the scores

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

- Divide by the number of scores

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

9



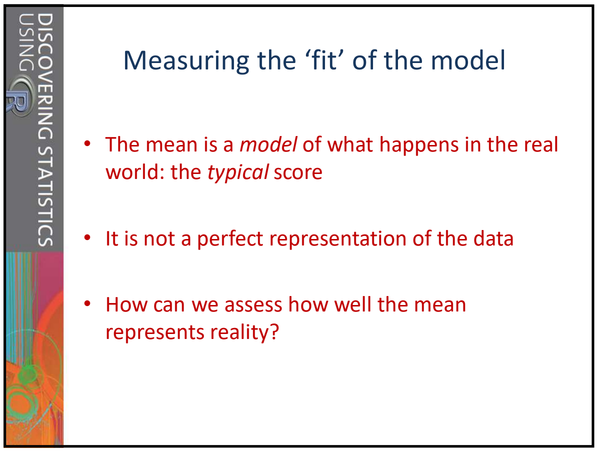
Example

- Let's look at this as a linear model
  $outcome_i = (model) + error_i$
- For the first lecturer in our sample, the model looks like

$$outcome_{lecturer1} = \bar{X} + error_{lecturer1}$$
$$1 = 2.6 - 1.6$$

?
THERE ARE TWO KINDS OF PEOPLE IN THE WORLD
STATISTICS TEACHER
BETTER THAN BOTH OF THEM

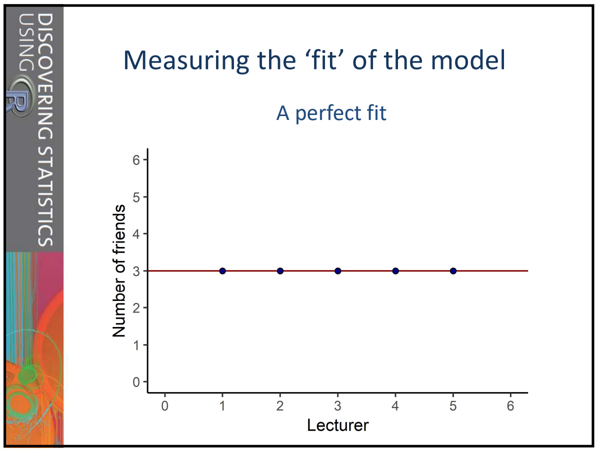
10



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?

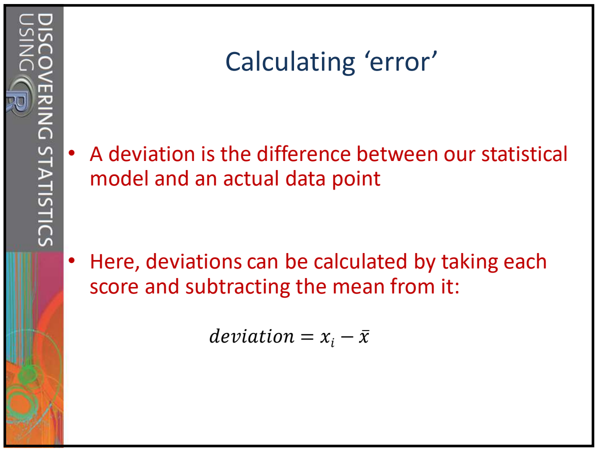
11



Measuring the 'fit' of the model

A perfect fit

12

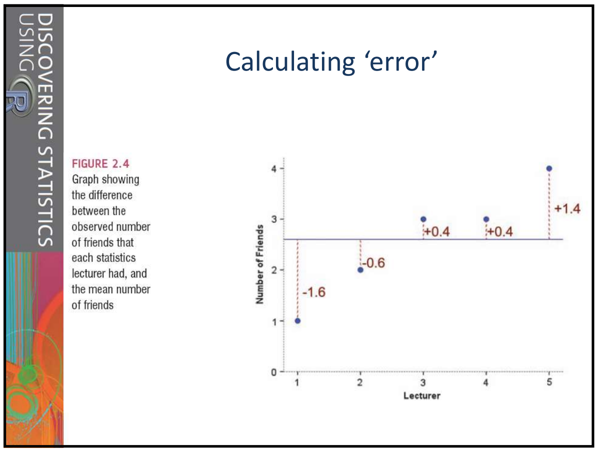


Calculating 'error'

