

# **MAST90007**

## **Statistics for Research Workers**

**June/July 2020**

### **WARNING**

This material has been reproduced and communicated to you by or on behalf of the University of Melbourne in accordance with section 113P of the Copyright Act 1968 (Act).

The material in this communication may be subject to copyright under the Act.

Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

**Do not remove this notice**



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The scope of statistical science . . . . .	1
1.1.1	Types of applications . . . . .	2
1.1.2	COVID-19 . . . . .	3
1.2	Types of research studies . . . . .	4
1.2.1	Paired samples designs . . . . .	5
1.2.2	Independent samples design, with two groups . . . . .	6
1.3	Types of data . . . . .	7
1.4	Exercises . . . . .	11
1.5	Answers . . . . .	13
<b>2</b>	<b>The language of statistics: describing, summarising and visualising</b>	<b>17</b>
2.1	Quality data presentation . . . . .	17
2.1.1	Graphs from software packages . . . . .	19
2.2	Visualising numerical data . . . . .	19
2.2.1	Understanding distributions . . . . .	19
2.2.2	Exploring relationships . . . . .	23
2.3	Summarising numerical data . . . . .	27
2.3.1	Measures of location . . . . .	28
2.3.2	Measures of spread . . . . .	32
2.3.3	Other summary measures . . . . .	34
2.3.4	Basic descriptives in MINITAB . . . . .	34
2.3.5	Summarising relationships: correlation . . . . .	35
2.4	Visualising summaries of numerical data: boxplots . . . . .	42
2.4.1	“Outliers” . . . . .	44
2.5	Visualising numerical data: making comparisons . . . . .	47
2.6	Visualising and summarising categorical data . . . . .	49
2.6.1	Understanding distributions . . . . .	49
2.6.2	Exploring relationships . . . . .	50
2.7	Visualising more complex data . . . . .	52
2.7.1	Panel graphs . . . . .	52
2.7.2	Scatterplot matrices . . . . .	53
2.8	Exercises . . . . .	55
2.9	Answers . . . . .	59
<b>3</b>	<b>Principles of communicating analytics</b>	<b>67</b>
3.1	Quality graphs . . . . .	67
3.2	Five principles of good graphs . . . . .	68

3.2.1	Show the data clearly . . . . .	68
3.2.2	Use simplicity in design . . . . .	71
3.2.3	Use good alignment on a common scale for comparison . . . . .	73
3.2.4	Keep the visual encoding transparent . . . . .	75
3.2.5	Prefer standard forms demonstrated to be effective . . . . .	77
3.3	Quality tables . . . . .	78
3.3.1	Tables of data . . . . .	79
3.3.2	Tables of summary statistics . . . . .	80
3.3.3	Tables of inferential statistics . . . . .	81
3.4	Exercises . . . . .	85
3.5	Answers . . . . .	87
<b>4</b>	<b>Foundations for inference: more advanced language</b>	<b>89</b>
4.1	Populations versus samples . . . . .	90
4.2	Understanding variation . . . . .	94
4.3	Probability . . . . .	95
4.4	Random variables and distributions . . . . .	97
4.4.1	Discrete random variables . . . . .	97
4.4.2	Binomial distribution . . . . .	100
4.4.3	Mean, variance and standard deviation . . . . .	102
4.4.4	Continuous random variables . . . . .	104
4.4.5	Normal distribution . . . . .	105
4.4.6	Abstractions . . . . .	109
4.4.7	An important notational convention . . . . .	110
4.5	Some important results . . . . .	111
4.5.1	Rescaling a random variable . . . . .	111
4.5.2	Sums and differences of random variables . . . . .	112
4.5.3	Mean and variance of $\bar{X}$ , the sample mean . . . . .	113
4.6	Other inference perspectives . . . . .	115
4.7	Exercises . . . . .	117
4.8	Answers . . . . .	123
<b>5</b>	<b>Confidence intervals</b>	<b>127</b>
5.1	Concept behind confidence intervals . . . . .	127
5.1.1	Choice of the confidence coefficient . . . . .	129
5.1.2	Considering precision . . . . .	130
5.1.3	Reporting confidence intervals . . . . .	131
5.2	Confidence intervals in Normal populations . . . . .	132
5.2.1	The mean of a Normal population, $\mu$ . . . . .	132

5.2.2	Degrees of freedom . . . . .	138
5.2.3	The difference between the means of two Normal populations — paired samples . . . . .	139
5.2.4	The difference between the means of two Normal populations — independent samples . . . . .	143
5.3	The amazing Central Limit Theorem . . . . .	147
5.3.1	The Theorem . . . . .	147
5.3.2	Confidence intervals for means . . . . .	149
5.3.3	Confidence interval for a population proportion, $\theta$ . . . . .	150
5.3.4	Confidence interval for difference between two population proportions, $\theta_1 - \theta_2$ . . . . .	154
5.3.5	A general form for confidence intervals . . . . .	155
5.4	Confidence interval for a correlation, $\rho$ . . . . .	156
5.5	Exercises . . . . .	159
5.6	Answers . . . . .	166
<b>6</b>	<b>Hypothesis testing</b>	<b>173</b>
6.1	The concept of a $P$ -value . . . . .	173
6.2	The null hypothesis . . . . .	174
6.3	Formulating statistical hypotheses . . . . .	176
6.4	An example: water quality . . . . .	177
6.5	Structure of the hypothesis test . . . . .	180
6.6	Misconceptions about hypothesis testing and $P$ -values . . . . .	182
6.7	Principled use of hypothesis tests and statistical inference . . . . .	184
6.8	Principled reporting of hypothesis tests and statistical inference . . . . .	185
6.9	One- or two-sided statistical tests . . . . .	190
6.10	Confidence intervals and hypothesis tests . . . . .	194
6.11	Exercises . . . . .	197
6.12	Answers . . . . .	199
<b>7</b>	<b>Models for data</b>	<b>201</b>
7.1	Distributions for modelling data . . . . .	201
7.2	Models for residual variation . . . . .	206
7.3	Model complexity . . . . .	206
7.3.1	Simple models . . . . .	207
7.3.2	Complex models . . . . .	207
<b>8</b>	<b>The linear model — a broad view</b>	<b>209</b>
8.1	Simple linear models . . . . .	209
8.2	Response variable . . . . .	211
8.3	Linear function . . . . .	211

8.4 Random error . . . . .	213
8.5 Types of explanatory variables . . . . .	215
<b>9 Inference — numerical outcome, paired samples</b>	<b>217</b>
9.1 Assuming Normality for the data: the $t$ test . . . . .	217
9.2 Distribution-free tests . . . . .	219
9.3 The sign test . . . . .	220
9.3.1 Hypothesis test . . . . .	220
9.3.2 Confidence interval . . . . .	222
9.4 Wilcoxon matched-pairs signed-rank test . . . . .	222
9.4.1 Hypothesis test . . . . .	223
9.4.2 Confidence interval . . . . .	224
9.5 Comparing the inferences . . . . .	225
9.6 Exercises . . . . .	226
9.7 Answers . . . . .	228
<b>10 Inference — numerical outcome and one categorical explanatory variable</b>	<b>233</b>
10.1 Independent samples . . . . .	233
10.2 $t$ -test for independent samples . . . . .	233
10.2.1 The case of equal variances . . . . .	233
10.3 Inference for $k$ means, independent samples . . . . .	238
10.3.1 Concept of the approach . . . . .	238
10.3.2 The $\chi^2$ and $F$ distributions . . . . .	240
10.3.3 Sums of squares and mean squares . . . . .	242
10.3.4 Developing the formal test of $H_0$ . . . . .	244
10.3.5 A connection with the two sample $t$ test . . . . .	246
10.3.6 ANOVA of the white pine data . . . . .	246
10.3.7 Exploring changes to the white pine data . . . . .	247
10.4 Confidence intervals and multiple comparisons . . . . .	250
10.4.1 Confidence intervals . . . . .	250
10.4.2 Multiple comparisons . . . . .	251
10.4.3 The multiple comparisons controversy . . . . .	254
10.5 Exercises . . . . .	257
10.6 Answers . . . . .	261
<b>11 Assumptions — numerical outcome and one categorical explanatory variable</b>	<b>271</b>
11.1 Two sample $t$ test, unequal variances . . . . .	271
11.2 Mann-Whitney test . . . . .	272
11.3 Assumptions of the model and model-checking . . . . .	274

---

11.3.1 Residuals . . . . .	275
11.3.2 Model-checking . . . . .	276
11.4 Dealing with model violations . . . . .	278
11.4.1 The Kruskal-Wallis test . . . . .	278
11.4.2 Transformations . . . . .	279
11.4.3 One-way ANOVA without assuming equal variances . . . . .	281
11.5 Exercises . . . . .	282
11.6 Answers . . . . .	285
<b>12 Inference — numerical outcome and two categorical explanatory variables</b>	<b>291</b>
12.1 Randomized blocks design . . . . .	291
12.1.1 Two-way analysis of variance . . . . .	292
12.1.2 The Friedman test . . . . .	297
12.2 Factorial experiments with two factors . . . . .	298
12.2.1 Interaction . . . . .	299
12.2.2 Analysis of a factorial design . . . . .	301
12.3 Exercises . . . . .	303
12.4 Answers . . . . .	306
<b>13 Inference — numerical outcome and numerical explanatory variables</b>	<b>317</b>
13.1 Simple linear regression: one predictor . . . . .	317
13.1.1 Estimation — the method of least squares . . . . .	319
13.1.2 Confidence intervals . . . . .	320
13.1.3 Prediction intervals . . . . .	321
13.1.4 Hypothesis testing . . . . .	322
13.2 Multiple regression . . . . .	323
13.2.1 Extension of simple linear regression . . . . .	323
13.2.2 $R^2$ . . . . .	324
13.2.3 Adjusting for other predictors . . . . .	325
13.2.4 Checking assumptions: residual analysis . . . . .	328
13.2.5 Polynomial regression . . . . .	329
13.3 Exercises . . . . .	330
13.4 Answers . . . . .	334
<b>14 Inference — binary outcome</b>	<b>349</b>
14.1 Introduction . . . . .	349
14.2 The logistic regression model . . . . .	351
14.2.1 One binary explanatory variable . . . . .	353
14.2.2 Categorical explanatory variables with more than two levels .	354
14.2.3 One continuous explanatory variable . . . . .	356

---

14.3 Estimated probabilities . . . . .	357
14.4 Assessing goodness of fit . . . . .	358
14.5 Exercises . . . . .	359
14.6 Answers . . . . .	361
<b>15 Inference — categorical outcome, simple methods</b>	<b>365</b>
15.1 Hypothesis test for comparing two population proportions . . . . .	365
15.2 Fisher's exact test . . . . .	367
15.3 $\chi^2$ test for $r \times c$ table . . . . .	370
15.4 Other methods . . . . .	375
15.5 Exercises . . . . .	376
15.6 Answers . . . . .	378
<b>16 Study planning and design</b>	<b>381</b>
16.1 Design of experiments . . . . .	381
16.1.1 Randomization . . . . .	381
16.1.2 Blocking . . . . .	382
16.1.3 Replication . . . . .	383
16.2 Some common designs . . . . .	383
16.2.1 Completely randomized design . . . . .	383
16.2.2 Randomized block design . . . . .	384
16.2.3 Latin square design . . . . .	384
16.2.4 Extensions . . . . .	387
16.3 Study size . . . . .	388
16.4 Use of confidence intervals to determine sample size . . . . .	389
16.5 The hypothesis testing decision making framework . . . . .	392
16.5.1 A competing alternative hypothesis ( $H_1$ ) . . . . .	393
16.5.2 Level of significance ( $\alpha$ ) and type I error . . . . .	394
16.5.3 Power . . . . .	395
16.5.4 Type II error . . . . .	397
16.5.5 A useful analogy . . . . .	398
16.6 Use of power to determine sample size . . . . .	398
16.6.1 Sample size for a comparison of means . . . . .	399
16.6.2 Sample size for a comparison of proportions . . . . .	403
16.6.3 General issues in sample size determination . . . . .	404
16.7 Exercises . . . . .	407
16.8 Answers . . . . .	409
<b>17 Extensions and modern methods</b>	<b>411</b>
17.1 Extensions of linear models . . . . .	411

17.2 Some modern methods . . . . .	412
17.2.1 Regression trees . . . . .	412
17.2.2 Random forests . . . . .	413
17.2.3 Training and testing . . . . .	414



# Statistics for Research Workers

Statistical Consulting Centre  
University of Melbourne

## 1 Introduction

This course's name includes the word "statistics". That word's origins are connected to the word "state", in the sense of a nation or society. 'Statistics' originally meant 'systematic information about the state'.

This meaning of the word persists today; the Australian Bureau of Statistics carries out a census of the Australian population every five years, and conducts surveys to document aspects of Australian society.

What we learn here is related to this activity, but is much more concerned with a slightly different meaning of the word: 'Statistics' is also used to refer to the scientific discipline of the design and analysis of quantitative studies.

To make this clearer, we may refer to 'statistical science'.

### 1.1 The scope of statistical science

The main purpose of research is to further knowledge. In some areas, such as history and literature, this mainly involves the consideration and interpretation of documents and other writings. In other areas, such as the physical, biological and social sciences, the majority of research involves either experiments or observational studies (including surveys) which generate various amounts and types of numerical data. Some experiments and observational studies produce very little data, e.g. an experiment to measure the speed of light may result in a single value, whereas others produce vast amounts of data, e.g. the government census, or a study of internet traffic.

For large, moderate and, in some cases, even quite small studies it can be very difficult to interpret the results and the role of statistics is to provide means for extracting information from data. Statistics is also concerned with the design and conduct of studies to ensure that they provide the required information as efficiently as possible.

A rich framework of design and foundational thinking about inference is a core element of the discipline. It helps us work out good methods of making conclusions about the world in the presence of variation, and identifies ap-

---

These course notes have been steadily developed over time. Contributions to the notes have been made by Ken Sharpe, Ian Gordon, Vicky Ryan, Hok Pan Yuen, Sue Finch and Ray Watson.

proaches that are likely to mislead us or provide biased information. These are the **design tools** of statistics. They involve concepts such as bias, confounding, interaction, sampling and randomisation, foundational design principles such as intervention, experiments and observational studies and specific designs such a stratified random sample, a randomised block design or a case-control study.

The **analytic tools** range from simple ways of presenting data, such as graphs, tables and summary statistics, to complicated analyses based on sophisticated probability models. Many can be thought of as data reduction techniques which enable us to summarize data in terms of a small number of quantities which capture key inferences.

The statistical design and analytic tools developed over the years are sometimes specific-purpose in that they are only appropriate for addressing particular types of data. They are, however, *subject* non-specific in that most of them can be applied in a wide variety of contexts.

The breadth of the application of statistical science is startling. It may be used whenever quantitative information entailing variation is used to address a research question. In a university this means a high percentage of academic disciplines: not only the obvious ones such as agriculture, biology and medicine but also history, education and fine arts (and everything in between). The small fraction excluded are research questions involving abstract thought, such as philosophy and pure mathematics.<sup>1</sup>

### 1.1.1 Types of applications

Statistics is concerned with the design and analysis of quantitative studies. Many modern societies are data rich, so there is a tendency to think that the sheer volume of data must have a corresponding amount of insight. This goal is often explicitly worded this way: the objective of an analysis (typically, of a large, unstructured database) is to “gain insights”. However, this goal is vague and there remains a need for careful, thoughtful approaches to data, in which defining the question and coming up with a good design is as important as ever.

Statistical science contributes many ways of thinking about data, all of them useful. These include:

- Description, using summary measures, tables and visualisations, but also more complicated techniques such principal components analysis or other multivariate methods;
- A principled approach to study design that appropriately address research questions;

---

<sup>1</sup>Ironically, since statistics has a strong foundation of mathematics.

- Modelling, in which underlying random processes are assumed and fitted to data, to identify dominant patterns and estimate quantities;
- Sampling, to infer information of interest about a population;
- Prediction or forecasting, where data at hand are used to build predictions of observations outside the data, or in the future;
- Causal inference, in which we gain understanding about the effects of proposed causes, classically, through randomised controlled trials.

A trend in data science is to emphasise analytic tools, such as methods of analysis and software applications. But analytic tools and methods do not define a discipline. We would not define meteorology as the discipline that uses weather stations; it is concerned with fundamental questions about the patterns of weather, making weather forecasts, and related research questions about such phenomena as climate change. To understand ‘statistical science’ it is important to focus on the questions that it asks, not the set of tools it uses. The broad aims and their associated questions are the defining characteristic of the academic discipline.<sup>2</sup>

### 1.1.2 COVID-19

The COVID-19 pandemic illustrates all of these applications of statistics. The examples here are a subset only:

- Description: Time series graphs of COVID-19 reported cases and deaths in different countries have featured prominently in mainstream and social media.
- Modelling: Models of the development of the pandemic make major contributions to policy, for many purposes; most compellingly, to understand what need there could be for medical resources such as Intensive Care Unit beds and ventilators.
- Sampling: Sampling has been a topic of some controversy; without random sampling there are important population quantities that we cannot infer reliably about the pandemic; reliable assays are vital here, too.
- Prediction from the models, with associated uncertainty, is key to the policy discussions, including decisions about the extent and longevity of lockdown measures.

---

<sup>2</sup>A good discussion of these issues of the scope of statistics is: Hernan, M.A., Hsu, J. and Healy B. (2019). A second chance to get causal inference right: a classification of data science tasks. *Chance* 32:22-29.

- Causal inference arises in research on possible anti-virals and vaccines, and also on the possible effects of alternative strategies for relaxing restrictive measures.

## 1.2 Types of research studies

A framework for the ‘design’ of studies is critical in understanding and evaluating data. The word ‘design’ is being used here in the broadest possible sense, defining the method or structure of the data collection. We describe two broad types of design, with sub-categories.

1. Intentional: the data were collected with some specific purpose in mind and a plan for how the collection was to be carried out. This includes:
  - (a) Monitoring: The data were collected as part of a routine process of monitoring.
  - (b) Observational: The data were collected without a deliberate intervention to change what is observed, but with a definite research goal in mind.
  - (c) Survey: Data from a survey are usually observational, so this can be sensibly regarded as a special case of the previous category. However, a survey is a particular data collection strategy with its own methodology and theoretical framework. There are many different types of surveys and understanding these types is part of our contextual understanding.
  - (d) Experiment: Any data collection that comes from a context in which an intervention has been used to change an outcome in some way, can be regarded as an experiment. An experiment usually and desirably involves a comparison between two or more interventions. As for surveys, there is a rich body of theoretical understanding of what makes a good experiment. In particular, experiments involving humans and animals have some challenges and constraints that are not relevant to experiments on plants and inanimate objects.
  - (e) Meta-analysis, or systematic reviews: In this approach, data and results from already-conducted studies of a research question are retrieved and analysed together. The aim is to bring together all relevant inferences on the question, to improve statistical efficiency and give a comprehensive perspective.
2. Opportunistic: This means that the data available to be analysed have arisen in some other way, but not as a consequence of a conscious

plan; the data have been assembled without articulating a planned approach, and not according to a defined protocol. For example:

- (a) Data 'scraped' from the web, when the data obtained are a subset of a larger database.
- (b) Re-use of an existing data set or part of it.

These aspects of the context of the data collection bear upon the suitability of quantitative answers to particular types of questions. This link is more transparent with intentional designs than with opportunistic ones. The relationship between design and interpretation must be carefully considered. For example, although data analytics often aims to identify causes of outcomes, only particular types of designs support claims of causality.

We will discuss some aspect of the design of studies later but first introduce two simple experimental designs. These are included here for two reasons. First, they are the simplest designs that illustrate some important design concepts and are therefore paradigms for many more complex designs. Second, by including them here we emphasise the importance of good design; versions of these designs, or extensions of them, will be seen as we think about description and inference in later chapters.

### 1.2.1 Paired samples designs

A very old and natural technique for controlling important factors in experiments is to use "pairing" or "matching". The idea is that if we are interested in comparing two treatments or interventions, it is good to find pairs, where the two members of a pair are very similar. Then we can give a different treatment to each member of the pair.

▷ **EXAMPLE.** Palatal injections in removing wisdom teeth

If you go the dentist for removal of your upper backmost wisdom teeth, it is likely you will receive an injection in the gum and also an injection in the roof of your mouth called a palatal anaesthetic. The palatal injection can be quite painful and can result in swelling and discomfort after the tooth is removed.

Dr. Mark Badcock wondered if the palatal anaesthetic was really helping to reduce pain and so carried out a randomised controlled trial that compared the effects of lignocaine and saline injections on the level of pain reported during and after upper molar surgery. Patients acted as their own controls; each patient had upper molars on both sides removed but received a different palatal injection on each side. The study was double blind as neither the dentist or the patient knew which side received lignocaine.

The study assessed subjective experience of pain which can be highly variable between individuals. Hence the use of a paired design where patients

experienced both treatments and the difference in their experience under the different treatments can be measured was appropriate.

The purpose and effect of this pairing or matching is not to remove bias. Rather, it is to make the statistical inferences more efficient. Inferences are always a matter of sorting out “signal” from “noise”; when we use pairing, we are not really changing the signal: we are reducing the noise.

Paired data may arise from a “before and after” study, when measurements are taken on the same entity before and after an intervention. The interpretation of such studies can be vexed, since, necessarily, the intervention has coincided with the passing of time, and it may be therefore difficult to attribute any observed effect to the intervention. Or pairing can arise because the pairs are assembled using characteristics which are presumed to lead to observations of similar size.

Suppose a drug trial is designed to test the effect of vitamin D on rheumatoid arthritis, to be measured by time-resolved fluoroimmunoassay. This technique can measure changes in the rheumatoid factor, an autoantibody that tends to have high values in people with any type of arthritis. Pairs of participants could be assembled according to age, sex and other medication, and then one member of the pair randomised to vitamin D, and the other to placebo. In the analysis of the trial, the data would be treated as paired.

