DISCOVERING STATISTICS USING R

Why do we need statistics?

Or:

Why is my evil lecturer forcing me to learn statistics?

1

DISCOVERING STATISTICS USING R

Aims and Objectives

- Know the different steps that make up the scientific research process
- Know the (broadly speaking) 3 methods of hypothesis testing in biology
- Know what the 'frequency distribution' of data is, and how it can be characterised

2

DISCOVERING STATISTICS USING R

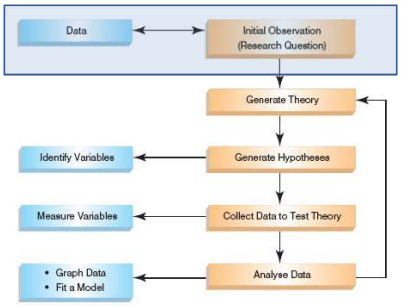
Types of data analysis

- Quantitative methods
  - Testing theories using numbers
- Qualitative methods
  - Testing theories using language
    - Magazine articles/Interviews
    - Conversations
    - Newspapers
    - Media broadcasts

3

DISCOVERING STATISTICS USING R

The research process



1

FIGURE 1.2 The research process

4

DISCOVERING STATISTICS USING R

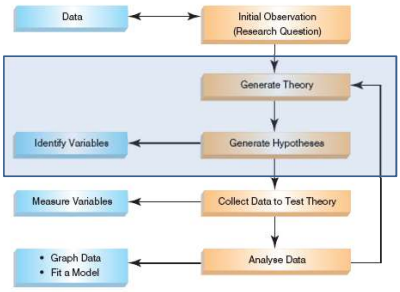
Initial observation

- Find something that needs explaining
  - Observe the real world
  - Read other research
  - In biology we often distinguish between 'proximate' and 'ultimate' explanations
- Test the concept: collect data
  - Collect data to see whether your hunch is correct
  - To do this you need to define variables
    - Anything that can be measured and can differ across entities or time

5

DISCOVERING STATISTICS USING R

The research process



2

FIGURE 1.2 The research process

6

DISCOVERING STATISTICS USING R

Generating and testing theories

- Theory
  - A hypothesized general principle or set of principles that explains known findings about a topic and from which new hypotheses can be generated
- Hypothesis
  - A prediction from a theory
  - Null hypothesis (H_0): there is no difference
  - Alternative hypothesis (H_1): there is a difference
- Falsification
  - The act of disproving a theory or hypothesis

7

DISCOVERING STATISTICS USING R

TABLE 1.1 A table of the number of people at the *Big Brother* audition split by whether they had narcissistic personality disorder and whether they were selected as contestants by the producers

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

8

DISCOVERING STATISTICS USING R

Types of hypotheses

- Null hypothesis, H_0
  - *Big Brother* contestants and members of the public do not differ in their scores on personality disorder questionnaires
- The alternative hypothesis, H_1
  - *Big Brother* contestants and members of the public differ in their scores on personality disorder questionnaires

9

DISCOVERING STATISTICS USING R

The research process

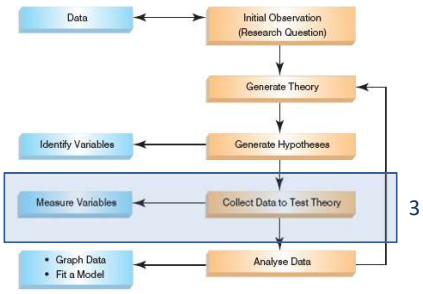


FIGURE 1.2 The research process

10

DISCOVERING STATISTICS USING R

Data collection (1): What to measure?

- Hypothesis:
  - *Coca-Cola kills sperm*
- Independent (or "predictor") variable
  - The proposed cause
  - A manipulated variable (in experiments)
  - Coca-Cola in the hypothesis above
- Dependent (or "outcome/response") variable
  - The proposed effect
  - Measured not manipulated (in experiments)
  - Sperm in the hypothesis above

(Umpierre, Hill & Anderson 1985)

11

DISCOVERING STATISTICS USING R

Levels of measurement

- Categorical:
  - Binary variable: only two categories
    - *e.g.* male or female
  - Nominal variable: more than two categories
    - *e.g.* forest, grassland, woodland, desert
  - Ordinal variable: a nominal variable with a logical order
    - *e.g.* low, medium and high predation risk
- Continuous:
  - Interval variable: equal intervals on the variable represent equal differences in the property being measured
    - *e.g.* the difference between 6 and 8 is equivalent to the difference between 13 and 15
  - Ratio variable: same as an interval variable, but the ratios of scores on the scale must also make sense
    - *e.g.* a weight of 6.2 kg means that the subject is twice as heavy as another subject weighing 3.1 kg

12

DISCOVERING STATISTICS USING R

Measurement error

- Measurement error
  - The discrepancy between the actual value we're trying to measure, and the number we use to represent that value
- Example:
  - I (in reality) weigh 80 kg
  - My bathroom scales say 83 kg
  - The measurement error is 3 kg

13

DISCOVERING STATISTICS USING R

Validity and Reliability

- Criterion validity
  - Whether an instrument actually measures what it set out to measure
- Content validity
  - Evidence that the content of a test corresponds to the content of the construct it was designed to cover
- Ecological validity
  - Evidence that the results of a study, experiment or test can be applied, and allow inferences, to real-world conditions
- Reliability
  - The ability of the measure to produce the same results under the same conditions
- Test–retest reliability
  - The ability of a measure to produce consistent results when the same entities are tested at two different points in time

14

DISCOVERING STATISTICS USING R

Data collection (2): How to measure?