- A deviation is the difference between our statistical model and an actual data point
- Here, deviations can be calculated by taking each score and subtracting the mean from it:

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

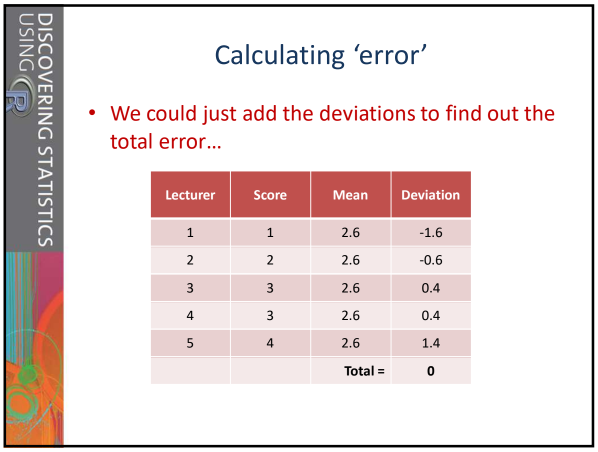
13



Calculating 'error'

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

14

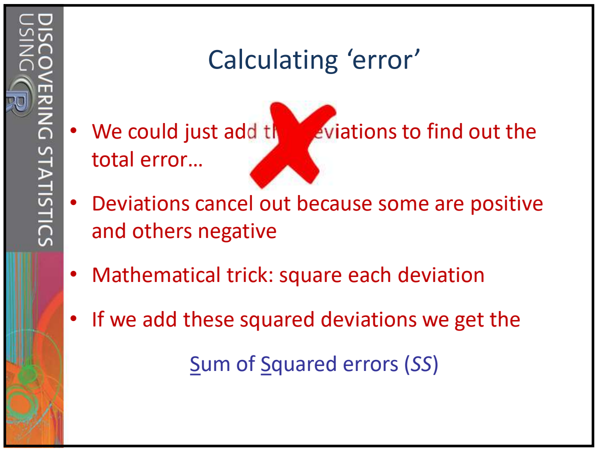


Calculating 'error'

- We could just add the deviations to find out the total error…

Lecturer	Score	Mean	Deviation
1	1	2.6	-1.6
2	2	2.6	-0.6
3	3	2.6	0.4
4	3	2.6	0.4
5	4	2.6	1.4
		Total =	**0**

15

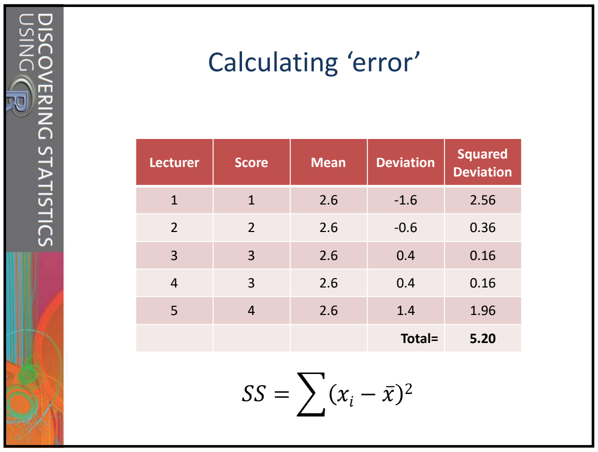


Calculating 'error'

- We could just add the deviations to find out the total error…
- Deviations cancel out because some are positive and others negative
- Mathematical trick: square each deviation
- If we add these squared deviations we get the