### **1.2.2 Independent samples design, with two groups**

We may wish to make an inference about the difference between two treatments or interventions, but often we do not have a readily available structure for pairing, or matching. It is more likely that there might be a large population, definable in some way, who could be used for such a study. In such cases, among a pool of potential subjects, allocation can be made at random to each of the two programs. In medicine, for example, randomised controlled trials where participants are randomly allocated to a treatment or a control (often taken to be usual care) are described as the “gold standard”. Randomisation ensures that the groups are balanced, on average, on possible causes of the outcome other than the treatment provided, and hence supports claim about the causal effect of the treatment on the outcome.

We discuss experimental design and the importance of randomisation in more detail in Chapter 16.

It is important to realize that with a set-up like this, measurements on subjects in the two groups are independent of each other, because they have no structural link with each other. In particular, they are not paired.

Independent samples can arise in other ways. Participants might be classified into groups according to some characteristic, for example, the presence of a hearing impairment. Participants might be offered the choice of

two treatments. In such cases, randomisation is not involved and so claims about the causal effect of the group membership or treatment on the outcome are difficult to justify. This is the case whenever the interventions being compared have been assigned in some conscious, non-random way; the most common of these is self-selection, leading to selection bias.

▷ **EXAMPLE.** Penicillin for bacterial infection

In 1940 Australian scientist Howard Florey and co-workers were developing penicillin in Oxford. There was some evidence that it could cure bacterial infection. Demonstrating this would be a potential finding of incredible importance, accentuated even more because it was during the Second World War when gas gangrene, bacterial infections and syphilis were killing many soldiers.

In May 1940, Florey and his team trialled penicillin in an experiment on eight mice.

They first injected all mice with lethal doses of streptococci (bacteria). Four of the eight mice also got the treatment, penicillin. Two of these received a single injection, and two received five injections. Florey is reported to have phoned his colleague, Margaret Jennings and said, “It looks like a miracle”. The controls had died within a matter of hours whereas the treated group lived for a minimum of two days.

But if penicillin seemed so promising, why not inject all eight mice with it? Florey understood the principle of a simultaneous control group: a comparison. This first study showed a dramatic result but was a relatively small experiment. And — even given the very small sample size — he included two doses of the treatment, to investigate a possible effect of the level of dose.

### 1.3 Types of data

The classification of data types is handled in a variety of ways. We may think of the type in terms of the area of application, such as medical or anthropological data. Sometimes the way the data are stored is the focus, such as a sound file, a video file or a database. Or we may think of the type of data in terms of its structure, such as spatial, chronological, relational or hierarchical. All of these perspectives are useful.

Data analytic tasks very often have to deal with data that has been digitised in some way; it is very common for the input to be an Excel file or text file with variables in columns and units of observation in rows, with many variations on this theme.

This basic form has an implicit structure of several or many variables, and in what follows we consider the important types of data that we encounter

in such variables.

It is useful to distinguish between two major types of data, quantitative and qualitative data, each of which can be further sub-divided into sub-categories as indicated below. Though some statistical tools can be used on data from more than one sub-category, others have been designed for use on data from a particular sub-category and it can be inappropriate to apply them to other types of data.

- qualitative → quantitative
- nominal or categorical → ordinal → discrete → continuous

A **nominal or categorical** variable has particular values that are not quantitative at all. The labels we give to these categories may be letters (A, B, C ...) or words (Melbourne, Sydney, Brisbane ...); if we use numbers as the labels for the categories this is only a convenience, and does not imply a numerical measure, or even an order in the categories.

Examples: sex, plant species, country of birth.

An **ordinal** variable is one whose values are in a meaningful order, without necessarily being on an interval scale. Such a variable is more than categorical, since the categories are in an order, such as “excellent”, “good”, “fair”, “poor”. For such a variable, the use of numbers may reflect the ordering, but we need to be careful, because the difference between two adjacent categories is not guaranteed to be constant at all. This makes averaging an ordinal variable, for example, a dubious procedure.

Consider a rating of a person’s condition that allows the following categories:

1. healthy, no evident disease or disability
2. unwell, usual behaviour is restricted by disease or disability
3. seriously ill and requiring constant medical attention
4. comatose
5. dead

This is an ordinal scale. It would be unreasonable to treat it as numeric, so that the “distance” between 4 and 5, for example, is the same as that between 1 and 2. There is a large “jump” between 4 and 5.

Examples of ordinal variables: severity of disease, Likert scale responses to a proposition (“strongly agree”, “agree”, “neutral”, “disagree”, “strongly disagree”).

A **discrete** variable takes distinct numerical values on a meaningful scale. It is almost always associated with counting in some way. Such variables can, in principle, be measured without error.

Examples: family size, number of typographical errors on a page, road toll.

**Continuous** variables take any value within a range of possible values. Because of the rounding inherent in the measurements of any quantity, there is a sense in which no actual variable in research is continuous: they are, at best, discrete. We can think of time as a continuous variable conceptually, but in fact we will measure it with error, and therefore record it to the nearest year, month, day, minute, second, or .... Nevertheless, some variables are measured with a fine enough scale that it is sensible and useful to treat them as if they were actually continuous. And the same applies to variables that are actually discrete but recorded to a fine level of detail, such as money.

Examples: time, blood pressure, weight, pressure, money.

For categorical and ordinal variables the descriptions of the categories are needed for meaning. For discrete and continuous variables, the numerical size of the variable and its units are vitally important. In many contexts, the units of a numerical variable are clearly implied, even by the name of the variable (e.g. "year"). If they are not, however, they must be specified: a distance of 72.6, without its units and devoid of context, cannot be properly interpreted.

The distinctions between the types of data are important because some statistical methods appropriate for one type are not appropriate to another. To take an extreme example, suppose that we used the following code for the variable country of birth:

1. Australia
2. U.K.
3. Europe (other than U.K.)
4. Asia
5. Other

The data recorded for this variable might consist of the number 1, 1, 2, 5, 1, 1, 1, 1, 3, 1, 1, 4, 1, ... But the average of these numbers is totally meaningless, because country of birth is a genuinely categorical variable and we cannot therefore interpret the numbers assigned as anything other than a convenience, usually for data entry purposes.

A less extreme example is the analogous situation for an ordinal variable. For example, the numbers 1 to 5 may be applied to the categories "strongly agree", "agree", "neutral", "disagree", "strongly disagree". Averaging in this case is also inappropriate, because it relies on the assumption that the difference between any of the categories is directly proportional to the number of gaps between the categories. It implies, for example, that the difference between "strongly agree" and the other end of the scale, "strongly disagree", is exactly four times greater than the difference between "strongly agree" and "agree". This is a very strong assumption that would usually be insupportable.<sup>3</sup>

---

<sup>3</sup>The University of Melbourne has analysed its survey on "Quality of Teaching" making

Our aim is to explain how and when to use some of the more commonly used statistical design and analytic tools and to explain the principles upon which they depend. The emphasis will be on understanding and applications rather than on formal mathematical derivations.

We hope that you will learn to:

1. recognize when statistical methods are required;
2. use a range of statistical tools;
3. understand the statistics reported in many (non-statistical) research papers;
4. recognize when it is necessary to consult a statistician.

There are many introductory statistics text books. It is a good idea browse these; you are likely to find one that suits your needs and tastes.

The following are some you might consider:

- Moore and McCabe: *Introduction to the Practice of Statistics*. This was one of the first of the new generation of introductory texts, focussing more on insight and understanding, and with a good deal of enrichment material.
- Altman: *Practical Statistics for Medical Research*. One of several texts designed for those in medical fields.
- Kirkwood and Sterne: *Essential Medical Statistics*. More advanced than Altman's book.
- Welham, Gezan, Clark and Mead: *Statistical Methods in Biology: Design and Analysis of Experiments and Regression*. A text for agriculture and biology with relevant examples.
- Utts and Heckard: *Mind on Statistics*. An excellent book on broader issues of statistical literacy, with many interesting examples and case studies.

---

precisely this assumption.

## 1.4 Exercises

*These problems are intended to enhance your learning in Statistics for Research Workers, by doing some statistical analysis. There are answers provided, but of course you'll benefit most from the problems if you think about them thoroughly before you look at the answers. You will need to use Minitab quite a lot; there are instructions to help you with that, since the primary aim is to learn statistics.*

*Notation:*

*\* denotes an optional problem. These problems are not required for assessment, but they will usually involve some extension of ideas or concepts. You might attempt these problems if you have done the set problems and you have some free time.*

*Computer instructions are generally indicated by square brackets.*

- 1.1 Consider the following studies, all of which involve a proposed music therapy intervention, intended to alleviate anxiety in communities of elderly people. Consider the merits and any problems with each design.
  - (a) A music therapist introduces the program in a residential aged care facility, and measures participants' anxiety before and after the therapy program.
  - (b) The program is offered to residential aged care facilities with some government sponsorship. Thirty-seven (37) facilities take up the offer, and the participants' anxiety is measured before and after the therapy program in each of the 37 homes.
  - (c) A comparison is made between a facility in which the program is used, and a nearby one in which it is not used; about 50 people are measured in each facility.
  - (d) Five facilities in Tasmania use the program. These are compared with five in New South Wales (NSW) that do not. There are about 150 participants in each state. The same anxiety measures are used in all ten facilities.
  - (e) Twenty facilities are recruited for study and consent is obtained from them to be randomised to either receiving the program, or an alternative control program involving a typical group meeting for games in a social environment.

- 1.2 Open the MINITAB worksheet OZ\_CARS.mwx.

- (a) Examine the worksheet.

How many rows have entries in them? How many variables are there? What is the worksheet about? What are all the variables? What are the units of the variables?

*Part of the point of this first problem is to illustrate the way that data are often presented to statisticians. The advent of Excel, in particular, has meant that researchers may choose to provide data to statisticians in essentially the form they have stored it themselves. However, a helpful data storage format, which might include text boxes, charts, and other information, is not of the standard format required for statistical analysis. When presenting data to a statistician, a researcher should expect to have to work on arranging the data in a suitable form for analysis, including descriptions of variables and the meaning of codes used. There are a host of minor considerations which can be time-consuming for a statistician to deal with, such as missing data, unclear variable definitions, inconsistent coding, data entry errors, text where there should only be numbers, etc., and it is helpful if researchers can learn about the need to present data in the right form for analysis.*

- (b) The data come from a study of the important pollutants from Australian cars in the early to mid-1990s. A survey was carefully designed so as to produce a representative sample of the main models of cars in use at the time, and the surveyed cars were measured in various ways, tuned, and then re-measured. Does this information help you to say what some of the variables might be? And their units?
- (c) Note the row of column labels above the columns, starting with C1-T; some have a “T” suffix, one has a “D” suffix and several have no suffix (e.g. c6). The “T” stands for “text” and denotes a categorical variable. The “D” stands for “date”. Note the categorical variables in the data. [Use Stat > Tables > Tally Individual Variables in Minitab to look at the categorical variables.] How many Fords are there in the data set? How many H/TOPs?
- (d)\* The survey was designed to provide data to assess the effect of tuning on car performance. This national survey was conducted in Melbourne and Sydney. Why do you think the national survey was carried out in only two cities? Is this an issue?
- (e)\* Households participating in the survey may have had more than one car, but only one car per household was eligible. Suggest a suitable method for selecting one car that can be reliably implemented by asking questions over the phone.

## 1.5 Answers

1.1 An overall design aspect here is the ‘unit of observation’, something we must always consider. When an intervention is applied to a group in the one environment, the unit is the group as a whole, not an individual in the group. There can and would be many features of a residential aged care facility that would influence the outcome, such as the general tone and demeanour, the quality of the staff and the living conditions, the dynamics of the place over the period of the program, and so on.

- (a) *A music therapist introduces the program in a residential aged care facility, and measures participants' anxiety before and after the therapy program.*

This study has an effective sample size of one. It is also a ‘before and after’ study, which means it can be readily influenced by phenomena other than the intervention; for example, a much-loved staff member leaving during the period of the program being provided.

- (b) *The program is offered to residential aged care facilities with some government sponsorship. Thirty-seven (37) facilities take up the offer, and the participants' anxiety is measured before and after the therapy program in each of the 37 homes.*

The facilities all self-selected themselves to have the program. There could definitely be an association between properties of the homes that are willing or keen to have the program, and the success of the program. For example, such homes could be better resourced and organised than average, and more likely to cooperate helpfully with the program provider. So we could easily get a result from this study that was more positive than would be the case across all facilities.

- (c) *A comparison is made between a facility in which the program is used, and a nearby one in which it is not used; about 50 people are measured in each facility.*

There are about 100 people in this study; this sounds like a reasonable sample size. But the unit of intervention is the facility as a whole, so the relevant sample size for replication is two, which is inadequate. There is no way to tell if the difference between the results at the two facilities is due to the program, or other differences between the facilities. With only one facility in each intervention group, this is a fatal design error.

- (d) *Five facilities in Tasmania use the program. These are compared with five in New South Wales (NSW) that do not. There are about 150 par-*

*ticipants in each state. The same anxiety measures are used in all ten facilities.*

There are some improvements here. But we are left with ‘confounding’ between state and program; we can’t tell if the comparison’s results are due to the program, the state, or a mix of these.

- (e) *Twenty facilities are recruited for study and consent is obtained from them to be randomised to either receiving the program, or an alternative control program involving a typical group meeting for games in a social environment.*

This study has several good features: random allocation of the intervention, a sensible comparison intervention, and some replication: ten facilities in each intervention group.

- 1.2 (a) There are 181 rows, each containing data on a single vehicle. There are 17 variables, one per column. The worksheet appears to be about some Australian used cars and some of their basic features, such as make, model, date (of manufacture?), mass (in kg?), odometer reading (in kilometres?) and some other variables. The identity and units of the other variables are not clear in isolation, although you might guess that the last two variables are fuel economy (in units of litres/100 km, perhaps?).

The point of this first part of the exercise is that “the data”, even in a MINITAB data file, are seldom enough for a proper statistical analysis. We need to know a lot more about the variables, including what was measured and what the units are. For a valid inference, we usually also need to know how the data were collected, for what purpose and so on.

- (b) This information helps a bit. If we get imaginative with the column headings of the variables after cylinders, we might be able to work out that hc\_pre refers to “hydrocarbon emissions, pre-tuning”, and hc\_post is the analogous post-tuning variable. CO is carbon monoxide, NOx is nitrogen oxides. But we would have little idea of the units of these variables; in fact, the units are grams/kilometre. The last two variables are fuel economy, measured in litres/100 km. In general, in addition to the data set, a list of the variables, their meaning, units (if numeric) and codes (if categorical) needs to be supplied, to allow a sensible statistical analysis.

- (c) There are 57 Fords and 1 “H/TOP” (whatever that is, presumably, a “hard-top”).

- (d)\* Presumably the ‘national’ survey was only carried out in two cities to limit the cost. However a survey of only two major cities

is not really a national survey. There may be, for example, differences between cars in urban and rural locations that could be important.

- (e)\* The method needs to be random or pseudorandom (i.e. effectively random). It should not be based on a judgment of the interviewer or the interviewee. Nor should it be related to characteristics of the car that might be associated with the effectiveness of the tuning. Asking for the newest car, for example, would potentially introduce biases into the sample. In the study, they asked for the car with the registration renewal due soonest, believing that this was pseudorandom.



## 2 The language of statistics: describing, summarising and visualising

The idea that *mathematics* is a language is common and useful. Symbols are used, which must be learned. Propositions are expressed using these symbols that reflect logical arguments and proofs. There is a grammar to the mathematical expressions and a vocabulary and lexicon of terms. This facilitates concise and precise communication of mathematical truths and arguments.

The same concept applies to *statistical science*. To communicate about data, a language is required, with agreed terms, symbols and basic representations. There are levels of sophistication in the language, as we shall see. In this chapter we consider some of the foundational elements of that language; it is useful to think of these as important statistical ‘words’. In some cases they are actual words, like ‘mean’ and ‘boxplot’. Symbols are also used. Data involve patterns, which leads to visual representations — graphs — being an essential and basic part of the statistical language.

Thoughtful data exploration and representation can provide the basis for good understanding of the main features of data, and can support appropriate statistical modelling. Data exploration is often taught in some way at many levels of education, including quite junior ones; it may be thought that the topic is trivial, easy or basic (“bar charts, mean, mode, median”). However, there are good and bad ways to present data succinctly and clearly, and developing skills that support *good* communication about data is foundational for anyone working in statistical data science. The area of graphical representation of data is evolving as new ideas and technology emerge, and is an active area of research.

Broadly speaking, we can represent data using:

- numerical summaries, and;
- graphs.

Both of these are heavily used in statistical inference. For this reason alone, studying data presentation is important.

### 2.1 Quality data presentation

Researchers have investigated good and bad ways of representing data, using a mix of empirical research and creative flair. An example of the kind of research carried out is given in Exercise 2.2. A ground-breaking and award-winning book on the topic is “*The Visual Display of Quantitative Information*”, by Edward Tufte (Graphics Press, Connecticut, 1983), and includes such memorable terms and expressions as “chartjunk” and the “data-ink ratio”, the latter defined as the ratio of the ink used to represent data to the to-

tal ink used to print the graphic (high in good graphics). “*The Elements of Graphing Data*”, by William Cleveland (Summit Hill, NJ, Hobart Press, 1994) develops a theory of graphical representation, based on empirical research. A text with a more modern focus is “*The Functional Art : An Introduction to Information Graphics and Visualization*”, by Alberto Cairo (New Riders, 2012).

Many of the principles that these writers have espoused amount to thoughtful common sense; others are less obvious and have arisen out of their research. Some graphics produced by commonly-used software adhere to these principles reasonably well, but others do not. In general, the default graphics from most software can be improved.

Edward Tufte wrote: “Data graphics are paragraphs about data.” Tufte was writing about the integration of data and text, but it is also possible to see the analogy the following way. A good graph should be about a single idea (the “paragraph”), and contain the data (the “words”) arranged in coherent and meaningful ways (the “sentences”).

Guidelines for good practice in data visualisation are:

- A good graph has clear and informative labelling. This applies to both the caption and the parts of the graph. The reader should not have to guess what anything means, or interpret the meaning of an abbreviation. Where it is reasonable and possible to do so, the units of a variable should be included.
- For the purposes of comparisons in graphs, as much as possible, line up the features to be compared along a common scale. Research shows humans are best at visual comparison when a common linear scale is used, as opposed to many other possible ways of representing things to be compared, for example, by using volumes, or angles, or lengths not lined up along a common scale. This is what Exercise 2.2 is about.
- Minimize the amount of ink, including colour shading, by eliminating the use of ink that does not communicate anything meaningful.
- Avoid distortions from spurious use of perspective and other artistic tricks.

Throughout, we present many examples of graphs that follow these general guidelines. Five principles of good graphics and the rules of thumb that follow from them are discussed in Section 3.2.

The goal of high quality data presentation includes the presentation of summary statistics and results of more formal analysis, usually in tables; this is discussed in Section 3.3.

### 2.1.1 Graphs from software packages

There is quite a lot of variation in the quality of graphs produced by statistical packages. The default provided by a package for a particular type of data or the default settings for a particular type of graph may not follow the guidelines and principles for good graphs espoused in this chapter. You can almost always improve on the defaults set in a package. You should be careful if using software that automatically generates graphs or that uses generic ‘drag-and-drop’ graphing functions. The lack of transparency about the way the data are summarised or the way the graph is constructed can result in misinterpretations of the information provided by the graph. In general, it is preferable to use software that allows you to specify the type of graph you require.

It is possible to get good graphs easily in MINITAB. However, sometimes a small amount of modification is required to adhere to the guidelines above; happily, for the most part, MINITAB allows these modifications.

Most of the graphs in these notes were produced in MINITAB, sometimes with some editing. We have changed the default appearance used by MINITAB to be more consistent with the principles above. In MINITAB 19 this is done by **File → Options**, and making choices under **Graphics** and **Individual Graphs**.

## 2.2 Visualising numerical data

In many applications a numerical (discrete or continuous) variable is the primary focus of interest; this might be considered to be an outcome variable, such as a measure of health in a medical trial. Here we consider visualisations that support our understanding of the distribution of the data, and that facilitate exploring the relationship between two numerical variables. Many of our examples come from the **Countries** data set, where the numerical variables of interest include life expectancy, fertility and rates of HIV infection.<sup>4</sup>

### 2.2.1 Understanding distributions

#### ▷ EXAMPLE. Countries data (MINITAB worksheet: countries.mwx)

The **Countries** data set came from the “EarthTrends” website, which was maintained for many years by the World Resources Institute. The data were accessed in 2008; one of the variables includes life expectancy at birth for

<sup>4</sup>“HIV” is the human immunodeficiency virus, the cause of acquired immunodeficiency syndrome (AIDS).

people born between 2000-2005, and also between 1980-1985, for each country.