Instead of section 1.6.1 in book

- Methods of hypothesis testing in biology:
  - Observational (correlational)
    - Observing what naturally goes on in a system without directly interfering with it
    - Quantitative description
  - Experimental
    - One (or more) variable is systematically manipulated to see its effect on an outcome variable
    - Make statements about cause and effect
  - Comparative (phylogenetic)
    - Compare traits among different taxa in relation to the same environmental variables
    - Look at 'experiments' done by natural selection

15

DISCOVERING STATISTICS USING R

Data collection (2): How to measure?

- Study design
  - Independent (between-subjects)
    - Different subjects are used in all conditions
    - Values of the response variable are independent
  - Repeated-measures (within-subject)
    - The same subjects take part in all (or most) conditions
    - Values of the response variable are not independent

16

DISCOVERING STATISTICS USING R

Types of variation

- Systematic
  - Differences in response due to explanatory variable
- Unsystematic
  - Differences in response due to confounding variables and noise
    - Age, sex, measurement error, *etc.*
- Randomization and counterbalancing
  - Minimize unsystematic variation
    - *e.g.* due to differences between subjects or practice/boredom effects

17

DISCOVERING STATISTICS USING R

The research process

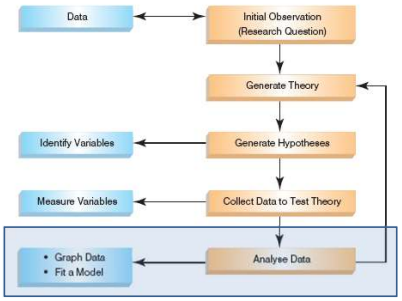


FIGURE 1.2 The research process

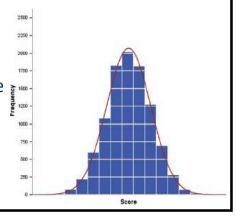
4

18

DISCOVERING STATISTICS USING R

Analysing data

- Frequency distributions (histograms)
  - A graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set
- The 'normal' distribution
  - Bell-shaped
  - Symmetrical around the centre



19

DISCOVERING STATISTICS USING R

Properties of frequency distributions

1. Skew
   - The symmetry of the distribution
   - Positive
     - scores bunched at low values, tail pointing to high values
   - Negative
     - scores bunched at high values, tail pointing to low values

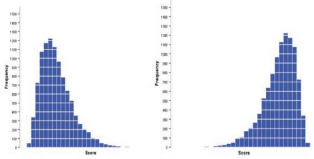


FIGURE 1.4 A positively (left figure) and negatively (right figure) skewed distribution

20

DISCOVERING STATISTICS USING R

Properties of frequency distributions

2. Kurtosis
   - The 'heaviness' of the tails
   - Positive
     - heavy tails, 'leptokurtic'
   - Negative
     - light tails, 'platykurtic'

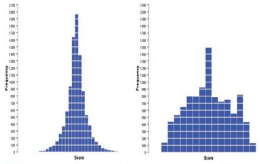


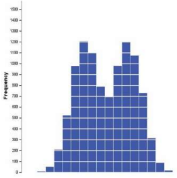
FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

21

DISCOVERING STATISTICS USING R

Properties of frequency distributions

3. Central tendency
   - Mode
     - The most frequent score
       - Unimodal: having one mode
       - Bimodal: having two modes
       - Multimodal: having several modes



   - Median
     - The middle score when scores are ordered

       22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252
   - Mean
     - The sum of scores divided by the number of scores

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

22

DISCOVERING STATISTICS USING R

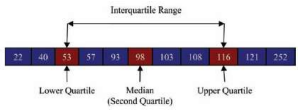
Properties of frequency distributions

4. Dispersion
   - The range
     - The smallest score subtracted from the largest

       22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252

       → 252 – 22= 230
     - Very biased by outliers
   - The interquartile range
     - Quartiles split the sorted data into four equal parts

FIGURE 1.7 Calculating quartiles and the interquartile range



23

DISCOVERING STATISTICS USING R

Beyond the frequency distribution

- Probability distributions
  - Another way to think about frequency distributions is not in terms of how often scores occur in a given dataset, but how likely they are to occur in general (*i.e.* probability)
  - Idealized frequency distributions that enable us to assess how likely a given value in our data is
    - *e.g.* the normal distribution
- *z*-scores
  - Standardise a score with respect to the other scores in the group
  - Express a score in terms of how many standard deviations it is away from the mean
  - The distribution of *z*-scores has a mean of 0 and *SD* = 1

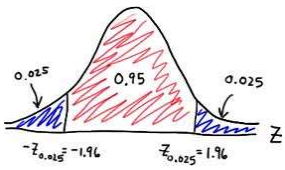
24

DISCOVERING STATISTICS USING R

The normal distribution and *z*-scores

$$z = \frac{X - \bar{X}}{s}$$

- 1.96 cuts off the top 2.5% of the distribution
- −1.96 cuts off the bottom 2.5% of the distribution
- Thus, 95% of *z*-scores lie between −1.96 and 1.96



25

DISCOVERING STATISTICS USING R

What did we discover about statistics?

Actually, not a lot because we haven't really got to the statistics bit yet. However, we have discovered some stuff about the process of doing research. We began by looking at how research questions are formulated through observing phenomena or collecting data about a 'hunch'. Once the observation has been confirmed, theories can be generated about why something happens. From these theories we formulate hypotheses that we can test. To test hypotheses we need to measure things and this leads us to think about the variables that we need to measure and how to measure them. Then we can collect some data. The final stage is to analyse these data. In this chapter we saw that we can begin by just looking at the shape of the data but that ultimately we should end up fitting some kind of statistical model to the data (more on that in the rest of the book). In short, the reason that your evil statistics lecturer is forcing you to learn statistics is because it is an intrinsic part of the research process and it gives you enormous power to answer questions that are interesting.

26