Sum of Squared errors (*SS*)

16

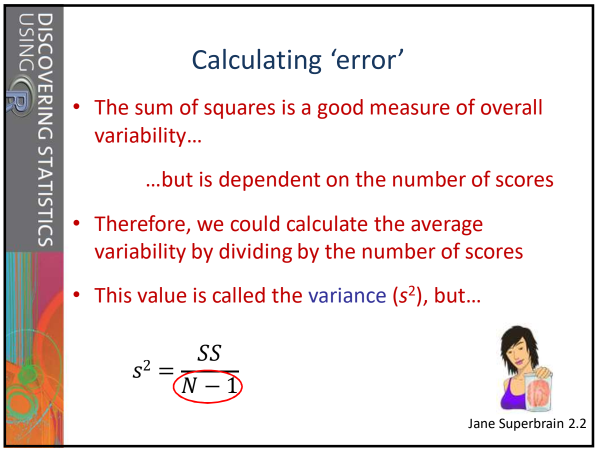


Calculating 'error'

Lecturer	Score	Mean	Deviation	Squared Deviation
1	1	2.6	-1.6	2.56
2	2	2.6	-0.6	0.36
3	3	2.6	0.4	0.16
4	3	2.6	0.4	0.16
5	4	2.6	1.4	1.96
			Total=	**5.20**

$$SS = \sum (x_i - \bar{x})^2$$

17



Calculating 'error'

- The sum of squares is a good measure of overall variability…

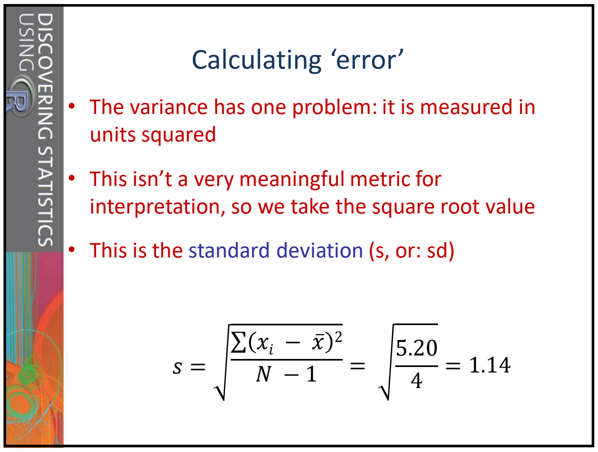
…but is dependent on the number of scores

- Therefore, we could calculate the average variability by dividing by the number of scores
- This value is called the variance (s^2), but…

$$s^2 = \frac{SS}{N-1}$$

Jane Superbrain 2.2

18

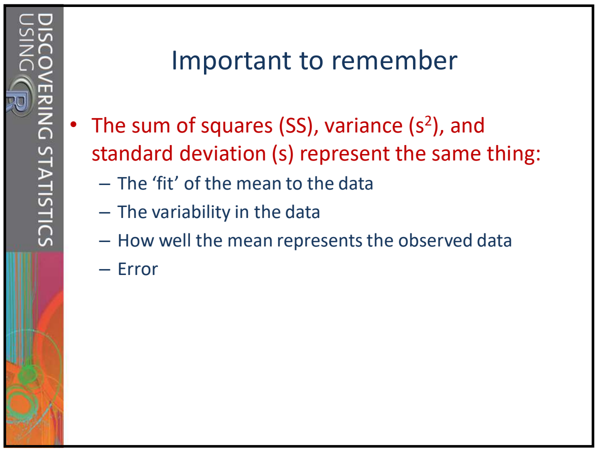

DISCOVERING STATISTICS USING R

Calculating 'error'