Here are the data for life expectancy at birth in 1980-85:

Armenia	72.5	Bosnia and Herzegovina	70.7	Afghanistan	40.0	Gabon	56.3	Dominican Rep	62.8
Azerbaijan	68.4	Bulgaria	71.2	Algeria	60.5	Gambia	44.1	El Salvador	56.6
Bangladesh	50.0	Croatia	70.5	Egypt	56.5	Ghana	53.6	Guatemala	58.0
Bhutan	47.7	Czech Rep	70.7	Iran, Islamic Rep	59.7	Guinea	40.2	Haiti	51.8
Cambodia	52.1	Denmark	74.6	Iraq	62.3	Guinea-Bissau	39.1	Honduras	60.8
China	66.6	Estonia	69.6	Israel	74.5	Kenya	55.7	Jamaica	71.2
Georgia	70.7	Finland	73.9	Jordan	63.7	Lesotho	52.0	Mexico	67.5
India	54.9	France	74.7	Kuwait	71.3	Liberia	44.9	Nicaragua	59.3
Indonesia	56.2	Germany	73.8	Lebanon	65.9	Madagascar	48.0	Panama	70.5
Japan	76.9	Greece	75.2	Libyan Arab Jamahiriya	62.2	Malawi	45.7	Trinidad and Tobago	70.2
Kazakhstan	67.0	Hungary	69.1	Morocco	58.3	Mali	44.4	Argentina	70.0
Korea, Dem People's Rep	69.1	Iceland	76.8	Oman	62.7	Mauritania	47.4	Bolivia	53.9
Korea, Rep	67.2	Ireland	73.1	Saudi Arabia	62.6	Mozambique	42.8	Brazil	63.0
Kyrgyzstan	65.6	Italy	74.5	Syrian Arab Rep	62.5	Namibia	55.2	Chile	70.6
Lao People's Dem Rep	45.8	Latvia	69.3	Tunisia	64.9	Niger	40.7	Colombia	66.6
Malaysia	68.0	Lithuania	70.8	Turkey	62.3	Nigeria	48.1	Ecuador	64.3
Mongolia	57.5	Macedonia, FYR	69.6	United Arab Emirates	68.6	Rwanda	46.1	Guyana	61.0
Myanmar	51.8	Moldova, Rep	64.8	Yemen	49.1	Senegal	46.3	Paraguay	67.1
Nepal	49.1	Netherlands	76.0	Angola	40.0	Sierra Leone	35.3	Peru	61.4
Pakistan	53.0	Norway	76.0	Benin	49.2	Somalia	43.0	Suriname	67.1
Philippines	62.1	Poland	70.9	Botswana	62.8	South Africa	57.7	Uruguay	70.8
Singapore	71.8	Portugal	72.2	Burkina Faso	46.1	Sudan	49.1	Venezuela	68.6
Sri Lanka	67.9	Romania	69.7	Burundi	46.6	Tanzania, United Rep	51.0	Australia	75.2
Tajikistan	65.9	Russian Federation	68.3	Cameroon	50.7	Togo	50.2	Fiji	64.7
Thailand	65.0	Serbia and Montenegro	70.2	Central African Rep	46.5	Uganda	47.2	New Zealand	73.7
Turkmenistan	63.2	Slovakia	70.6	Chad	42.3	Zambia	52.0	Papua New Guinea	49.7
Uzbekistan	66.6	Slovenia	71.2	Congo	56.8	Zimbabwe	59.6	Solomon Islands	60.6
Viet Nam	58.7	Spain	75.8	Congo, Dem Rep	47.1	Canada	75.9		
Albania	70.4	Sweden	76.3	Côte d'Ivoire	50.0	United States	74.0		
Austria	73.1	Switzerland	76.2	Equatorial Guinea	43.8	Belize	71.2		
Belarus	70.7	Ukraine	69.1	Eritrea	43.3	Costa Rica	73.5		
Belgium	73.7	United Kingdom	74.0	Ethiopia	42.7	Cuba	73.4		

Data presented in this way can be difficult to interpret and various methods for organising, summarising and displaying such data have been developed to help with the interpretation.

The graphical displays of data we consider here are meant to represent the “distribution” of the data. Sometimes we may call this the “empirical distribution”, to distinguish it from the “theoretical distributions” we meet later. An empirical distribution shows the values observed, together with their corresponding frequencies. This can be shown in several ways, not all of them giving the same level or amount of information. We may think of an empirical distribution as a way of looking at the variation in the data, while overlooking the fine detail of the individual observations. In the above table, we may notice, for example, that Slovenia and Jamaica have the same life expectancy, namely, 71.2 years. In some contexts this might be an interesting fact. But the specific nature of this information is not a feature of the various representations of a distribution, because it is the overall pattern that is the point.

Useful visualisations for a single numerical variable such as life expectancy, include the:

- histogram
- dotplot (and its close relation, the individual value plot), and
- boxplot.

We discuss the first two of these in this section.

### Histograms

A histogram is a pictorial display which gives the frequencies (or relative frequencies) of various categories or intervals.

The essential feature of a histogram is that the area of the bars are proportional to the corresponding frequencies.

Histograms enable us to see the ‘shape’ of a data set; whether it is symmetric or skewed, whether it is unimodal, bimodal or multimodal, whether it has short, medium or long tails and whether there are any outliers. If data are grouped in intervals to draw a histogram then it is necessary to take care when choosing the number (and width) of intervals; too many intervals and the histogram will be too ragged, too few intervals and the histogram will be too smooth. Either way, the main features of the data will not be adequately displayed. In most cases 10–15 intervals is desirable.

There is no standard convention for dealing with the endpoints of the bars of a histogram. In Figure 1, a bar with endpoints  $a$  and  $b$  say, corresponds to values that belong to the interval  $[a,b)$ , i.e.  $a$  is included and  $b$  is excluded.

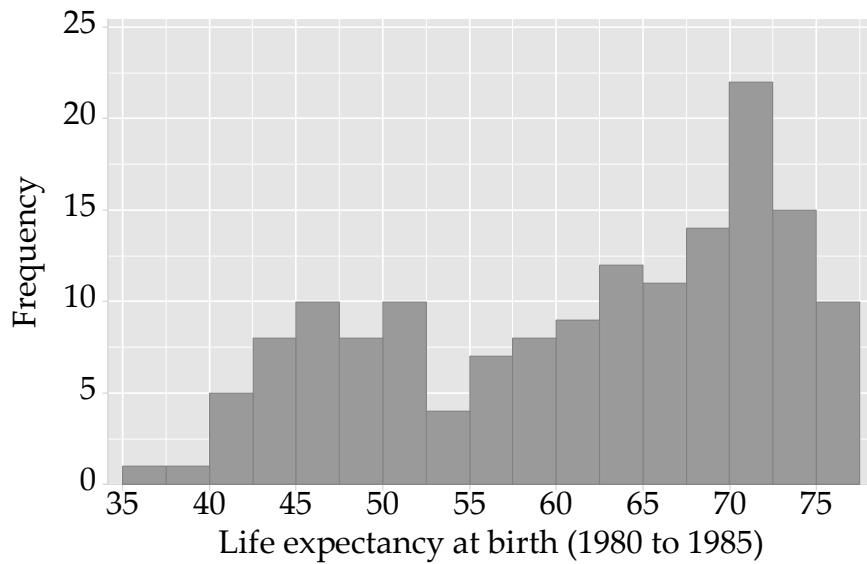


Figure 1: Histogram of life expectancy at birth in 1980-1985 in 155 countries

## Dotplots

The dotplot is an alternative to the histogram that is much more useful when the data set has a small to modest size. For such a data set we can comprehend the whole distribution of the data without any grouping, and this is what the dotplot represents. It shows every point in the sample, to the level of resolution possible, by putting a dot at a point at which there is an observation, and stacking dots above one another if required.

Figure 2 is the dotplot for the life expectancy data shown above. Close inspection shows that the resolution in this dotplot uses the nearest whole number (... 60, 61, 62, ...).

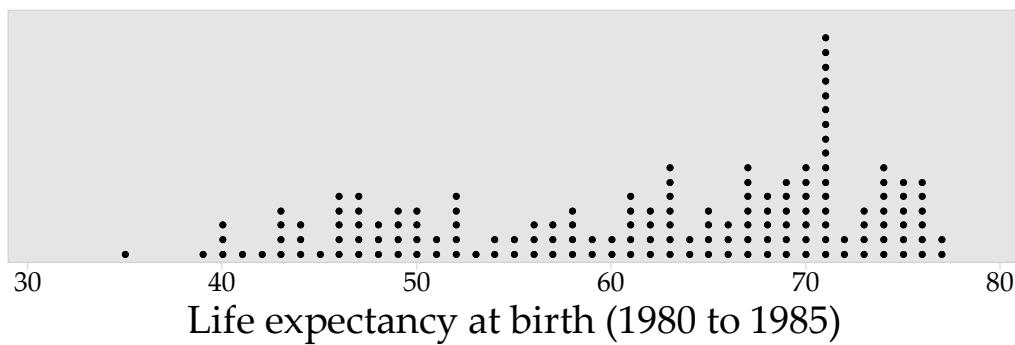


Figure 2: *Dotplot of life expectancy at birth in 1980-1985 in 155 countries*

## Individual value plots

An **individual value plot** is a variation on a dotplot. Usually a dotplot is constructed by stacking dots; this becomes cumbersome if there are many data points. An individual value plot either “collapses” the dotplot or randomly “jitters” the points in the direction perpendicular to the axis showing the scale of the variable. The jittering makes the points more distinguishable, without misrepresenting the distribution. Some software provides a separate option to create such a plot; others provide this functionality by editing a standard dotplot. Figure 3 illustrates a collapsed and a jittered version for the life expectancy data.

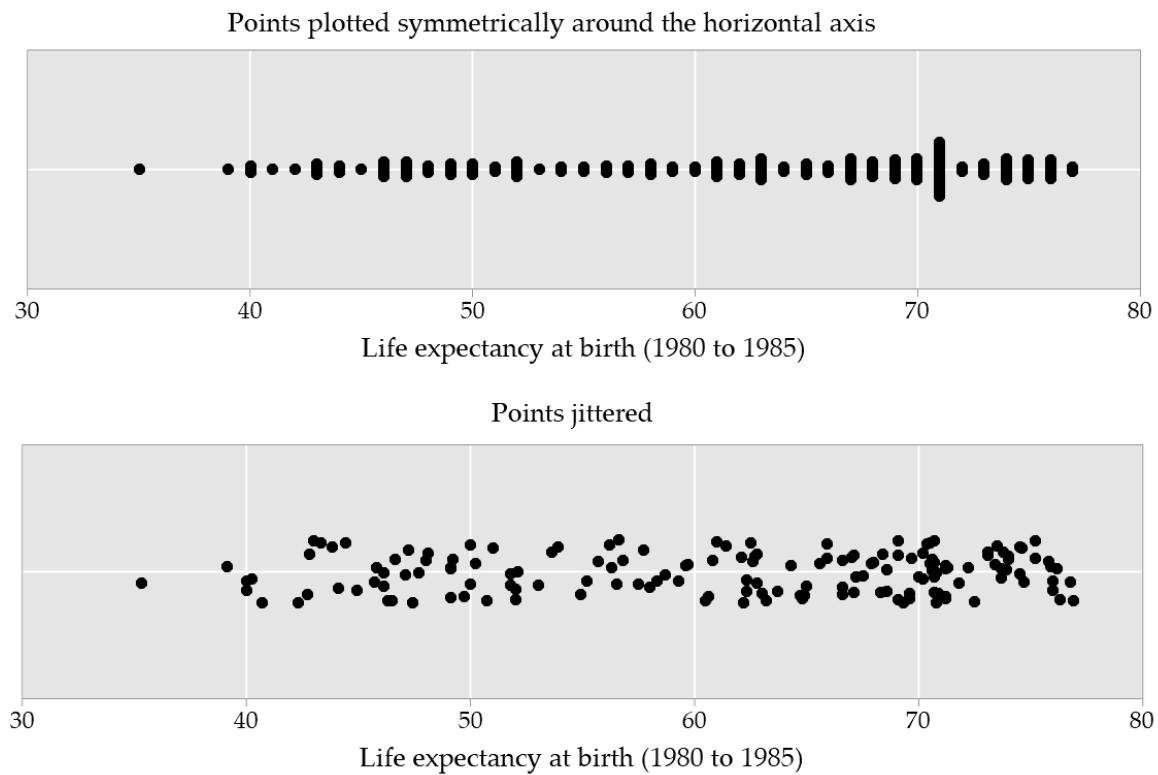


Figure 3: Two individual value plots of life expectancy at birth in 1980-1985 in 155 countries

As we will illustrate in Section 2.5, individual value plots are most useful in efficiently displaying many data points broken down by groups. However, it should be noted that there is a compromise — individual value plots (particularly if jittered) generally do not provide the same insight into features of the distribution as histograms or dotplots. Hence for plotting a single variable alone, like life expectancy at birth, we would rarely use an individual value plot.

### 2.2.2 Exploring relationships

Bivariate data, as the name suggests, are data in pairs, where we simultaneously consider two variables. Here we consider ways of using visualisation to explore the relationship between two numerical variables.

Useful visualisations include the:

- scatterplot
- line plot

- scatterplot with marginal boxplots.

We discuss the first two of these in this section.

### Scatterplots

▷ **EXAMPLE. Countries data** (MINITAB worksheet: countries.mwx)

The **Countries** data set includes two variables measured for the years 2000 to 2005: life expectancy and fertility rate (number of children per woman).

We can examine the relationship between these two numerical variables with a **scatterplot**, as shown in Figure 4. A scatterplot is used to plot bivariate data for which it is possible to have more than one  $y$ -value for each distinct  $x$ -value.

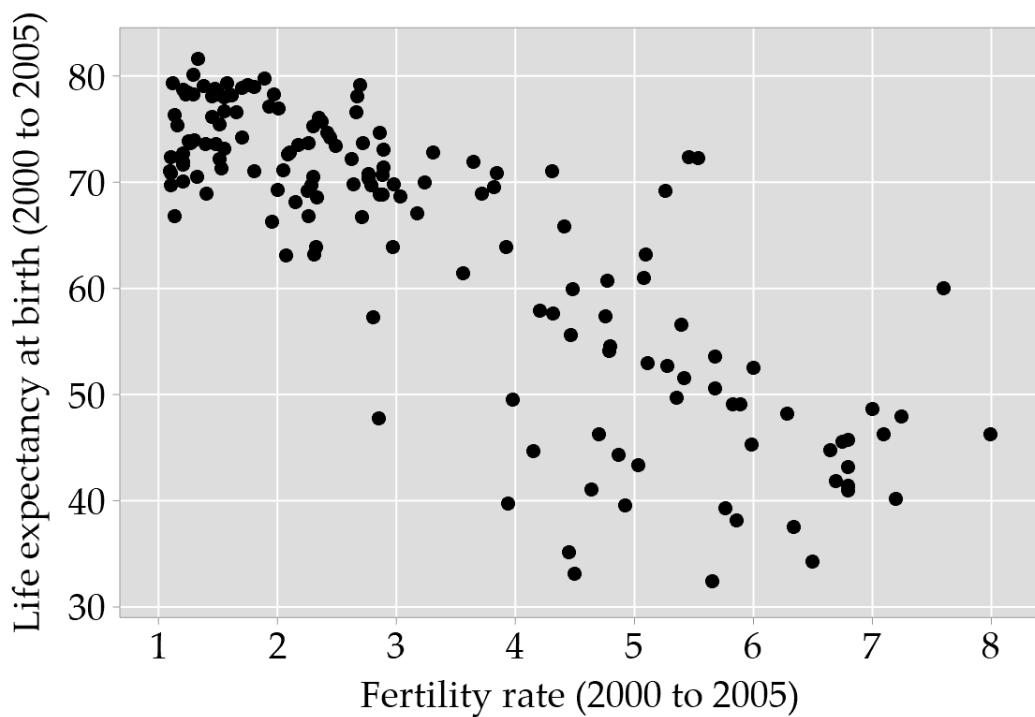


Figure 4: Scatterplot of fertility rate versus life expectancy at birth in 2000 to 2005 in 155 countries

We see that countries that have a higher number of children per woman have a lower life expectancy.

Examining a scatterplot is often the first step in fitting a curve (often a line) using a parametric model for the data. This is covered in Chapter ???. A useful embellishment of the scatterplot is a scatterplot "smoother", which aims to put a curve through the data without assuming a particular model, to suggest an overall relationship descriptively. Figure 5 shows the data plotted in Figure 4 with a smoother added; this suggests that the relationship is roughly linear.

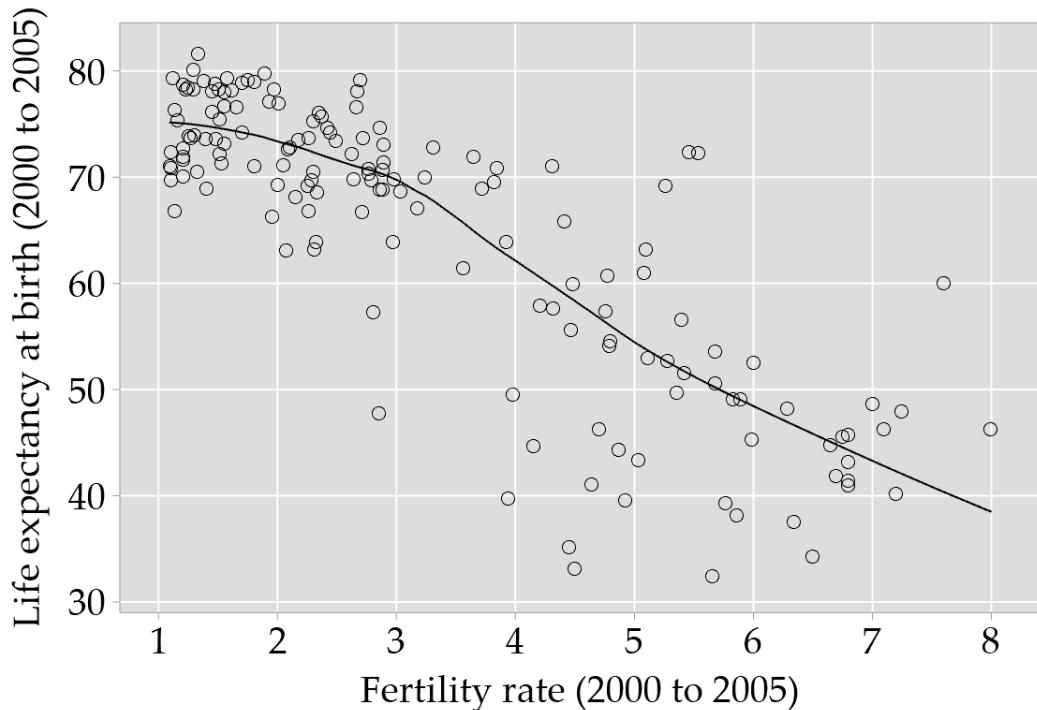


Figure 5: Scatterplot of fertility rate versus life expectancy at birth in 2000 to 2005 in 155 countries, with a smoother added

If we are only interested in exploring the association between the variables, then it doesn't really matter which variable goes on the horizontal axis and which goes on the vertical axis. However, the labelling of the variables has some importance when we come to modelling a numerical variable in terms of another numerical variable. In this framework, sometimes referred to as prediction, the practice is to plot the predicted variable on the horizontal axis.

Further, when one of the variables is time, the standard practice is for time to be on the horizontal axis and the other variable to be on the vertical one (for example, Figure 7 below).

A variable that is used for prediction (and sometimes is under experimental control) is often referred to the explanatory variable and the predicted is called the response variable. Sometimes the response variable is called the dependent variable and the explanatory variable is called the independent variable, but these (the latter, in particular) are not very helpful names.

In sum, a scatterplot is always drawn with the explanatory variable on the horizontal axis, and the response variable on the vertical axis.

## Line plots

Sometimes one of the numerical variables of interest has a natural ordering; time is a common example. If one of the variables (the  $x$ -variable) has such a natural ordering and there can be only one  $y$ -value for each distinct  $x$ -value, a suitable graph is a **line plot**, like that shown in Figure 6. In this context, it is appropriate to put time on the  $x$ -axis of the graph and to join the points with a line.

### ▷ EXAMPLE. Ice core data

The data come from ice core samples obtained and analysed in the 1990s from the Law Dome, near Casey Station, Antarctica. The measurements are of carbon dioxide concentrations from air samples in the past, trapped in the Antarctic ice.

Source: [cdiac.ornl.gov/trends/co2/lawdome.html](http://cdiac.ornl.gov/trends/co2/lawdome.html).

Figure 7 shows a second example which includes the Law Dome data described above and data from another site at Vostok, also in Antarctica. These can be found at: [cdiac.ornl.gov/trends/co2/vostok.html](http://cdiac.ornl.gov/trends/co2/vostok.html). This plot puts the rapid rise of the last 200 years or so in more historical context.