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

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

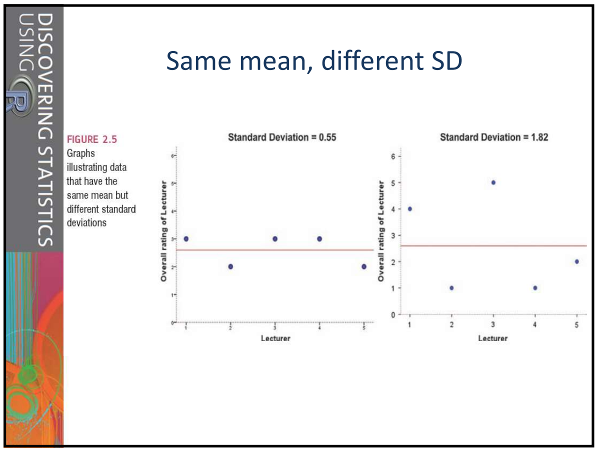
19


DISCOVERING STATISTICS USING R

Important to remember

- The sum of squares (SS), variance (s^2), and standard deviation (s) 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

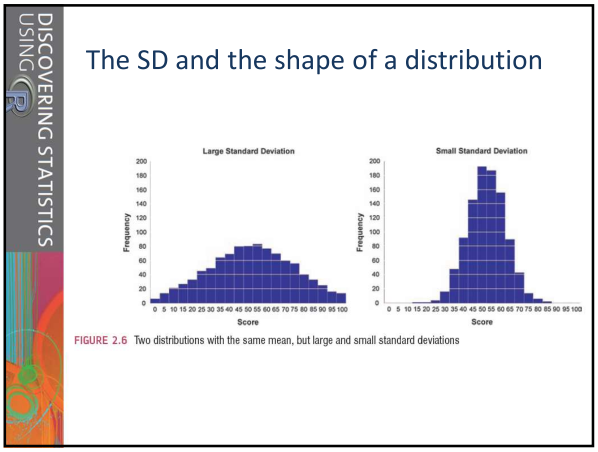
20



Same mean, different SD

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

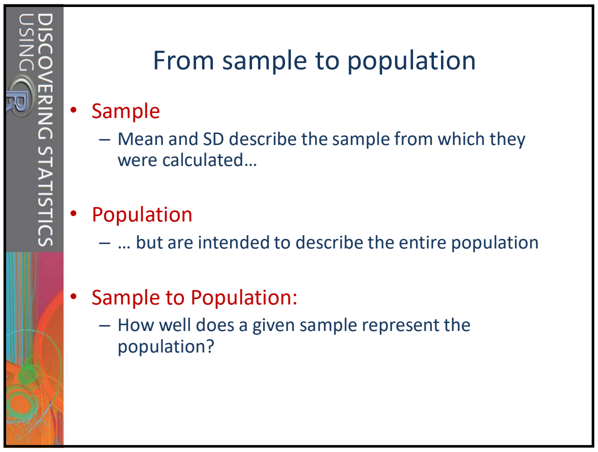
21



The SD and the shape of a distribution

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

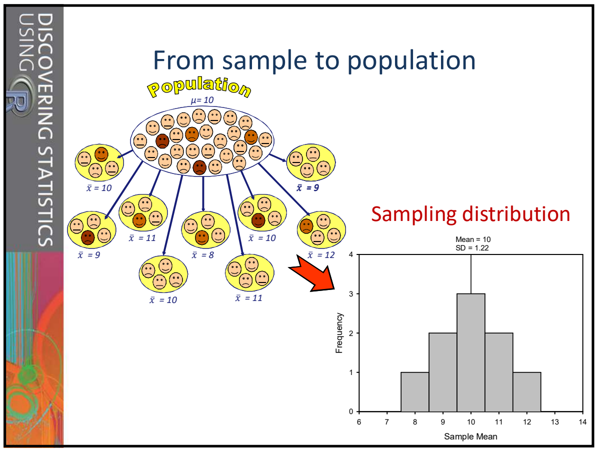
22


DISCOVERING STATISTICS USING R

From sample to population

- Sample
  - Mean and SD describe the sample from which they were calculated…
- Population
  - … but are intended to describe the entire population
- Sample to Population:
  - How well does a given sample represent the population?

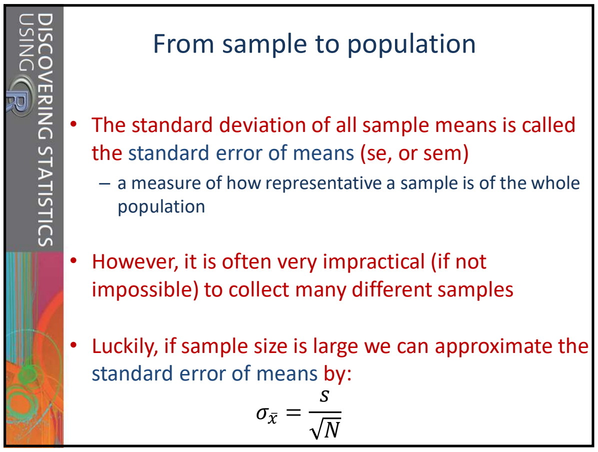
23



From sample to population

Sampling distribution

24

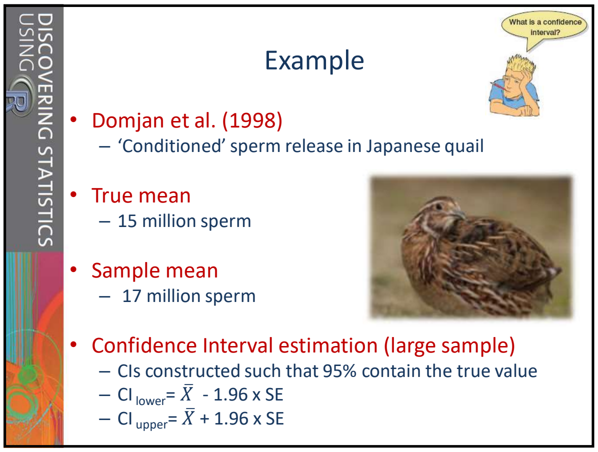


From sample to population

- The standard deviation of all sample means is called the standard error of means (se, or sem)
  - a measure of how representative a sample is of the whole population
- However, it is often very impractical (if not impossible) to collect many different samples
- Luckily, if sample size is large we can approximate the standard error of means by:

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

25



What is a confidence interval?

Example

- Domjan et al. (1998)
  - 'Conditioned' sperm release in Japanese quail
- True mean
  - 15 million sperm
- Sample mean
  - 17 million sperm
- Confidence Interval estimation (large sample)
  - CIs constructed such that 95% contain the true value
  - $\text{CI}_{\text{lower}} = \bar{X}$ - 1.96 x SE
  - $\text{CI}_{\text{upper}} = \bar{X}$ + 1.96 x SE

26

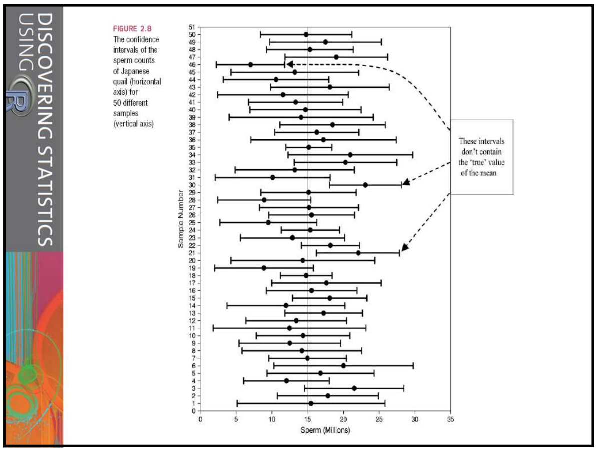
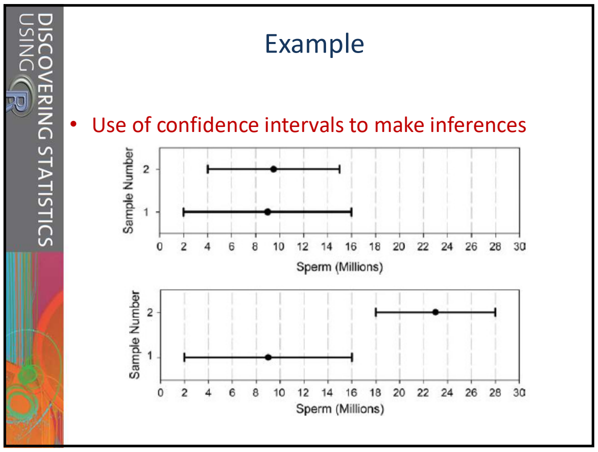