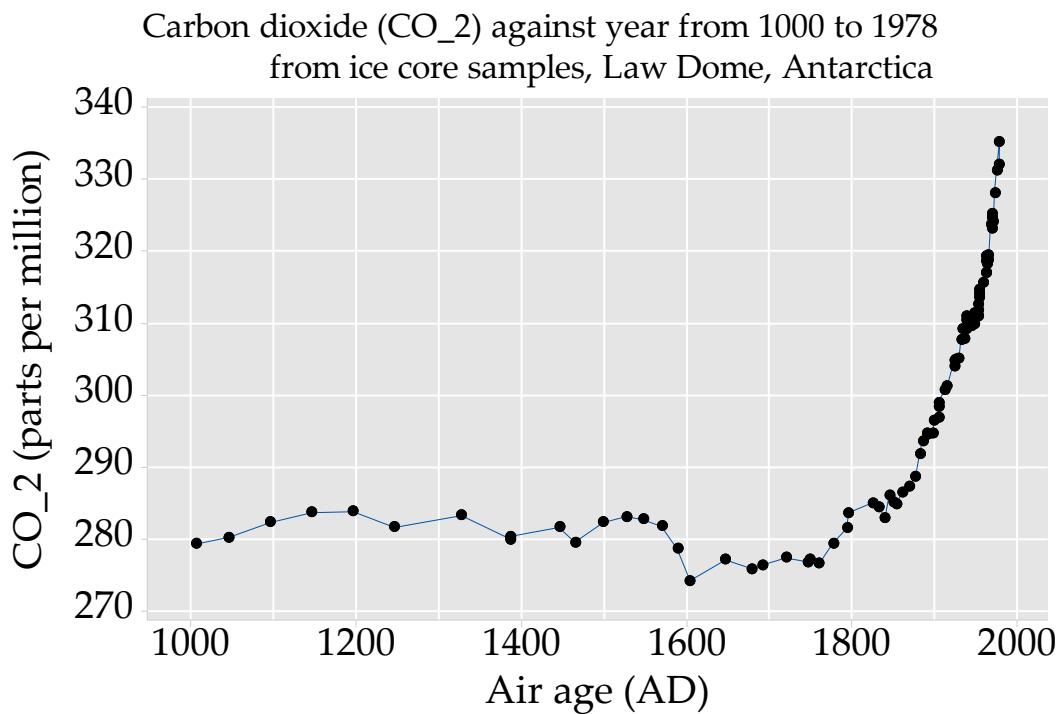


Figure 6: Carbon dioxide levels by year: Law Dome samples, Antarctica

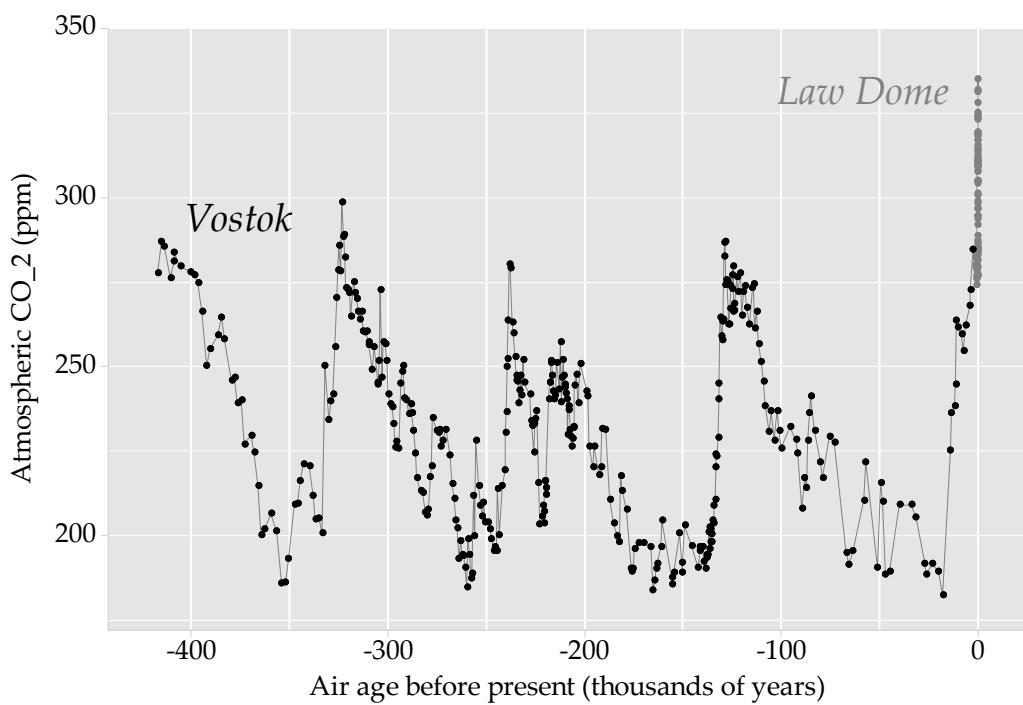


Figure 7: *Carbon dioxide levels by air age before present: Law Dome and Vostok samples*

## 2.3 Summarising numerical data

The two most important concepts in summarising numerical data are ‘location’ and ‘spread’.

Location refers to the position of the data, in a general sense, along the number line. Measures of location are often referred to as measures of central tendency. The mean and the median are both measures of location. The quartiles also measure location, but not central tendency.

Spread, as the name suggests, refers to the variation or dispersion in the data: how spread out it is. The standard deviation, the variance and the interquartile range are measures of spread. Measures of spread are sometimes called measures of scale.

When data are analysed, it is common to use these measures; they are fundamental features in much practical data analysis. We will look at these measures in more detail in a moment, and illustrate how we commonly measure location and spread with either:

- the mean and standard deviation, respectively, or
- the median and interquartile range, respectively.

### 2.3.1 Measures of location

First we provide some notation in order to define the measures of location.

We typically label a set of  $n$  data values for a numerical variable  $x$  thus:  $x_1, x_2, \dots, x_n$ . The subscript does not denote a size ordering; it just reflects the order in which the data were written down, which might be the order in which they were collected.

The most commonly used measures of location and central tendency are:

- mean (or average); usually denoted by  $\bar{x}$ . The mean is given by:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}.$$

- median ( $M$ ); arrange the values in ascending order and the median is given by the ‘middle’ observation or the value at a ‘depth’ of  $\frac{1}{2}n$  from either end of the data set. If  $n$  is even the median is taken to be the average of the two middle values, i.e. the average of the two values which are at a depth of  $\frac{1}{2}n$ .

There are other measures of location which are not measures of the centre of a set of values; these include:

- the minimum and maximum;
- other such measures are percentiles and quartiles, whose meaning, at least approximately, is suggested by their names. The 15th percentile of the data is the point below which 15% of the data falls, and so on.

Two important measures of this type are the lower and upper quartiles,  $Q_1$  and  $Q_3$ . The median is the second quartile, or  $Q_2$ , but this name is not used for it.  $Q_1$ , the median and  $Q_3$  divide the data into four quarters. That is the essential idea, but in practice there are a number of alternative ways of defining  $Q_1$  and  $Q_3$ , and they are used in different software.

We use the MINITAB definition, which is a sensible one. To calculate quartiles, first order the data from smallest to largest. So there is an observation at position  $1, 2, \dots, n$ . Then  $Q_1$  is at position  $(n+1)/4$  and  $Q_3$  is at position  $3(n+1)/4$ . If the position is not an integer, interpolation is used. For example, suppose  $n = 10$ . Then  $(10+1)/4 = 2\frac{3}{4}$  and  $Q_1$  is between the 2nd and 3rd ordered observations (call these  $x_{(2)}$  and  $x_{(3)}$ ),  $\frac{3}{4}$  of the way up. Thus  $Q_1 = x_{(2)} + \frac{3}{4}(x_{(3)} - x_{(2)})$ . The same approach applies to the median; it is at position  $(n+1)/2$ .

▷ **EXAMPLE. Countries data: Sub-Saharan Africa** (MINITAB worksheet: countries.mwx)

Consider another variable in the **Countries** data file — the difference in life expectancy, calculated over 20 years: (2000 to 2005) minus (1985 to 1980). In general, we might expect this difference to be a positive number as health and well-being improves over time. Here we investigate measures of location for the difference in life expectancy in Sub-Saharan Africa; the data are shown in a dotplot in Figure 8.

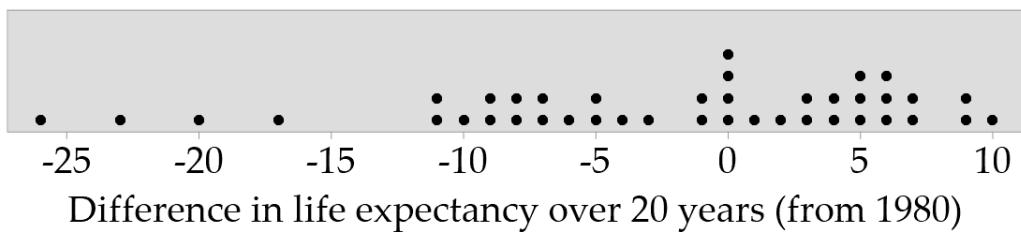
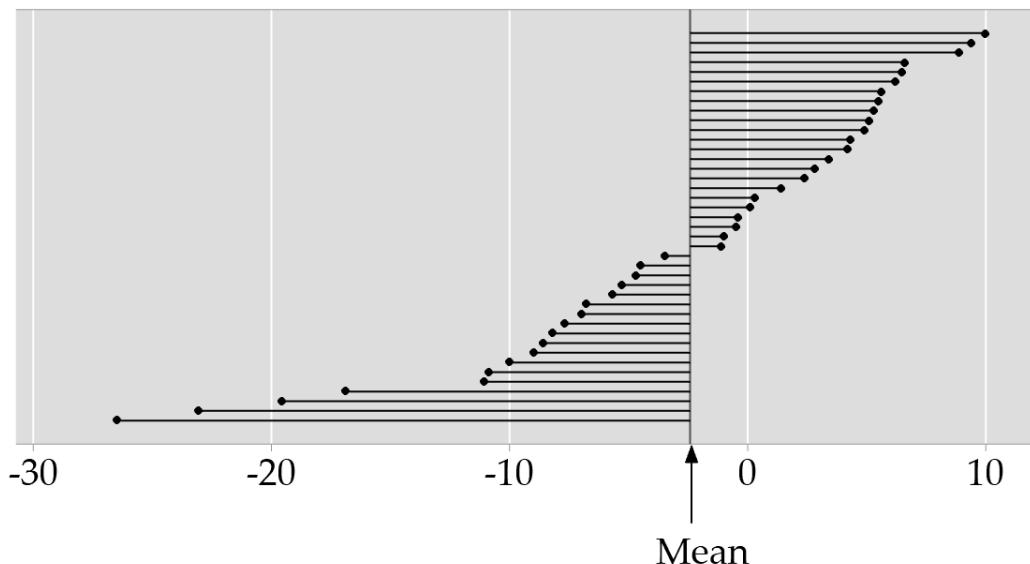


Figure 8: Dotplot of life expectancy difference in 41 Sub-Saharan countries

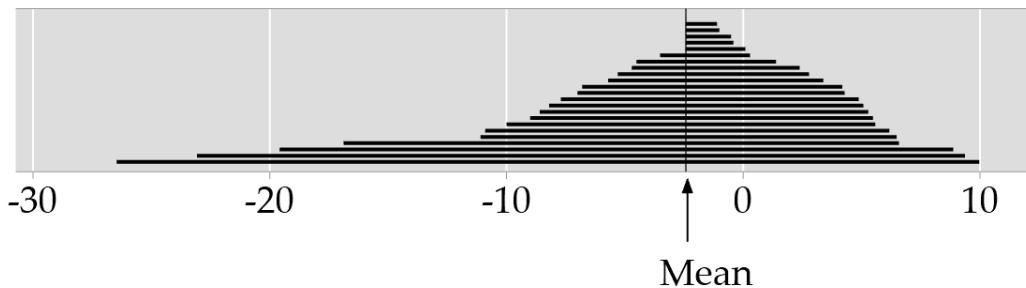
There are 41 Sub-Saharan African countries, and the mean  $\bar{x} = \frac{-99.2}{41} = -2.42$  years.

One way of thinking about the mean is as the balance point in the distribution. To achieve balance, the total distance of the data points to the left of the mean must be the same as the total distance of the data points to the right of the mean. This is illustrated in Figures 9 and 10 for the difference in life expectancy in the 41 Sub-Saharan countries. Figure 9 shows the distances of each value in the data from the mean; the distances from largest to smallest. Figure 10 shows the distances stacked, reflecting the concept of balance. This way of thinking about the mean can be a useful way of estimating the mean from a visualisation of the distribution.



Difference in life expectancy over 20 years (from 1980)

Figure 9: *Dotplot of life expectancy difference in 41 Sub-Saharan countries showing distances from the mean*



Difference in life expectancy over 20 years (from 1980)

Figure 10: *Life expectancy difference in 41 Sub-Saharan countries illustrating the mean as a balance point of the total distance*

Given that there are 41 Sub-Saharan countries, the middle value will be the 21st observation in the ordered data. There will be 20 observations below and 20 observations above this middle value. Figure 11 is a jittered individual value plot of the differences in life expectancy in the 41 Sub-Saharan countries; the number shown adjacent to each point is the rank in the ordered data. In Figure 11 the 21st observation is just below zero; the value of this point, the value of the median is  $-0.5$ . We think of the median as the ‘cut point’ in the distribution that creates two groups of equal size.



Figure 11: *Life expectancy difference in 41 Sub-Saharan countries illustrating the median as a cut point*

Figure 12 illustrates an extension of this idea to the upper and lower quartiles. The first quartile  $Q_1 = -7.95$ , and the third quartile  $Q_3 = 5.00$ .

(In Figures 11 and 12 a considerable amount of vertical jittering has been used, just to allow the identification of individual points. The only meaningful variation is on the horizontal scale.)

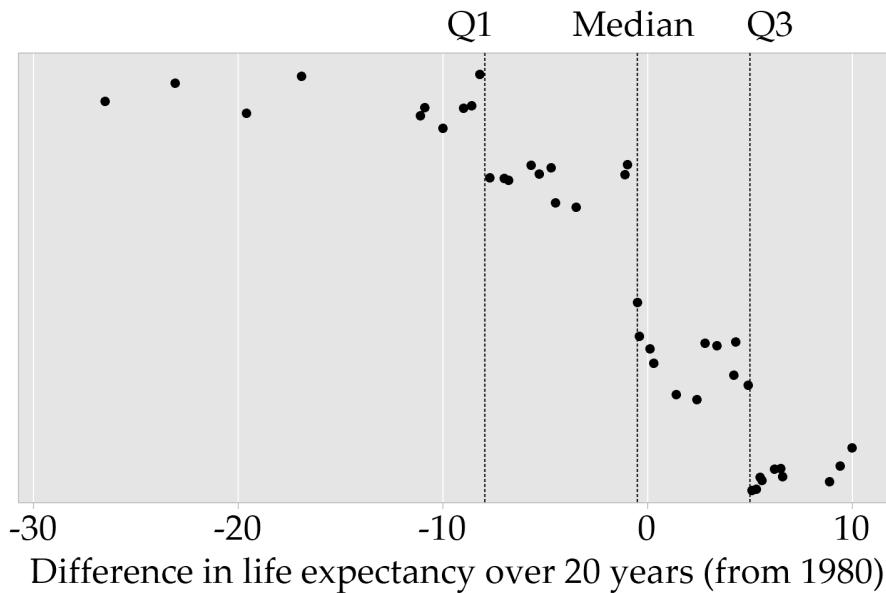


Figure 12: Life expectancy difference in 41 Sub-Saharan countries illustrating the quartiles as cut points

The choice of measure depends upon the type of data being considered and the use to be made of the measure. For measures of central tendency, as a rough guide, the mean is usually preferred with symmetric data and the median with skewed data. This is why you often see the median used for house prices; the distribution of house prices is skewed to the right (the very expensive mansions) and the mean would be a bit to the right of the median.

However, this choice needs to be made carefully. If you want to get a good estimate of the mean of a population from a sample, then the median may not be useful, especially if the data are skewed. And you might want to estimate the mean rather than the median of the population; for example, if you want to scale up the mean to estimate a total.

### 2.3.2 Measures of spread

These measures are also referred to as measures of scale. The most commonly used are:

- **standard deviation**,  $s$ ; the square-root of the variance (defined next). The standard deviation (and not the variance) is in the same units as the original values. For many samples, approximately 95% of the sample will be within 2 standard deviations of the mean. From a practical point of view, the standard deviation is the most important measure of spread.
- **variance**,  $s^2$ ; roughly, the mean of the squares of the differences between the data values and  $\bar{x}$ .

$$\begin{aligned}s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} \\ &= \frac{\sum(x_i - \bar{x})^2}{n - 1}\end{aligned}$$

You might wonder why  $(n - 1)$  is used and not  $n$ . We will discuss this in more detail later; for now, it may be useful to observe that when  $n = 1$  and we have just one observation, the sample variance using  $n - 1$  is the undefined value  $\frac{0}{0}$ , which makes more sense than the sample variance using  $n$  in the denominator, which is  $\frac{0}{1} = 0$ .

Variance, as a measure of spread, may seem somewhat strange—and it is! For one thing, it is not even in the same units as the data. It is important to think about the variance, generally, as the route to working out another measure of spread, namely, the standard deviation.

Why variances are used at all will become clearer when we look at some theory in Chapter 3.

- **range**; probably the simplest measure of spread, generally used only for small samples as it is too much affected by ‘outliers’. This is also the easiest understood measure.
- **interquartile range (IQR)**; the range of the middle 50% of the data.  $IQR = Q_3 - Q_1$ , where  $Q_1$  and  $Q_3$  are the lower and upper quartiles. Note that both the range and the interquartile range are numbers, not intervals. They reflect the lengths of defined intervals.

▷ **EXAMPLE. Countries data: Sub-Saharan Africa** (MINITAB worksheet: countries.mwx)

We return to the difference in life expectancy, calculated over 20 years, in Sub-Saharan African countries. The range can be estimated from the dotplot in Figure 8; it is 36.5 years.

The quartiles were  $Q_1 = -7.95$  and  $Q_3 = 5.00$ , so the interquartile range  $IQR = 12.95$ . This is reported when the median is used as the measure of location.

The variance  $s^2 = 77.4$ , and so the standard deviation  $s = 8.8$ . Common practice would be to report the mean  $\bar{x} = -2.42$  years with  $s = 8.8$ .

Consider applying the rule of thumb that approximately 95% of the sample will be within 2 standard deviations of the mean to these data;  $2s = 17.6$ , so:

- $\bar{x} - 2s = -2.4 - 17.6 = -19.8$
- $\bar{x} + 2s = -2.4 + 17.6 = 15.2$

Here  $n = 41$ ;  $41 \times 0.95 = 38.95 \approx 39$ . Hence we expect most of the data (39 of the 41 observations) to fall within the range  $-19.8$  to  $15.2$ . You can check this on Figures 8 or 9. As it happens, for this dataset the rule works very well; obviously, all of the data are less than  $15.2$ , and on the negative side there is an observation at  $-19.6$  (just greater than  $-19.8$ ) and two observations less than  $-19.8$ . Hence 95% (39/41) of the observations do fall within two standard deviations of the mean. The rule is an approximation — it won't always be 95% — but is surprisingly robust.

The choice of measure depends upon the type of data being considered and the use to be made of the measure. The standard deviation is by far the most commonly used measure though the interquartile range is preferable for some purposes, especially if the data are skewed or contain outliers.

### 2.3.3 Other summary measures

Summary measures of other characteristics of a data set, such as skewness (asymmetry) and kurtosis (thickness of tails), are available but are seldom used. Skewness and kurtosis are sometimes used to assess normality, but there are better ways to do this, which are discussed later.

### 2.3.4 Basic descriptives in MINITAB

There are two main ways to get basic descriptive summaries from MINITAB.

1. The simplest is obtained from Stat → Basic Statistics → Display Descriptive Statistics. For the subset of data for Sub-Saharan Africa, and the column containing the difference in life expectancy, we get:

#### Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Difference in life expectancy	41	0	-2.42	1.37	8.80	-26.50	-7.95	-0.50	5.00
Variable	Maximum								
Difference in life expectancy	10.00								

We have now seen all of these except for SE Mean, which we will deal with later. There are a large number of descriptive statistics that may be obtained, by varying what is displayed from this command; click on Statistics.

This basic description of a variable can also be obtained for each level of another categorical variable, by checking the By Variable box in the dialogue box, and selecting the relevant categorical variable. For example, we could obtain the summary statistics by region, using the entire data set.

2. The second way to get descriptive statistics on a variable gives you a lot more information, including some visual representations. Some of

this information needs careful interpretation in some cases; in fact, it is not all merely descriptive. It is obtained via Stat → Basic Statistics → Graphical Summary.

### 2.3.5 Summarising relationships: correlation

The most common descriptive measure corresponding to a scatterplot is  $r$ , the correlation between the two variables, sometimes known as Pearson's correlation coefficient. This is a measure of the strength of the *linear association* between two variables.

Obviously, examining a scatterplot is important when considering the possible relationship between two variables.

▷ **EXAMPLE. Countries data** (MINITAB worksheet: countries.mwx)

Figure 13 shows three scatterplots using variables from the **Countries** data set, the first of which we have seen in Figure 4. In the top left, the relationship is negative, in the top right it is positive, and there is little evidence of a systematic relationship in the bottom panel.

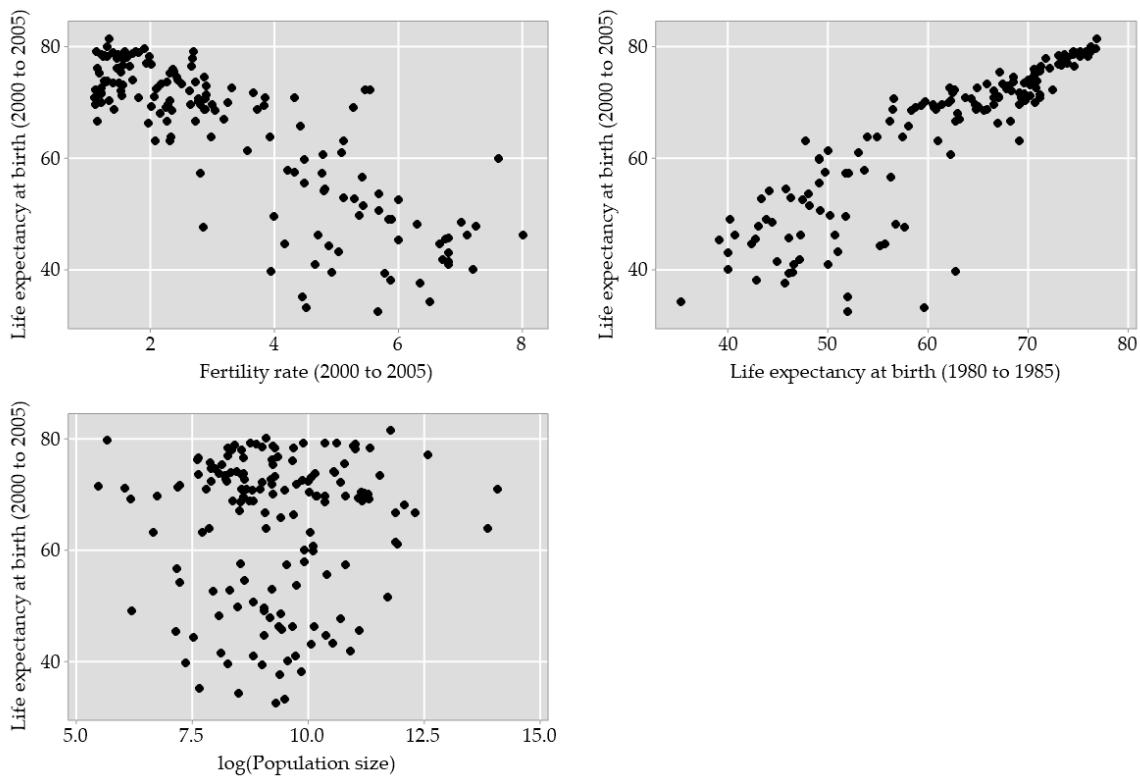


Figure 13: Three scatterplots for variables from the Countries data set

Pearson's correlation falls in the range:  $-1 \leq r \leq 1$ . A value of  $r$  close to

$\pm 1$  indicates a strong (linear) relationship and values close to 0 indicates a weak (linear) relationship. The association is positive (one variable tends to be large as the other is large) if  $r > 0$ , and negative (one variable tends to be large as the other is small) if  $r < 0$ .