FIGURE 2.8 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)

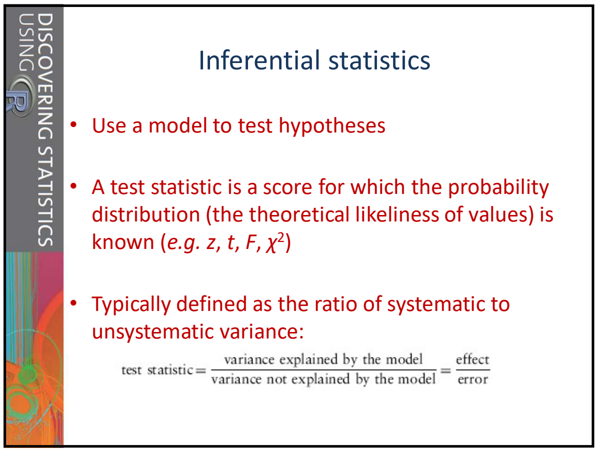
27



Example

- Use of confidence intervals to make inferences

28

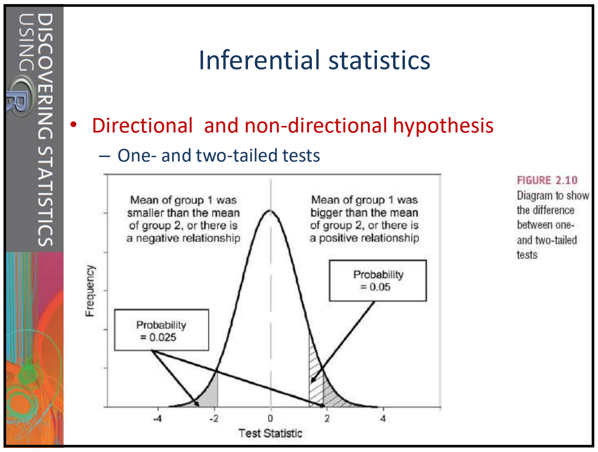

DISCOVERING STATISTICS USING R

Inferential statistics

- Use a model to test hypotheses
- A test statistic is a score for which the probability distribution (the theoretical likeliness of values) is known (*e.g. z*, *t*, *F*, χ^2)
- Typically defined as the ratio of systematic to unsystematic variance:

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

29

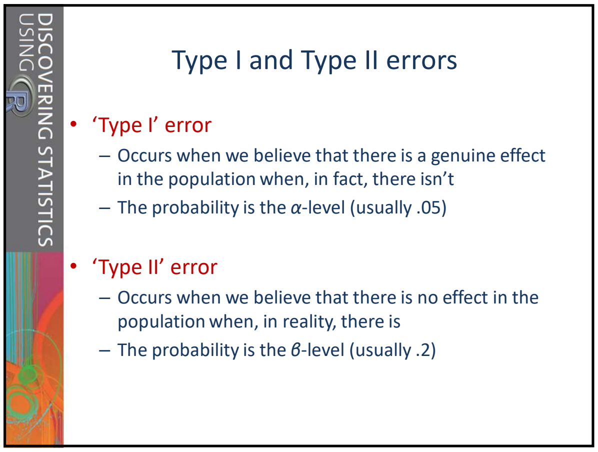

DISCOVERING STATISTICS USING R
Mean of group 1 was smaller than the mean of group 2, or there is a negative relationship
Mean of group 1 was bigger than the mean of group 2, or there is a positive relationship
Probability = 0.05
Probability = 0.025
Frequency
-4
-2
0
2
4
Test Statistic

Inferential statistics

- Directional and non-directional hypothesis
  - One- and two-tailed tests

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

30

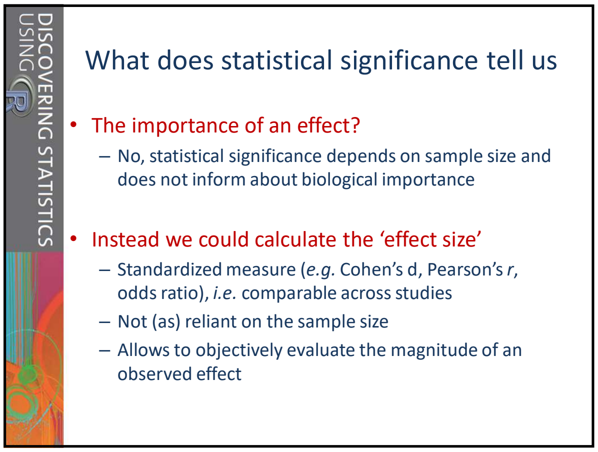


DISCOVERING STATISTICS USING R

Type I and Type II errors

- 'Type I' error
  - Occurs when we believe that there is a genuine effect in the 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 (usually .2)

31

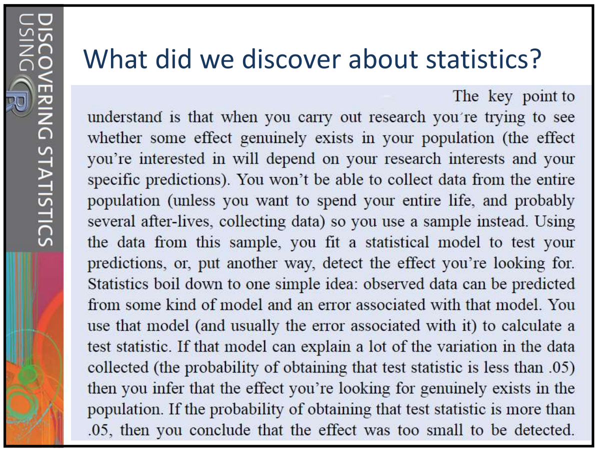


DISCOVERING STATISTICS USING R

What does statistical significance tell us

- The importance of an effect?
  - No, statistical significance depends on sample size and does not inform about biological importance
- Instead we could calculate the 'effect size'
  - Standardized measure (*e.g.* Cohen's d, Pearson's *r*, odds ratio), *i.e.* comparable across studies
  - Not (as) reliant on the sample size
  - Allows to objectively evaluate the magnitude of an observed effect

32



DISCOVERING STATISTICS USING R

What did we discover about statistics?

The key point to understand is that when you carry out research you're trying to see whether some effect genuinely exists in your population (the effect you're interested in will depend on your research interests and your specific predictions). You won't be able to collect data from the entire population (unless you want to spend your entire life, and probably several after-lives, collecting data) so you use a sample instead. Using the data from this sample, you fit a statistical model to test your predictions, or, put another way, detect the effect you're looking for. Statistics boil down to one simple idea: observed data can be predicted from some kind of model and an error associated with that model. You use that model (and usually the error associated with it) to calculate a test statistic. If that model can explain a lot of the variation in the data collected (the probability of obtaining that test statistic is less than .05) then you infer that the effect you're looking for genuinely exists in the population. If the probability of obtaining that test statistic is more than .05, then you conclude that the effect was too small to be detected.

33