In Figure 13, the correlations are:

- $r = -0.83$  for life expectancy (2000 to 2005) and fertility rate (2000 to 2005);
- $r = 0.88$  for life expectancy (2000 to 2005) and life expectancy (1980 to 1985);
- $r = 0.04$  for life expectancy (2000 to 2005) and log(population size).

$r = -1$  if and only if all of the data points lie exactly on a line with negative slope, and  $r = +1$  if and only if all of the data points lie exactly on a line with positive slope.

Note that the correlation does not depend at all on the units of  $x$  and  $y$ . It is “unit-free”. If  $x$  is “time”, for example, then the same correlation would be obtained, regardless of whether we used minutes, seconds or hours. This can be understood by considering the way the correlation is defined.

For data  $x_1, x_2, \dots, x_n$  with sample mean  $\bar{x}$  and sample standard deviation  $s$ , the standardised values are defined to be  $\frac{x_i - \bar{x}}{s}$ ; we subtract off the mean and divide by the standard deviation. The standardised values have a mean of zero and a variance (and standard deviation) of 1, and are unit-free.

Suppose we have  $n$  pairs of observations  $(x_1, y_1), \dots, (x_n, y_n)$ , and we define  $s_x$  to be the sample standard deviation of the  $x$ s, and  $s_y$  to be the sample standard deviation of the  $y$ s. Then the sample correlation coefficient is given by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right).$$

The formula for the correlation coefficient  $r$  involves summing the product of standardized values; hence  $r$  itself is also unit-free.

If the two variables tend to be either both large or both small, they are considered to be positively correlated. For a given pair  $(x_i, y_i)$ , if  $x_i$  and  $y_i$  are both large the two standardized scores will be positive, and so will their product; if they are both small, the two standardized scores will be negative, and their product will be positive. This will make the correlation,  $r$ , positive. This can be seen, for example, in the top right panel in Figure 14. The three pairs of variables shown in the scatterplots in Figure 13 were standardised to produce the scatterplots in Figure 14; note that patterns are unchanged — only the scale has changed.

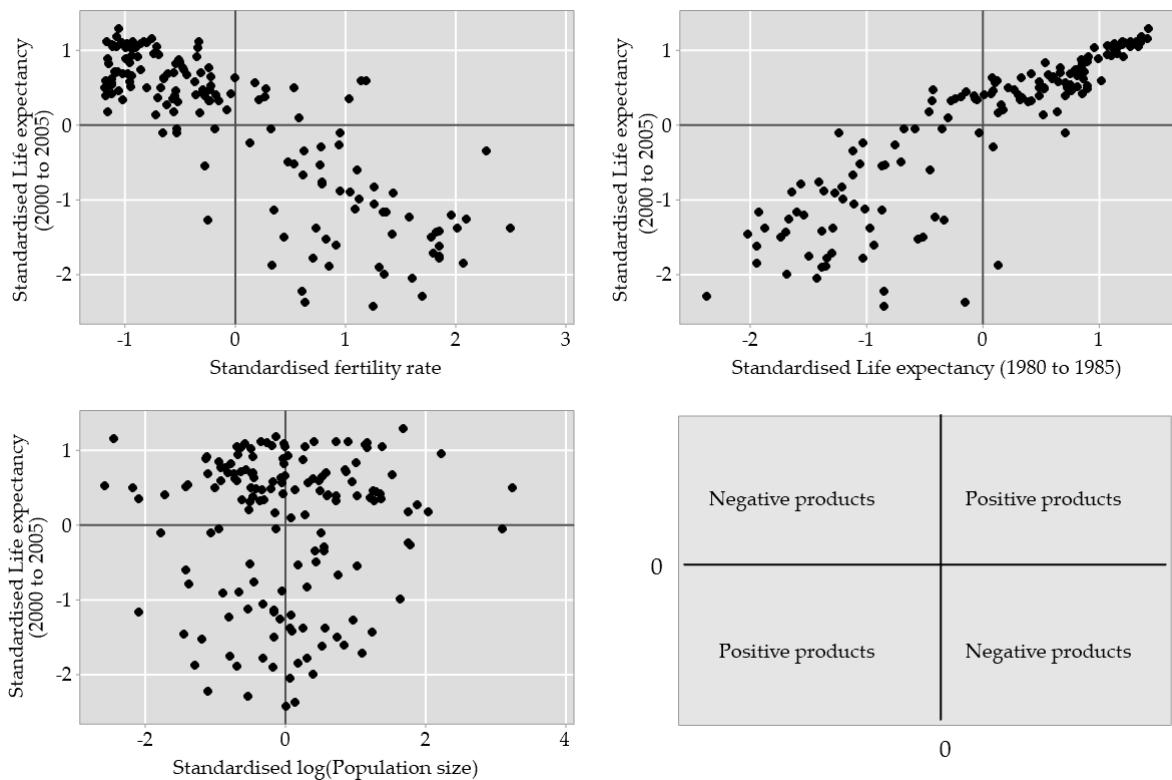


Figure 14: Three scatterplots for standardized variables from the Countries data set; black lines shown at the mean, 0, of the standardized scores

In MINITAB: Stat → Basic statistics → Correlation.

**Correlation measures linearity** The magnitude of the correlation coefficient tells us about the strength of the linear association. However, a very small correlation does not mean that two variables are not associated. For example, there may be a curved relationship between the variables in which case Pearson's  $r$  may give a misleading picture about the association as it is only a measure of **linear** association. Examining the scatterplot is therefore essential.

▷ **EXAMPLE. Commission on lunacy, 1854**

In 1854, US physician and psychiatrist, Edward Jarvis lead the 'Commission on lunacy' which aimed to determine accommodation needs and management plan for people with mental illness. A survey was carried out in Massachusetts which aimed to collect information about individuals with mental illness. 1,702 copies of a "lithographic" letter with survey of 15 questions were sent to physicians; the response rate was nearly 100%: 1,315 of the 1,319 eligible physicians surveyed returned their survey forms. Others working in mental health provided corroborating evidence, and survey enumerators took care to avoid double counting and to verify identities and data.

Figure 15 shows the data collated for the 14 counties in Massachusetts. For each county the average distance to the nearest mental health care facility was measured (in miles), and the percentage of people with mental illness who were cared for at home was determined. The correlation between these two variables is  $r = 0.41$ . Does this fully reflect the relationship in Figure 15? Is the pattern strictly linear?

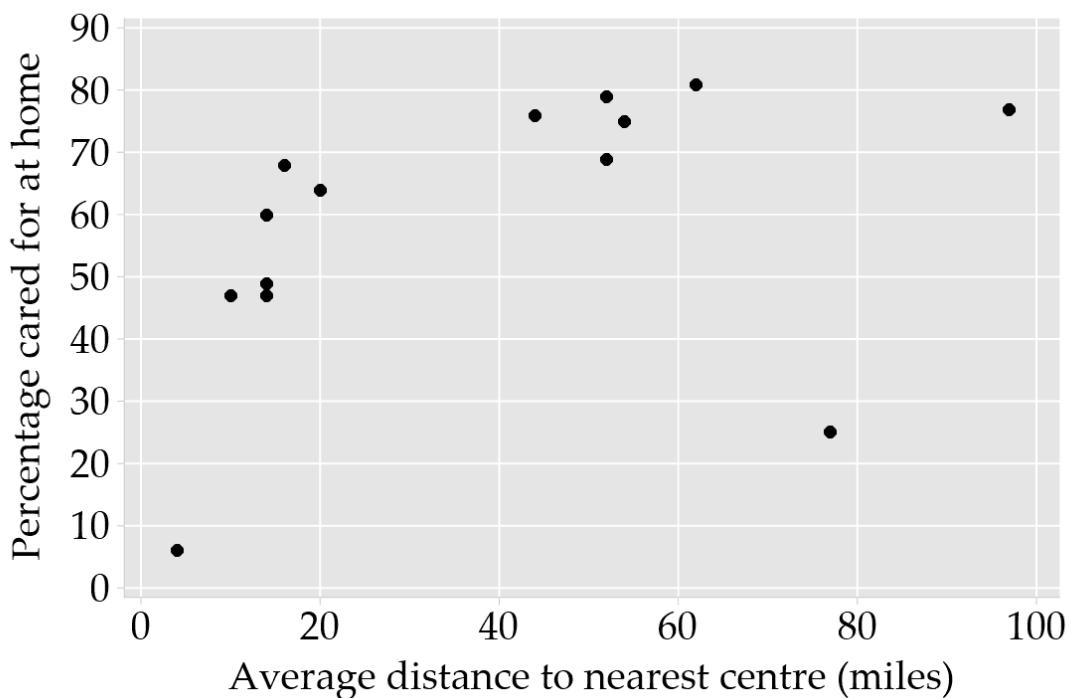


Figure 15: Scatterplot showing the relationship between average distance to the nearest mental health care facility and percentage cared for at home, in 14 counties in Massachusetts in 1854

**Correlation and causation** A large correlation reflects an association between two variables, it does *not* prove causation. One possibility is that the correlation between two variables may be a consequence of the association between both variables and some other variable (or variables). For example there is a strong (positive) correlation between the number of churches and the number of pubs in country towns. The reason is that both of them are related to population size.

This is an obvious example, but researchers, nevertheless, are often tempted to make the leap from correlation to causation. An example that illustrates the subtleties involved is the negative correlation between blood lead levels in children and IQ. One way that this can be explained is that the children who ingested more lead when babies, suffered adverse consequences in terms of intelligence (lead exposure caused IQ deficit). Another explanation is that babies of lower intelligence are less discriminating, and therefore

more likely to ingest lead-carrying substances (lower IQ caused higher lead exposure). And these are not the only possible explanations. Which explanation is correct? The existence of the correlation, *per se*, cannot answer this question.

▷ **EXAMPLE. Chocolate and Nobel prizes** Figure 16 plots data from Wikipedia on the number of Nobel laureates per 10 million people by country in 2017 and data from Statistica on the per capita chocolate consumption (in kilograms). Does eating more chocolate produce more Nobel prizes? What might explain the pattern in the data?

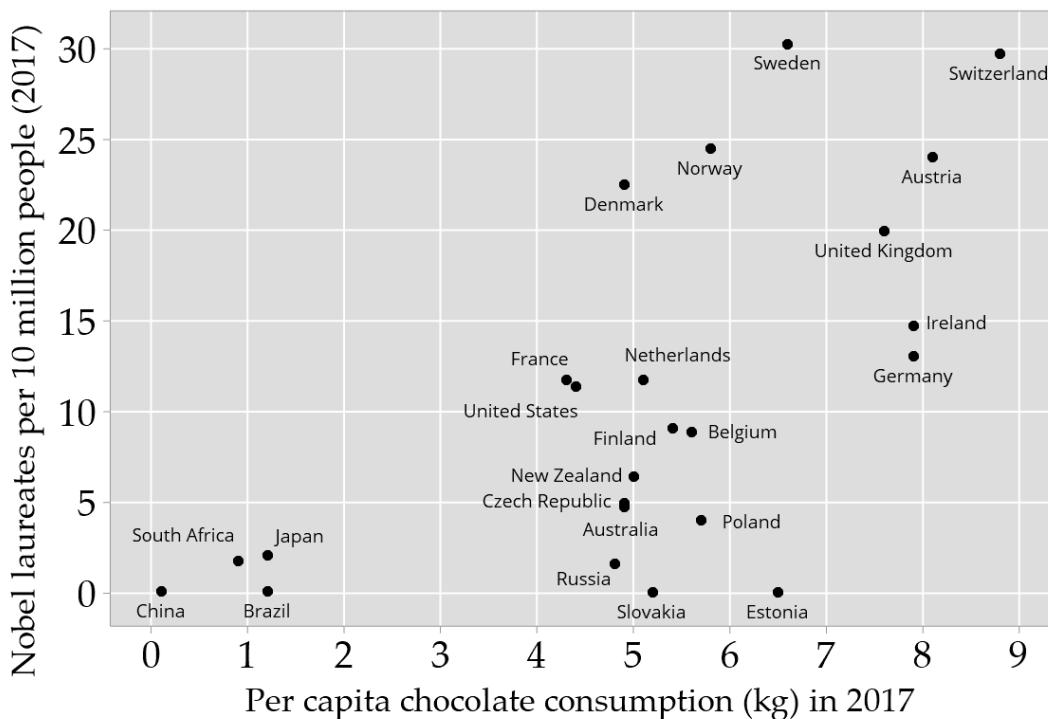


Figure 16: Scatterplot showing the relationship between chocolate consumption and Nobel prizes in 24 countries in 2017

If two variables reflect phenomena that are causally related, there will often be a strong correlation between the variables. This may tempt us to make the inference in the reverse direction. Study designs that are “observational”, in which no experimental intervention is used, and there is no randomisation, essentially entail the examination of associations between variables. The analysis amounts to a more elaborate version of examining correlations, and for that reason, traditionally, such studies have been seen as limited in their value as scientific evidence.

Over time, many disciplines have thought more carefully about what we can learn from observational studies, and there is a body of theory and modelling that attempts to address this issue in a principled way. A classic

reference on establishing causality in observational studies is: Hill, Austin Bradford (1965). The Environment and Disease: Association or Causation?. *Proceedings of the Royal Society of Medicine*. 58 (5): 295–300.

**Assessing agreement** Correlation is sometimes used to assess the extent of agreement between two alternative methods for measuring the same quantity. This is an inappropriate use of correlation. In fact, it is possible to have a correlation of exactly 1 between the two methods, and still have poor agreement.

▷ **EXAMPLE. Blood pressure measurements**

Figure 17 is a scatterplot of blood pressure (mmHg) from 200 patients, measured in two different ways: arm systolic pressure and finger systolic pressure (mmHg). The data are reported on in Bland & Altman (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. *The Lancet*, 1085-1087.

The correlation between these two measurements is  $r = 0.83$ . In this example, there is good agreement between the methods of measurement: the mean difference is small (around 4mmHg). However the correlation is not good evidence of this agreement. Consider if the finger systolic pressure measurements had all been systematically lower by 30 mmHg, as shown in Figure 18; what do you estimate the correlation to be here? How good is the agreement?

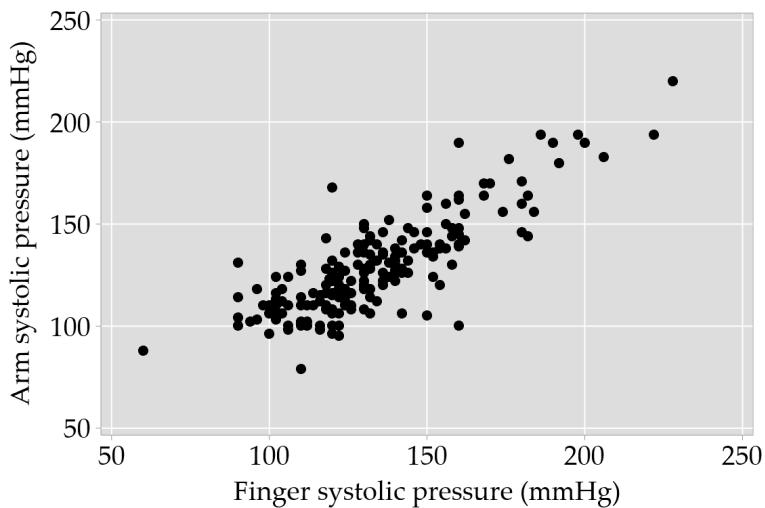


Figure 17: Scatterplot of two methods of measuring blood pressure

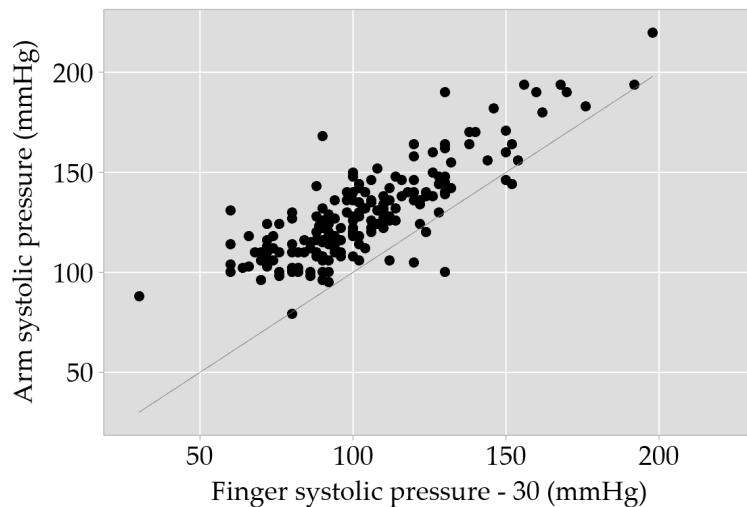


Figure 18: Scatterplot of two methods of measuring blood pressure, with adjustment; the line of perfect agreement is shown

We do not cover appropriate methods of assessing agreement, but a useful starting reference is: Bland & Altman (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 307-310.

**Correlations in sub-samples** The context of a correlation needs careful thought. Sometimes a correlation is calculated for a set of data made up of a number of sub-samples. The correlation in the sub-samples may be quite different from the overall correlation.

▷ **EXAMPLE. Fisher's iris data**

The famous statistician R.A. Fisher reported measurements made on 150 iris petals in a paper in 1936: “The use of multiple measurements in taxonomic problems” (*Annals of Eugenics*, 7 (2): 179–188.) Figure 19 plots the data. What do you estimate the correlation to be in each of the panels? What are the implications of this?

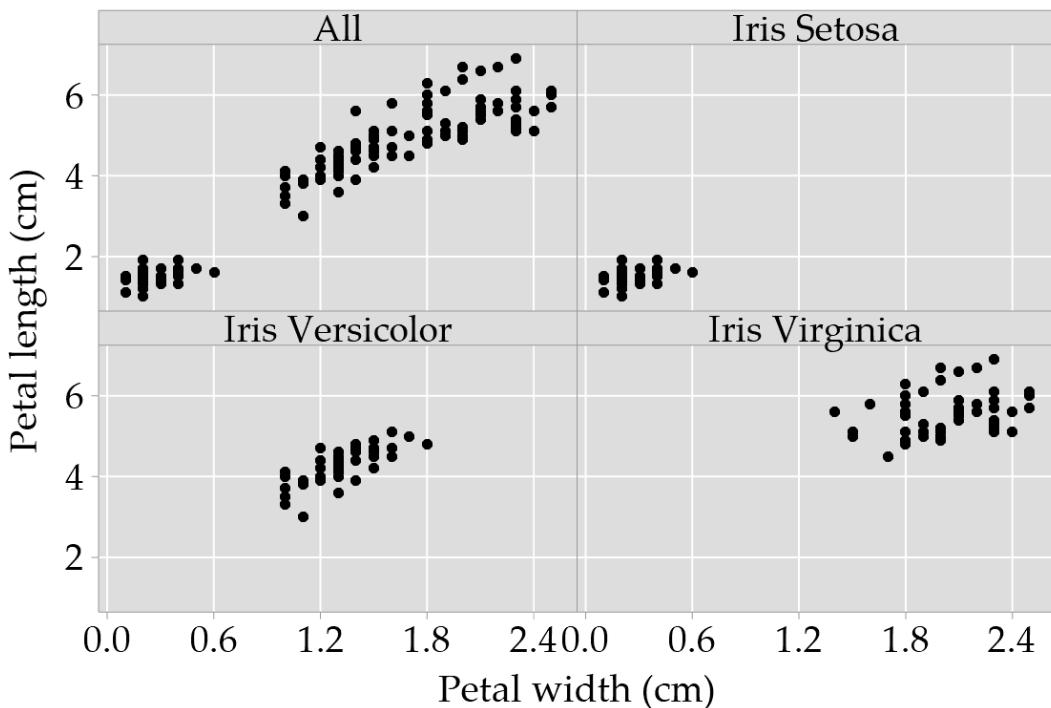


Figure 19: Scatterplots showing the relationship between petal length and petal width

## 2.4 Visualising summaries of numerical data: boxplots

A boxplot is a visual summary of a data set based on the median, quartiles<sup>5</sup> and extreme values, or “outliers”. Boxplots are especially useful for comparing data sets from two or more populations.

▷ **EXAMPLE. Countries data: Difference in Life expectancy** (MINITAB worksheet: countries.mwx)

We consider the variable ‘Difference in life expectancy’ again; this is the Life expectancy in 2000-2005 minus the Life expectancy in 1980-1985, reflecting the change over that period. We have examined this variable for a subset of countries — those in Sub-Saharan Africa; now we consider the entire data set.

Figure 20 shows eight separate boxplots of differences in life expectancy by region. The rectangle in a boxplot has ends at  $Q_1$  and  $Q_3$ , and the central line is the median. The rectangle is referred to as the ‘box’. So it is possible

---

<sup>5</sup>The definition of the quartiles used for a boxplot is sometimes slightly different from the definition given above; MINITAB uses the same definition in both cases.

to read off the three quartiles from the box. The lines which extend from the sides of the box are called ‘whiskers’.

MINITAB produces a boxplot via Graph → Boxplot ...

The rules for the data points beyond the box are:

1. The whiskers are drawn from each end of the box to the furthest observation in the data that is not greater than  $1.5 \times \text{IQR}$  from the ends of the box. This means that the whiskers extend to the maximum and/or the minimum points, if these points are no further than  $1.5 \times \text{IQR}$  from the box.
2. Observations which are more than 1.5 times the interquartile range (IQR) from the box are designated by an individual symbol, and are referred to as **outliers**.

In Figure 20 there are no outliers in Sub-Saharan Africa, for example. In South America, there is one outlier. The median is at 5.1,  $Q_1 = 4.1$ ,  $Q_3 = 6.3$  and the interquartile range is 2.2.

So  $1.5 \times \text{IQR} = 3.3$ . On the lower side, the whisker goes to 2.2, which is the smallest observed value not less than 0.8 ( $= 4.1 - 3.3$ ). The maximum endpoint for the upper whisker is  $9.6 = 6.3 + 3.3$ , so the upper whisker extends to the largest value less than 9.6, which is 8.4. The observed value greater than 9.6 is plotted as an outlier; this is an improvement in life expectancy of 10 years for Bolivia.

In MINITAB, if you hover the cursor over the boxplot, a small box of descriptive statistics appears, and if you hover over an outlier, the value of the outlier is shown. This feature is a general one in MINITAB graphs.

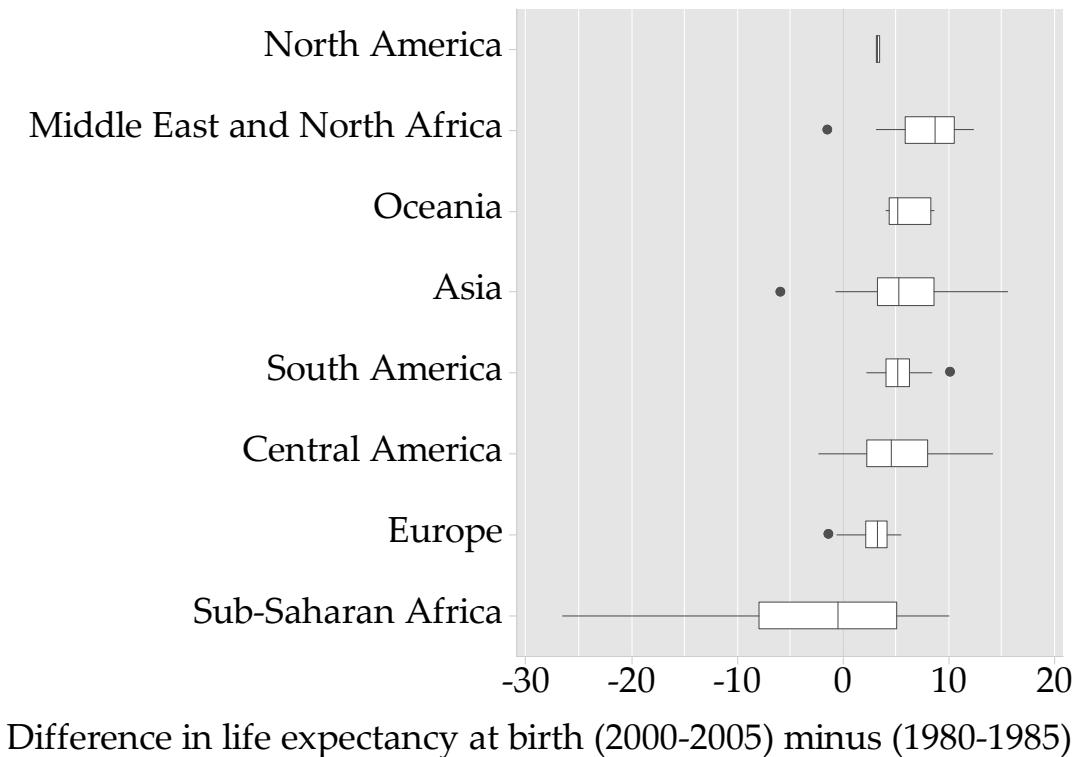


Figure 20: *Boxplots of the difference between the life expectancy at birth (2000–2005) and the life expectancy at birth (1980–1985) in 155 countries, by region*

Two important features of the boxplot are:

- It always shows the minimum, maximum and quartiles of the data, sometimes known as the “5-number summary”;
- It reflects actual points along the distribution, rather than derived and less tangible quantities such as mean  $\pm$  standard deviation. In particular, the ends of the whiskers are at an actual value observed in the data.

#### 2.4.1 “Outliers”

How should outliers be treated? The concept of an outlier depends on a model for what was expected. Something can only appear unusual when the usual pattern has been identified. One rule to follow is: do not automatically discard outliers. They may be extremely informative. To make data conform to a desired or convenient statistical model by removing outliers as part of an analytic strategy may mean losing a research insight.

Some textbooks recommend removing outliers or even “adjusting” outliers. This is a bad idea. In a telling example of why not to do this, Antarctic satellite collection data systems automatically deleted outliers, and as a result the hole in the ozone layer was detected much later than it could have been.

It is good practice to identify outliers and study them for the insight they may offer.

Removing data which are simply at the extremes or adjusting values at the extremes is a dubious scientific practice. Some people regard it as scientific fraud.

Outliers are most typically thought about in relation to boxplots. However, as mentioned, identification of outliers depends on the context considered. Here we introduce another useful graph to illustrate this point; this is a variation of a scatterplot, referred to as a **marginal plot**. It adds the two boxplots of the  $x$  and  $y$  variables to the margins (hence the name) of the scatterplot, using the scales of the scatterplot. This is an informative representation of bivariate data. (In some software it is also possible to add histograms or dotplots in the margins, but boxplots are generally most efficient.)

▷ **EXAMPLE. Countries data: Sub-Saharan Africa**

Figure 21 shows a scatterplot of difference in life expectancy over 20 years versus percentages of adults living with HIV in Sub-Saharan countries; boxplots are added in the margins. In the boxplot of the percentage of adults with HIV, a number of outliers are shown. However, when considering the bivariate relationship in the scatterplot — between difference in life expectancy and percentage of adults with HIV — we see that all data points are essentially consistent with a negative linear relationship. Here we see univariate outliers (in one variable), but little evidence of bivariate outliers (in the scatterplot).

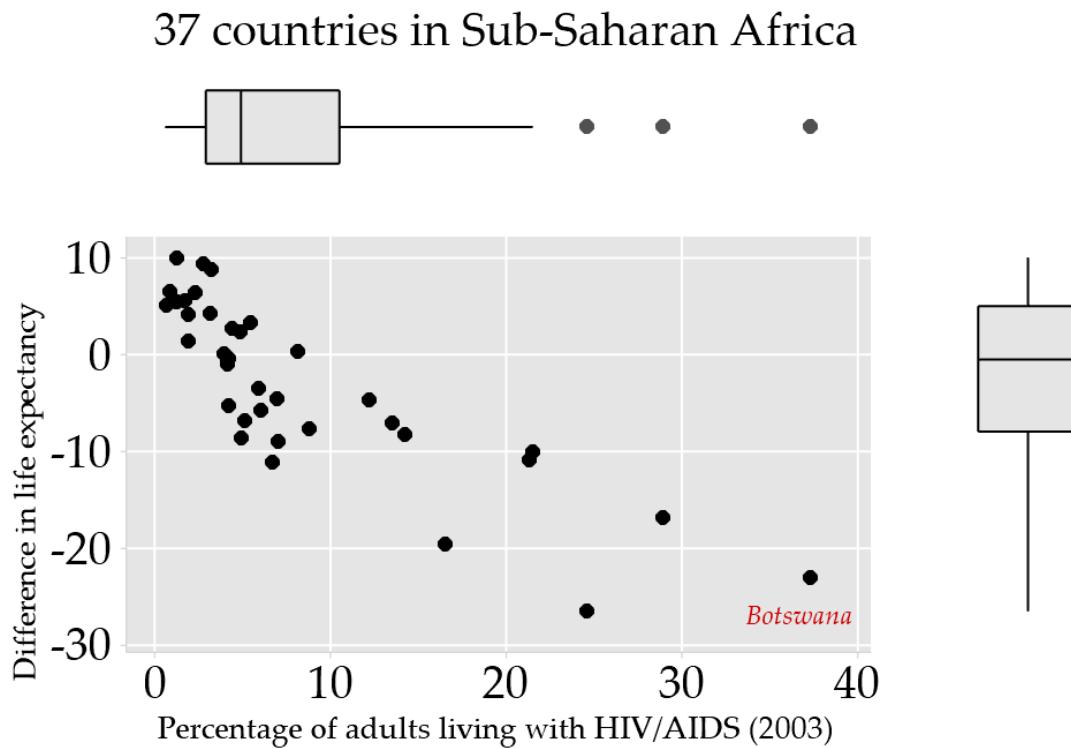


Figure 21: Marginal plot of difference in life expectancy over 20 years versus Percentages of adults living with HIV in Sub-Saharan countries

▷ **EXAMPLE. Commission on lunacy, 1854**

Figure 22 is a marginal plot of the data collected by Jarvis, introduced earlier; note here that there are no univariate outliers. However if we consider fitting a model to these data, in this case a non-linear curve, it is clear that one of the data points is inconsistent with this model — Nantucket; it is a outlier in this context.

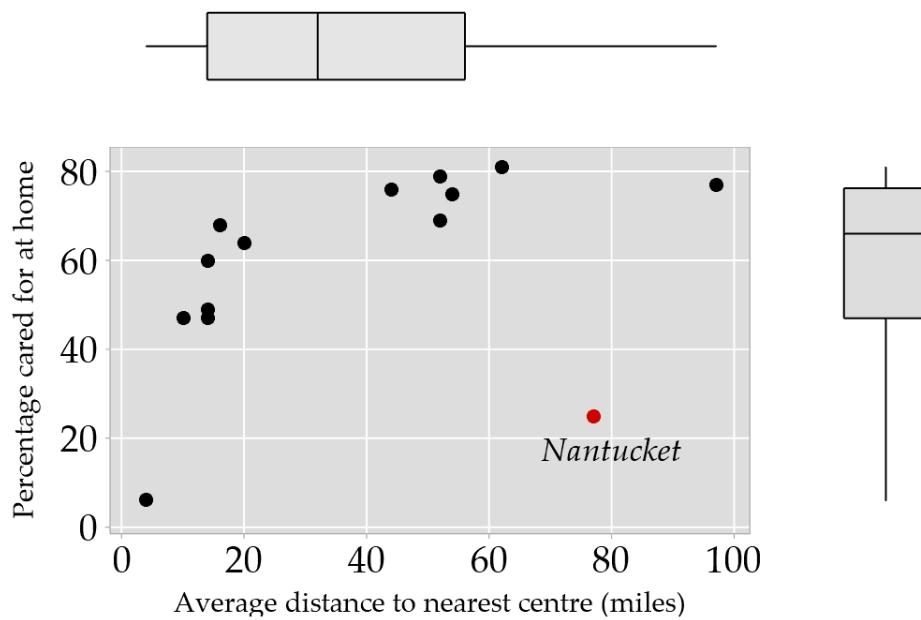


Figure 22: Marginal plot showing the relationship between average distance to the nearest mental health care facility and percentage cared for at home, in 14 counties in Massachusetts in 1854

## 2.5 Visualising numerical data: making comparisons

Figure 20, based on the **Countries** data set shows boxplots of the difference in life expectancy over 20 years according to region of the world. Constructing a graph of a numerical variable by groups, that is, broken down by a categorical variable, is a very common analytic practice that serves the goal of making comparisons between groups. Note that we see a limitation of using boxplots in this example — the size of the groups (the number of countries per region) varies substantially, and is limited for some of the very small regions.

Figure 23 and Figure 24 provide alternative visualisations that support the goal of exploring data to make comparisons.

Figure 23 is an individual value plots with the  $x$  values are randomly “jittered” in a vertical direction, by a small amount, to make them more distinguishable. Provided the degree of jittering is much less than the distance between the categories, there is no risk of confusion. The jittering in individual value plots means that they can be used for larger datasets.

Figure 24 shows dotplots by region; in this case, the dotplot is the best of the three representations we consider.

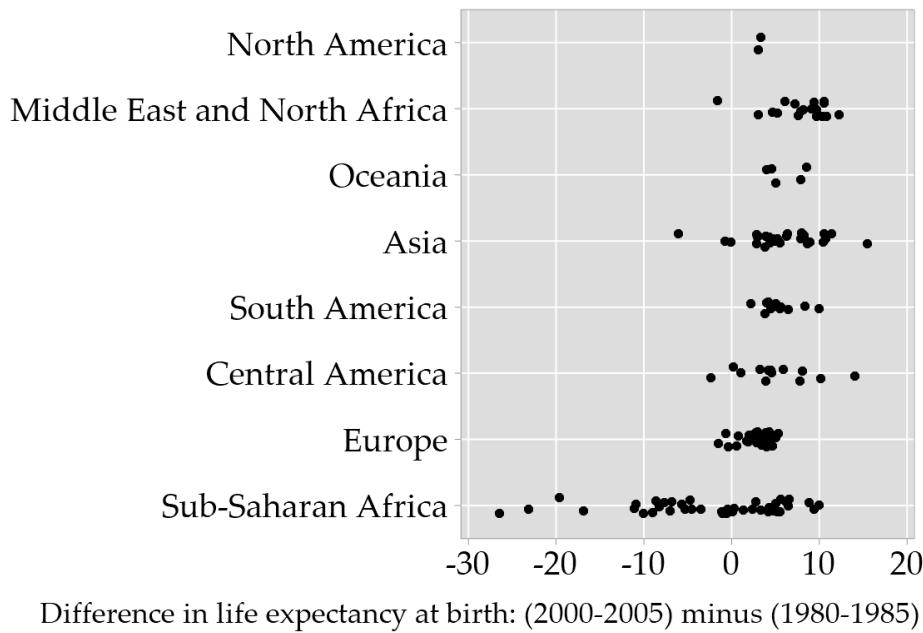


Figure 23: Individual value plot of the difference between the life expectancy at birth (2000–2005) and the life expectancy at birth (1980–1985) in 155 countries, by region

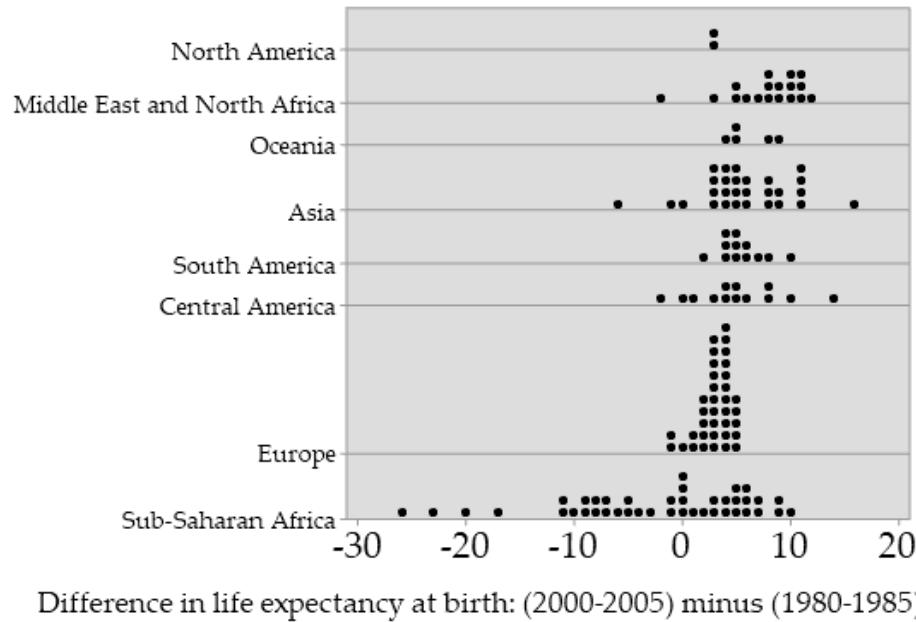


Figure 24: Dotplot of the difference between the life expectancy at birth (2000–2005) and the life expectancy at birth (1980–1985) in 155 countries, by region

## 2.6 Visualising and summarising categorical data

### 2.6.1 Understanding distributions

#### Dotcharts and barcharts

The distribution of a single categorical variable is typically graphed with a bar chart. Each bar represents the frequency (or percentage) of observations in each category.

##### ▷ EXAMPLE. Countries data

Here are data describing the number of countries in each region in the Countries data.

Asia	28	Central America	13
Europe	36	Middle East and North Africa	18
North America	2	Oceania	5
South America	12	Sub-Saharan Africa	41

Figure 25 shows a barchart of the data above. There can be gaps between the bars, as they represent separate categories. The bars are horizontal in this case; this facilitates estimating the values represented and assists in making comparisons, and allows horizontal axis labels, which are easier to read. At the top is a barchart based on counts; underneath this is a barchart based on percentages.

Figure 25 shows, at the bottom, a dotchart of the same data. The percentage in each category is represented as a point, rather than a bar. The dotchart uses less ink than the barchart and hence conforms better to the principles for good graphical practice advocated here. A dotchart differs from a dotplot — a dotplot represents numerical, rather than categorical, data.

Note that a bar chart is different from a histogram. In a histogram the intervals are constructed from a numerical variable; the intervals are plotted without any gaps between them, reflecting the underlying numerical variable. Bar charts show the frequency, or percentage, of observations in different categories.

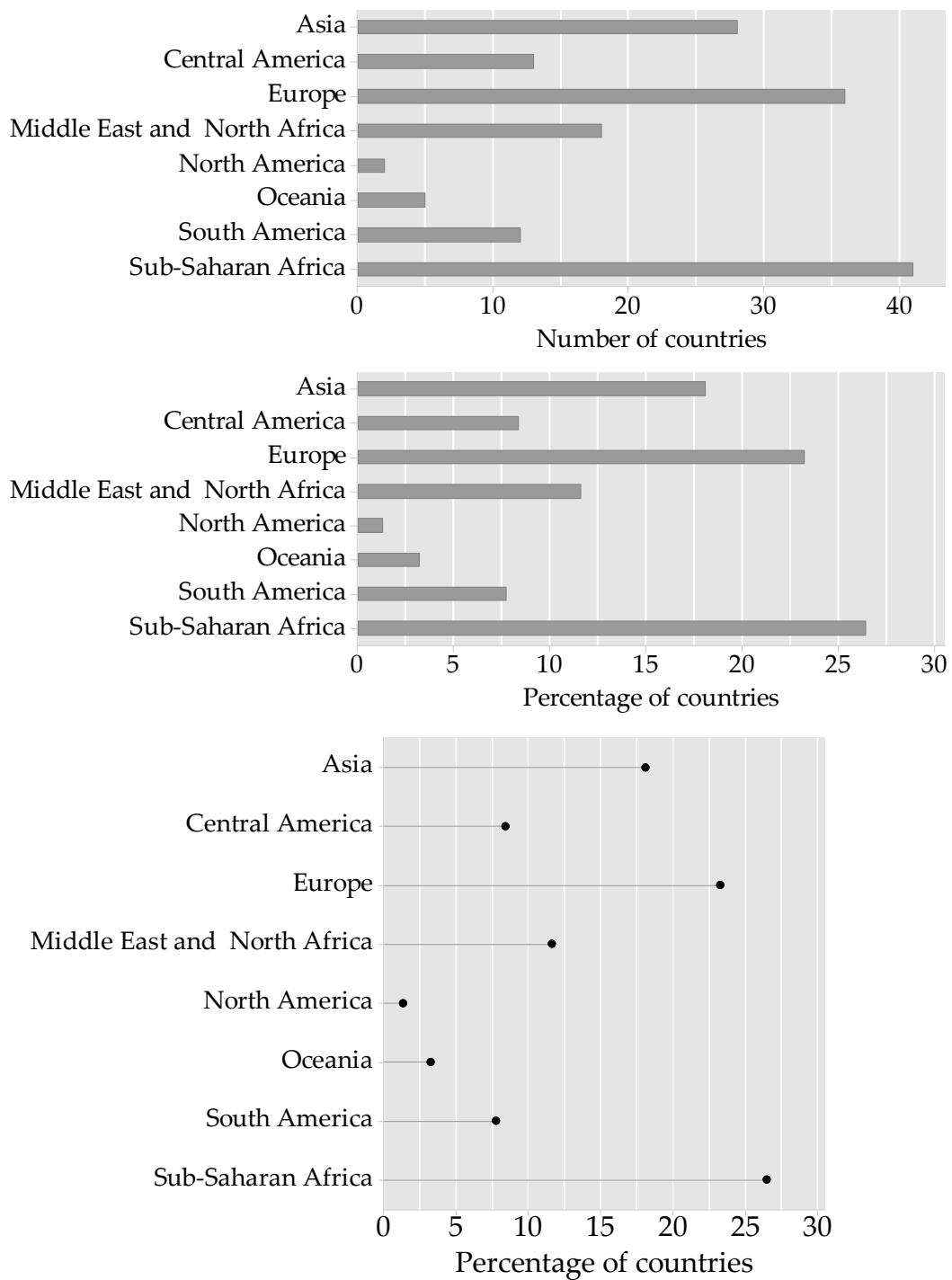


Figure 25: *Number and percentage of countries in regions of the world — Barcharts and the corresponding percentage dotchart*

## 2.6.2 Exploring relationships

Simple extensions of the barchart and dotchart can support exploration of the relationship between categorical variables. In this section, we provide a simple example.

▷ **EXAMPLE. Alligator food preferences**

Researchers captured 219 alligators in the Florida lakes in 1985 to study alligator food preferences.<sup>6</sup> The stomach contents were examined and the primary type of food eaten (in terms of volume) was classified as either: fish, invertebrate, reptile, bird or other (amphibian, mammal, plant material, stones, debris, or no food, or no dominant type of food). The explanatory variables included the location of capture (four locations), gender and size of the alligator (two categories). The outcome — type of food eaten — has five levels, and there are three explanatory variables of interest. The table representing the results has  $5 \times 4 \times 2 \times 2 = 80$  cells (5 types of food, 4 locations, 2 genders and 2 sizes). In the figures illustrated below, results are summarised across the four locations.

In Figure 26 percentages of the different food types are plotted; note that the percentages are calculated within each gender and size combination. This facilitates comparisons as the numbers of alligators within these classifications varies.

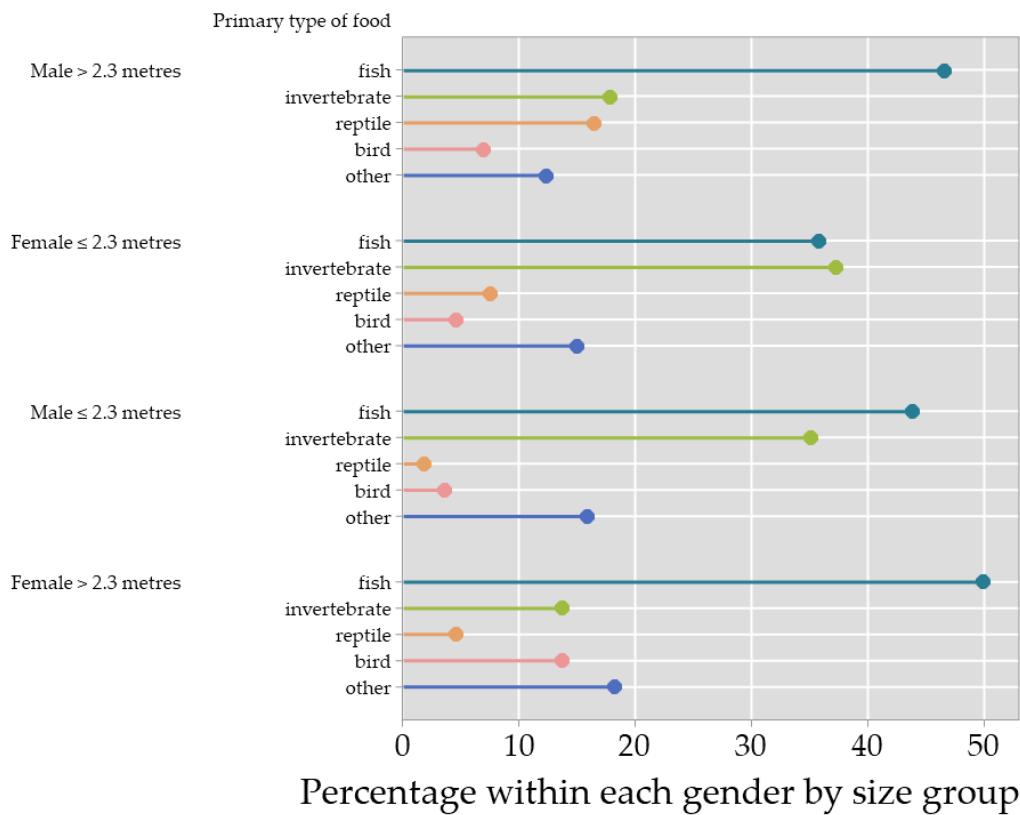


Figure 26: Dot chart of alligator food preference by gender and size

<sup>6</sup>These data come from Agresti A. (2013) Categorical data analysis (3rd edition), Hoboken: John Wiley & Sons.

## 2.7 Visualising more complex data

Multivariate data are the general case: any number of variables are considered simultaneously. A number of techniques for displaying three and higher dimensional data have been developed but many of them tend to be difficult to interpret. In this section we introduce two very useful graphs for multivariate data: panel plots and scatterplot matrices.

### 2.7.1 Panel graphs

Panel graphs are a useful general form of graph, applicable to many particular types. In a panel graph, a basic form of graph, such as a scatterplot, dotplot or an individual value plot, is shown for the values of a categorical (grouping) variable, or the combinations of two categorical (grouping) variables. A simple example of a panel plot is shown in Figure 27 where panels are defined by combinations of gender and size; this is the data we have just seen in Figure 26, used now to illustrate the general principle of constructing a panel plot.

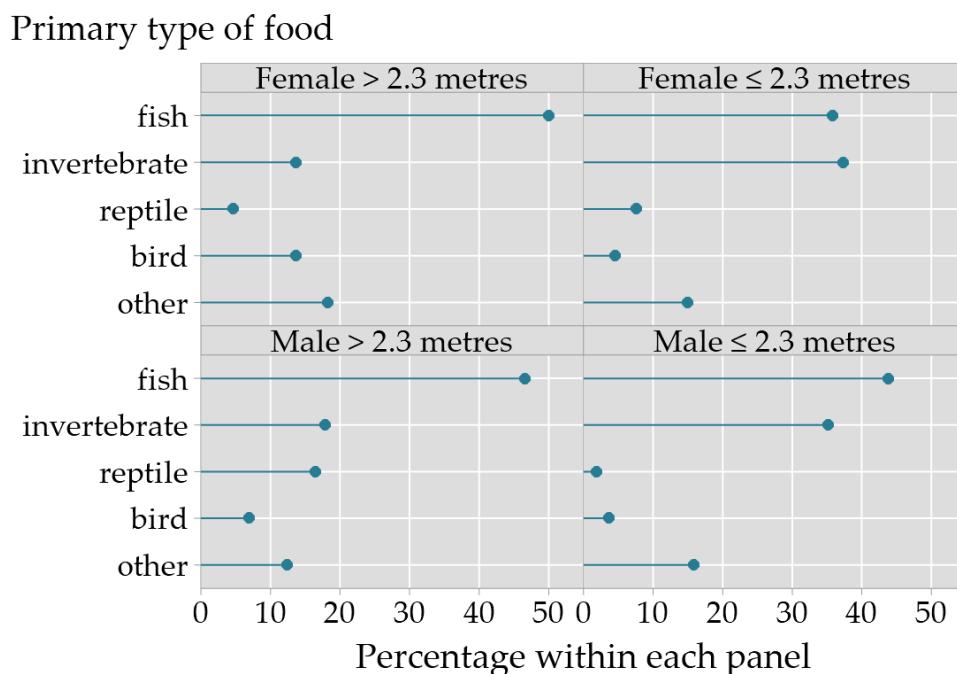


Figure 27: Dot chart of alligator food preference with panels defined by gender and size

Panel graphs support exploration of more complex data, as the next example illustrates.

▷ **EXAMPLE. Countries data** (MINITAB worksheet: countries.mwx)

Figure 4 is a scatterplot showing the relationship between fertility rate in 2000-2005 and life expectancy at birth (2000-2005) for all countries in the **Countries** dataset. In Figure 28 the same relationship is shown for the different regions in the data set: each region is in a separate panel. Note the richness of this plot. One sees the basic form of the relationship for region; it appears to be stronger in some regions than in others. The number of countries per region can also be seen, approximately, and hence the fact that there are so few observations from North America.

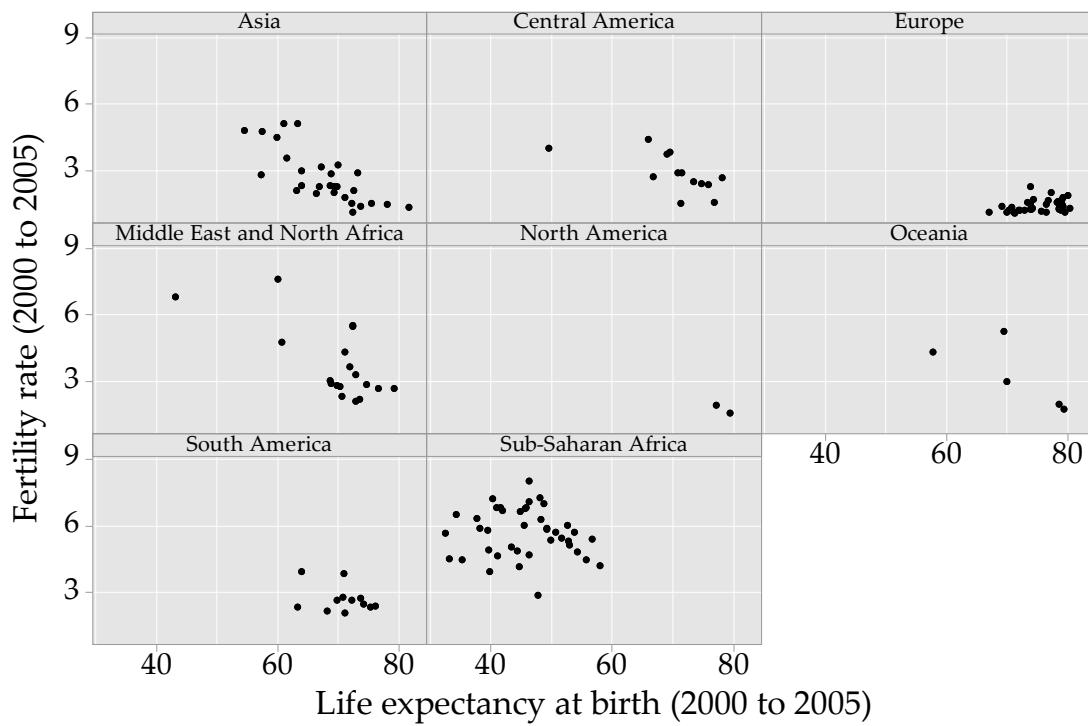


Figure 28: *Fertility rate versus life expectancy at birth (2000–2005) by region, 155 countries*

Panels can be defined by one, two or more categorical variables.

### 2.7.2 Scatterplot matrices

When three or more numerical variables are of interest, a graph which does not capture the full multivariate nature of the data but has nevertheless proven useful in applications, is the **scatterplot matrix**.

An example is shown in Figure 29, using three variables for the Sub-Saharan countries in the countries data. It consists of each variable plotted against each other variable, and can be supplemented (as in Figure 29) by the addition of scatterplot smoothers.

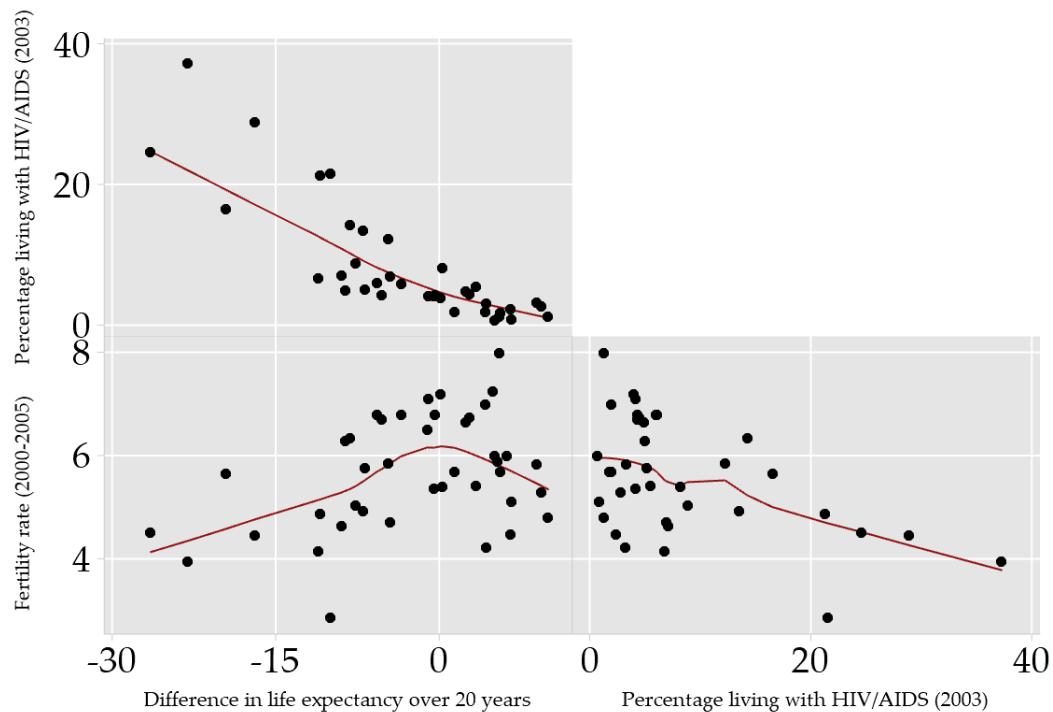


Figure 29: Scatter-plot matrix: percentage living with HIV/AIDS, fertility rate, and life expectancy difference over 20 years, with scatterplot smoothers shown

## 2.8 Exercises

- 2.1 A monitoring process at a lead smelter is used to measure the daily emission levels from the plant into the atmosphere. The data in the data file `lead_smelter.mwx` represent daily airborne lead emission amounts, in kg, for a year. The plant operates continuously.<sup>7</sup> The daily emissions are in the column labelled `emission`.
- Obtain a histogram and a dotplot of these data, across the whole year.

[ To produce a histogram, use `Graph > Histogram > Simple`; select `emission` by clicking `emission` followed by `Select`; click `OK`.  
To produce a dotplot, use `Graph > Dotplot > (One Y) Simple`; select `emission` by clicking `emission` followed by `Select`; click `OK`. ]
  - Based on the shape of the distribution, which do you expect to be larger, the mean or the median? Estimate both of these values visually, and then obtain them.

[ `Stat > Basic Statistics > Display Descriptive Statistics`; select `emission` and click `OK`. ]
  - Find the standard deviation and interquartile range of the emission data, and find the percentage of the observations that lie within 2 standard deviations of the mean, i.e. within the interval  $(\bar{x} - 2s, \bar{x} + 2s)$ . How well do these data conform to the rule of thumb in the notes, used to provide a tangible interpretation of the standard deviation?

[ *Interval notation:  $(a, b)$  denotes the numbers in the interval between  $a$  and  $b$ ; another common notation for an interval is  $a-b$ , but we tend to avoid that, as it can be confused with a difference. Thus the interval  $(14.4, 19.2)$  denotes the numbers between 14.4 and 19.2.* ]

*Hint: Work out  $\bar{x}-2s$  and  $\bar{x}+2s$  using a calculator, then find the observations inside (or outside) this range by inspecting the dotplot. To identify the exact number of observations outside the range, it is useful to temporarily change the x-scale to home in on a small interval. To do this, double click on the chart, click on the x-axis and temporarily modify the scale, so you are looking only at a small section of the data.*
  - Obtain a boxplot of the data.

[ `Graph > Boxplot > (One Y) Simple`; select `emissions` and click `OK`. ]
  - Use boxplots to compare the data from the different months, and use these to describe the pattern of emissions across the year.

[ `Graph > Boxplot > (One Y) With Groups`; select `emission` for `Graph variables`; click in the `Categorical variables for grouping` box, select `months`; click `OK`. ]

<sup>7</sup>The Australian National Pollutant Inventory provides figures for annual amounts of airborne lead emitted, by facility. In 2009 the annual figures for the Nyrstar smelter in Port Pirie and for Mount Isa Mines were 39 and 210 tonnes respectively; see [www.npi.gov.au/npidata/](http://www.npi.gov.au/npidata/).

- (f) Obtain summary statistics for the emissions for January and June.  
[ Stat > Basic Statistics > Display Descriptive Statistics; select emission for Variable; then click in the By variables box, select months; click OK. If you want the graph to display the months in date order, return to the worksheet and highlight the months column. Then click Column Properties > Value order > User-specified order; select the appropriate data order, then click OK. Return to the graph and Update it. ]
- (g)\* The daily emission levels are in kilograms. What questions would you ask about the data collection process in order to better understand how the data were obtained and to judge their quality?

2.2 Someone posted the following question on a discussion list. The relevant data are in lichen.mwx.

"I have measured two variables, Fe concentration and Protein, in three lichen species: S1, S2, S3. I took 30 measurements per species. The question was: Is there any correlation between Fe and Protein ? I took the correlation coefficients and the regression lines between these two parameters for each species separately as well as the correlation coefficients and the regression lines of pooled data independent of species. The problem is that while the correlation in any species is positive the overall correlation is negative!!!!!!! In the following data, that are given as an example, the correlation coefficients are  $r_1 = .88$ ,  $r_2 = 0.79$ , and  $r_3 = 0.90$  while the overall  $r = -0.67!!!!!!$  The question of the experimenter remains: Is the correlation between these two variables positive or negative ? What is the conclusion?"

- (a) Produce a scatter plot of Fe against Protein that identifies the three species on the plot.  
[ Graphs > Scatterplot ▶ With Groups.]
- (b) Check the results asserted by the questioner:
- (a) What is the correlation overall?  
[ Stat > Basic Statistics ▶ Correlation ... Variables: Fe and Protein.]
- (b) What are the correlations within each species?  
[ Data > Split worksheet ... By variables: Species and click . Then repeat the correlation procedure.]
- (c) What is the conclusion?

2.3 As cheddar cheese matures a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the La Trobe Valley of Victoria, samples of cheese were analyzed for their chemical composition and were subjected to taste tests. The data below were obtained from one type of cheese manufacturing process; there were 30 samples of cheese. "Taste" is the outcome variable of

interest. The taste scores were obtained by combining the scores from several tasters. Taste scores could range from 0 to 60; a higher score reflects a more positive taste rating. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide and lactic acid. For acetic acid and hydrogen sulfide (natural) log transformations were taken. The data are stored as `cheese.mwx`.

- (a) Look at the scatterplot matrix. How many different correlation coefficients are there? Guess them.  
[ To obtain a scatter plot matrix use the following: Graph > Matrix Plot > Simple; select all 4 variables; click Matrix Options, select Upper right, click OK. To obtain separate graphs for each pair of explanatory variables use Graph > Scatter plot > Simple.]
- (b) Obtain the correlation coefficients. [Use Stat > Basic Statistics ▶ Correlation ... ]
- (c) What would be the effect on the correlation coefficient of the following changes to the variables taste and acetic (relative to the correlation between the original two variables)? Specifically, which of the following will lead to a different correlation coefficient? If you are unsure of your answer, you can find out by making new variables and examining the relevant correlations.
  - (a) Each taste score is increased by 5.
  - (b) The taste scores are squared.
  - (c) Instead of acetic, its reciprocal is used, i.e.  $(1/\text{acetic})$ .
  - (d) acetic is measured in different units, e.g. in values that are one hundredth the size of the original variable.

2.4 *In most cases, data can be viewed as a sample, which has been obtained from some population. The population might be real, but more often it is hypothetical. Our statistical analysis of the sample is intended to enable us to draw inferences about this population. In many cases, we would like the inference to be even broader. For example 45 first-year psychology students at the University of Melbourne undertake a task and their times to completion are measured.*

*This can be regarded as a sample from the population of first-year psychology students at the University of Melbourne. We may wish to apply our results to all undergraduate students of the University of Melbourne; maybe all university students; or even all adults.*

For each of the following data sets:

- i. What population would correspond to this sample? Is this population real or hypothetical?

- ii. Under what circumstances would you be prepared to apply conclusions drawn from analysis of these data to a larger (more general) population?
- (a) 20 items from a production line at Grokkle Manufacturing are tested for defects.
  - (b) 16 women attending the Omega weight loss program have their weight loss recorded after six months.
  - (c) For 64 trains arriving at Flinders Street station (during 19–25 July 2010), the difference between their times of arrival and the scheduled arrival times are recorded.

*Consider a sample with some treatment applied. For example*

*45 first-year psychology students at the University of Melbourne undertake a task (having smoked a marijuana joint) and their times to completion are measured.*

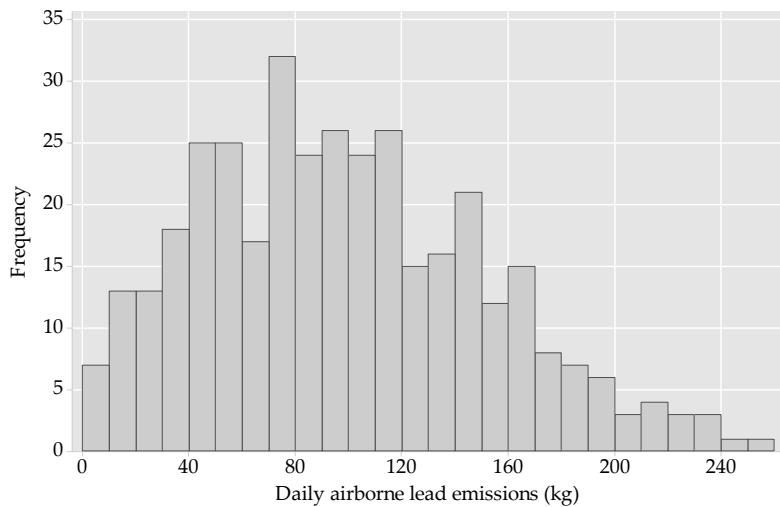
*This can be regarded as a sample from the population of first-year psychology students at the University of Melbourne, having smoked a marijuana joint. This is hypothetical: we have to imagine “what if ...”. We may wish to apply our results to all undergraduate students of the University of Melbourne; maybe all university students; or even all adults ... in each case, having smoked a marijuana joint.*

Answer the above questions (i. and ii.) for each of the following:

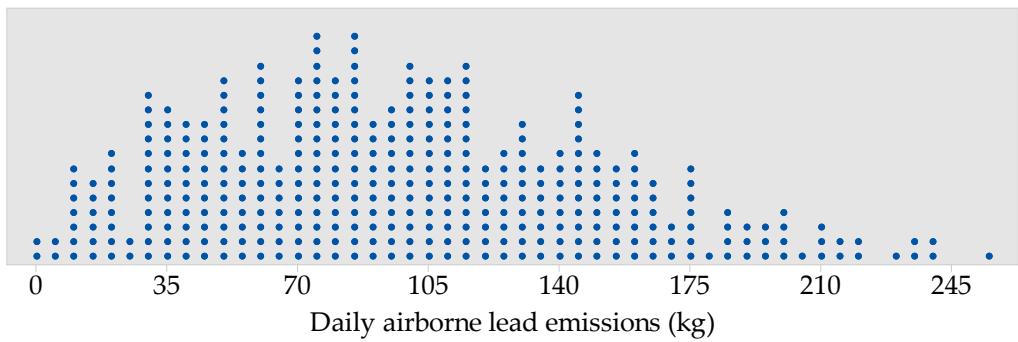
- (d) 30 patients in a Melbourne geriatric care facility were cared for using a new more physically active (PA) regime and their bewilderment ratings are recorded.
- (e) 40 blocks of concrete made using 0.5% of a new additive (XST) are tested for strength.
- (f) 24 women with breast cancer requiring surgery at the Metropolitan Hospital in 2004 were treated with radiation during surgery. Their five-year survival outcomes were observed.

## 2.9 Answers

### 2.1 (a) The histogram:



The dotplot:



- (b) As the data are somewhat skewed to the right, we would expect the mean to be slightly larger than the median. The relevant output is shown below:

#### Descriptive Statistics: Daily airborne lead emissions

Statistics

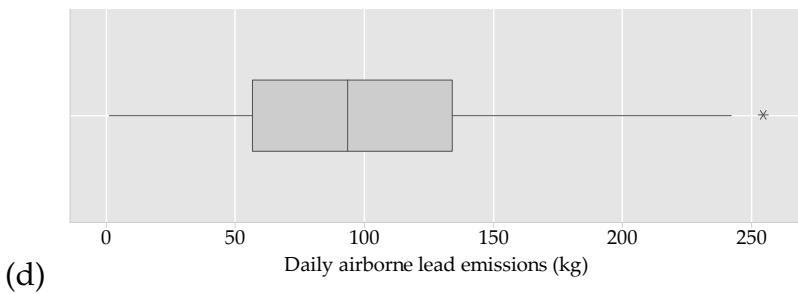
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Daily airborne lead emissions	365	0	98.13	2.80	53.51	1.30	56.65	93.40	134.05
<hr/>									
Variable	Maximum								
Daily airborne lead emissions	254.40								

The mean daily airborne lead emission ( $\bar{x}$ ) is 98.1 kg, and the median is 93.4 kg.

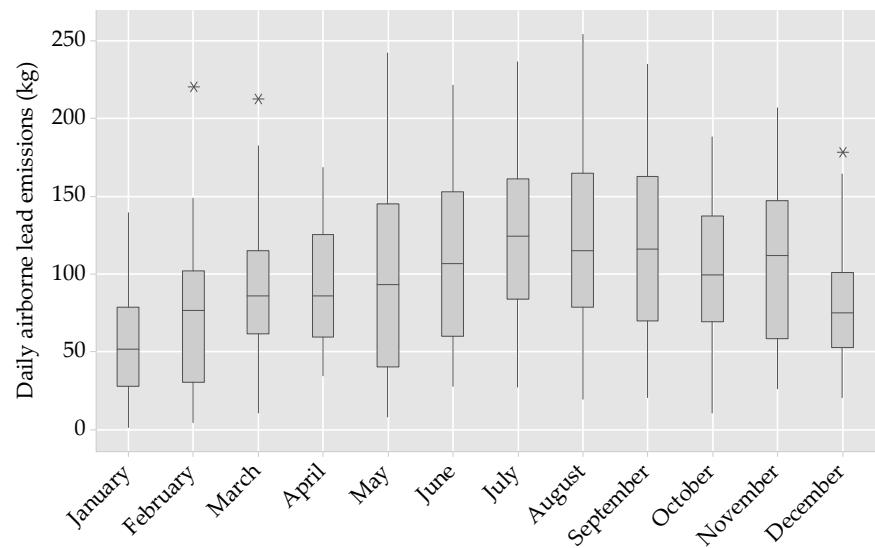
- (c) From the output above, the standard deviation ( $s$ ) = 53.5 kg  
Interquartile range is 77.4 kg

$$\bar{x} \pm 2s = 98.1 \pm (2 \times 53.5) = 98.1 \pm 107.0 = (-8.9 \text{ to } 205.1).$$

There are 365 observations in the data file. The interval (-8.9 to 205.1) contains 351 of the 365 observations (= 96%).



(d)



(e)

From the boxplots, there appears to be variation across the year, with lower levels in summer and higher levels in the mid-year months. However day to day variation in any month is also quite substantial.

(f) Here is the relevant output:

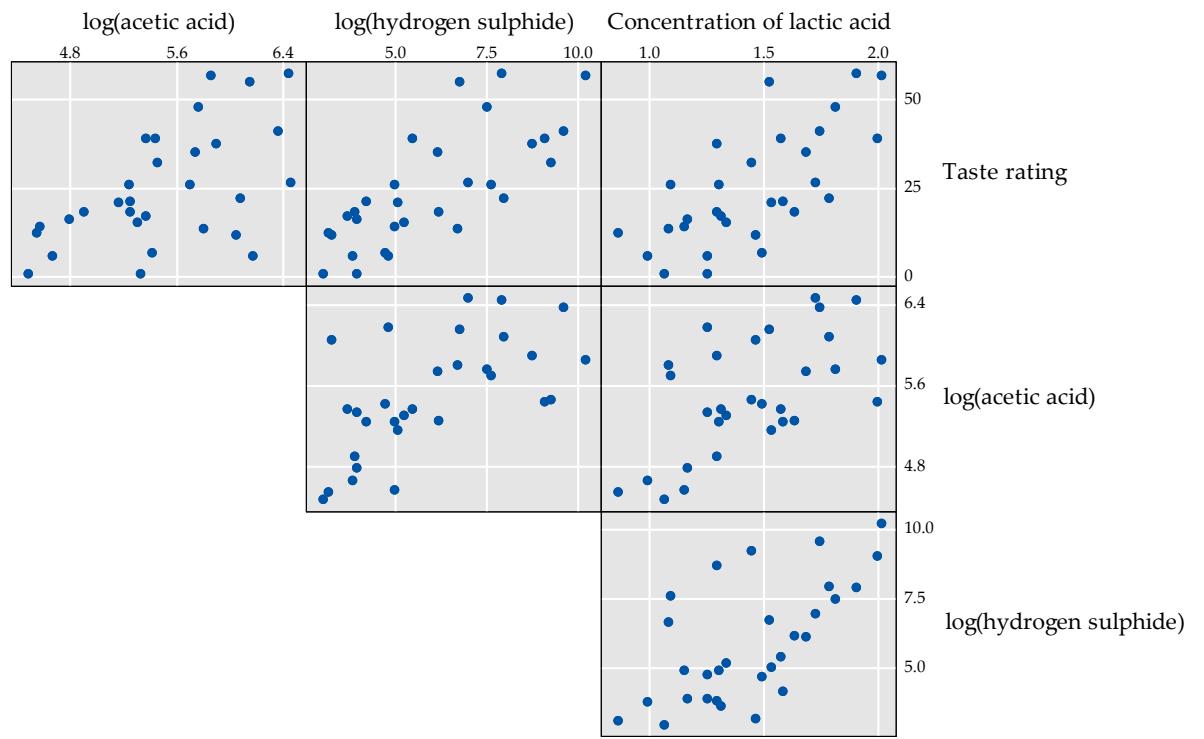
### Descriptive Statistics: Daily airborne lead emissions Statistics

Variable	month	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median
Daily airborne lead emissions	January	31	0	56.98	6.30	35.05	1.30	27.70	52.00
	February	28	0	76.54	9.55	50.55	4.40	30.43	76.60
	March	31	0	87.32	8.48	47.21	10.60	61.90	85.90
	April	30	0	94.05	7.25	39.72	34.80	59.63	85.80
	May	31	0	97.1	11.1	62.0	8.3	40.2	93.4
	June	30	0	107.53	9.15	50.11	28.00	60.30	106.80
	July	31	0	124.9	10.2	56.6	27.5	84.1	124.6
	August	31	0	124.2	11.6	64.7	19.4	79.0	114.9
	September	30	0	118.61	9.88	54.13	20.50	69.73	116.15
	October	31	0	100.47	8.34	46.43	10.80	69.40	99.30
	November	30	0	107.31	9.95	54.52	26.40	58.65	112.20
	December	31	0	81.63	6.44	35.88	20.40	52.70	75.30
Variable	month			Q3	Maximum				
Daily airborne lead emissions	January			78.80	139.50				
	February			102.38	220.30				
	March			115.30	212.30				
	April			125.45	168.60				
	May			145.1	242.2				
	June			153.03	221.60				
	July			161.4	236.4				
	August			165.0	254.4				
	September			162.90	235.00				
	October			137.60	188.40				
	November			147.43	206.90				
	December			101.00	178.30				

In January, the mean daily airborne lead emissions was 57.0 kg; the median was 52.0 kg. In June, the mean daily airborne lead emissions was 107.5 kg; the median was 106.8 kg.

(g)\* There are many important issues to consider. The protocol for collecting daily samples should be examined to identify factors that might introduce unwanted variation into the measurements. You might ask, for example, if measurements were taken at particular times each day. Or was there a continuous data-logger in place? What was the physical process used to collect the emissions? How are the data processed to obtain the measurement emissions in kg? Was a single measurement taken each day?

- 2.2 (a) The scatterplot matrix shows positive relationships between all the variable pairs.



There are six different correlations; they all look moderately strong and positive.

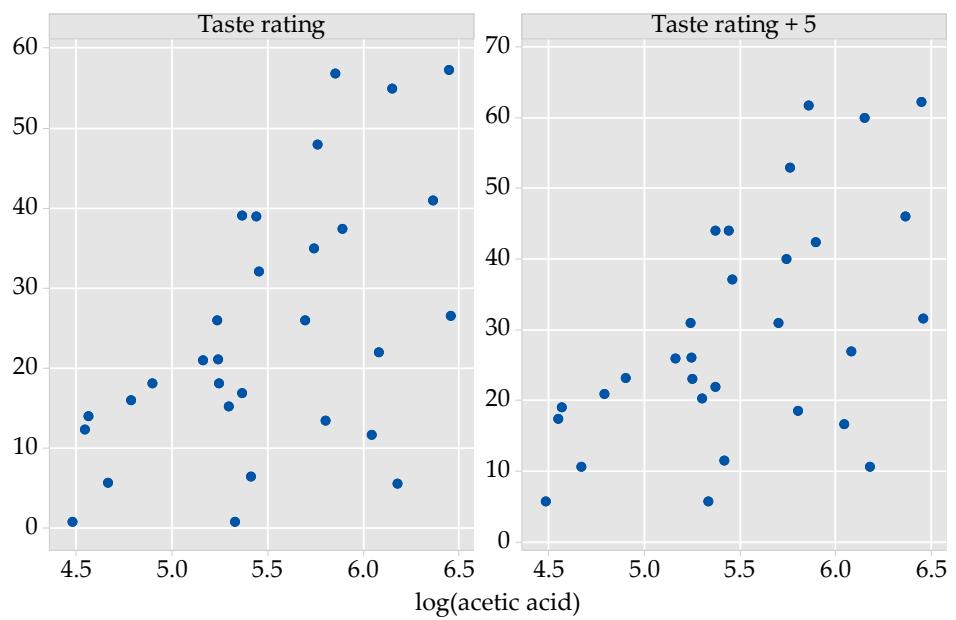
- (b) The relevant output is:

**Correlation: taste, acetic, H2S, lactic**  
**Correlations**

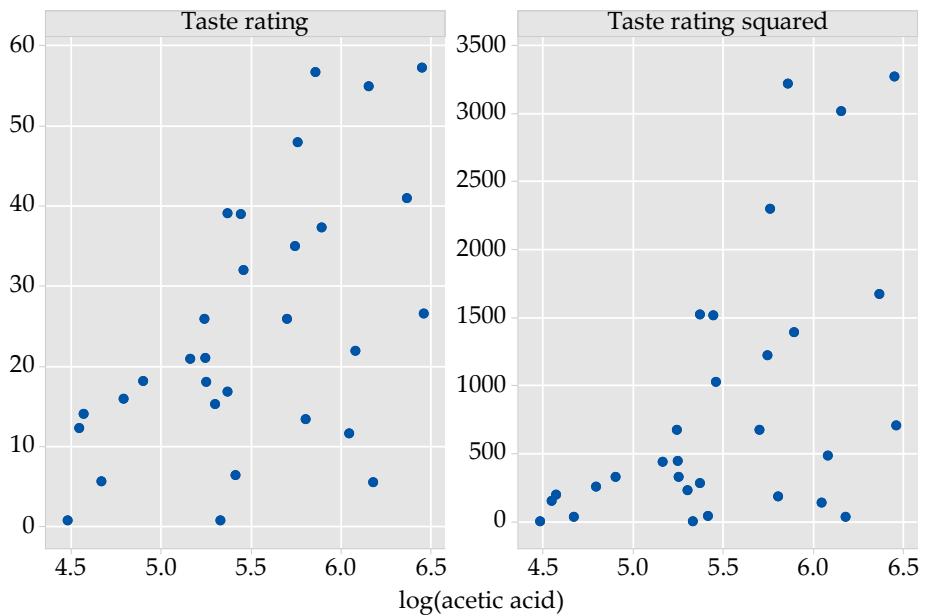
	taste	acetic	H2S
acetic	0.550 0.002		
H2S	0.756 0.000	0.618 0.000	
lactic	0.704 0.000	0.604 0.000	0.645 0.000

*Cell Contents*  
*Pearson correlation*  
*P-Value*

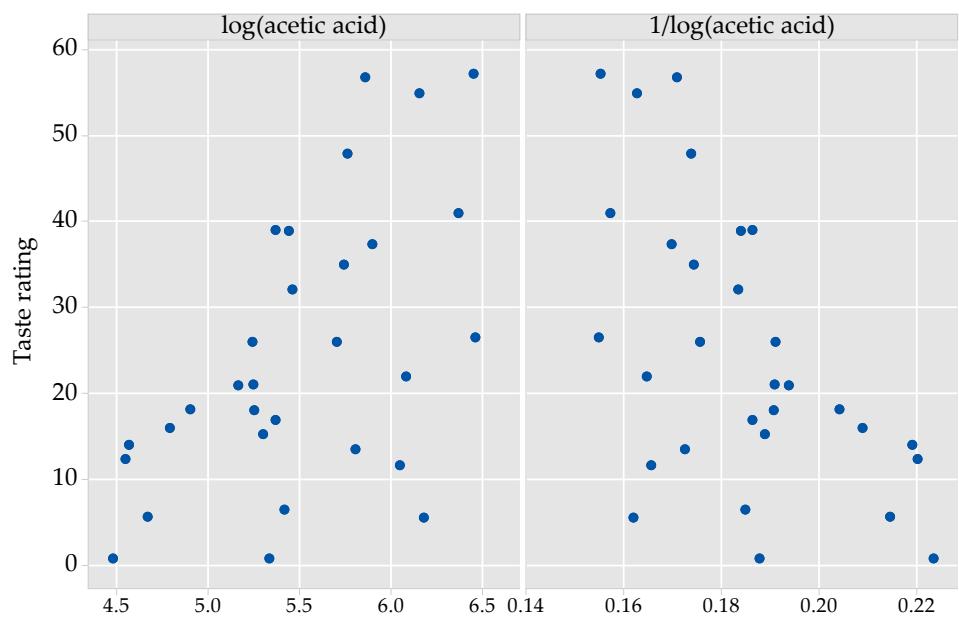
- (c) (a) Taste rating + 5



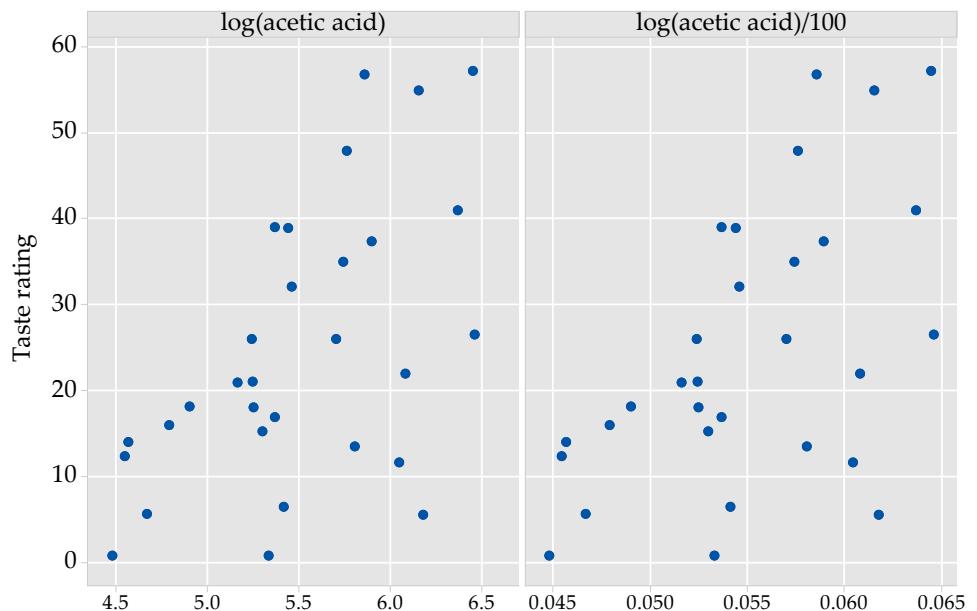
(b) Taste rating squared



(c) Reciprocal of log concentration of acetic acid



(d) log concentration of acetic acid over 100



## 2.4 (a) Grokkle

- (i) The population of items from Grokkle's production line. The items should be sampled at random. An important issue here is time. You can only sample in a particular time period. Strictly speaking, the population could then be real: the population of items in the time period from which you sampled.
- (ii) If the production line process is believed to be stable over time (usually, a very brave assumption!) you might consider applying the conclusions to a longer time period than that sampled. In practice, this is often done: a sample is taken in a week in March, and an inference is drawn about the whole year. This is a rather dangerous practice.

## (b) Omega

- (i) The women should be chosen at random. The population could then be the women attending the Omega program. Issues of time, location, program leader are all relevant here.
- (ii) If the program had a strict protocol for how it was delivered that was followed everywhere you might consider the conclusions to apply to any Omega program. Would you?

## (c) Trains

- (i) The trains should be chosen at random from among all trains arriving in that calendar period. Then the population corresponding to the sample is real and is all trains arriving at Flinders Street during July 19-25 2010.
- (ii) See the Grokkle discussion; it is easy to see that using this sample to make an inference about 2010 (for example) could be quite problematic. New train operator? Weather? Strikes? Special events?

## (d) Geriatric patients

When an intervention has been used, the circumstances in which it was applied are usually very important. We could say that the population is all geriatric patients “like the ones in the Melbourne facility”, but that begs the question. This is why randomization is so important in assessing an intervention. When we have a randomized trial, and therefore some patients with the intervention and some not, it can be reasonable to apply the conclusions more widely, to all geriatric patients. Effectively, this is often done.

## (e) Concrete

If the 40 blocks were chosen at random from a production line of blocks, then we have a similar situation to Grokkle, and we could apply conclusions to the population of blocks made using this additive.

## (f) Breast cancer

Similar issues to the geriatric patients arise.



### 3 Principles of communicating analytics

Care in data exploration and analysis should be matched with high quality communication of the features of the data and the results of the analysis. In this chapter, we cover principles for producing high quality graphs and summary tables. This includes graphs and tables of raw data, and of the results of analysis. It is important to establish these principles now and we endeavour to apply these principles to the graphs and tables we present throughout this course. You should refer back to this chapter and these principles in communicating any of the analytics you learn about here.

#### 3.1 Quality graphs

Computers and statistical software have a lot to answer for in the production of bad graphs. When a graph could only be produced slowly and painstakingly, by a graphic artist, a lot of thought went into the design and construction of the graph, before it was produced. The ease with which statistical graphs can now be produced can lead to careless preparation and design, and while they can also be relatively quickly edited, often this is not done. of your data is “Is the space required for the figure justified?” Sometimes a graph is not a good idea. If the data you wish to discuss amounts to a very small number of values, it may be better to use a table. “A picture is worth a thousand words”; but a picture is not necessarily worth one or two words. A useful rule of thumb is to consider using sentences for displaying 2-5 numbers. Graphs and tables should be considered for more information, graphs being particularly effective for showing complex relationships. This is not a strict rule; in some contexts, a graph may be the most

One of the first questions to ask when considering producing a graph appropriate presentation even for only five numbers.

When considering your graph, you should think about an explanation of your graph in the text. At the least, the graph needs to be appropriately referenced at the right spot in the flow of your argument; it is still possible to find journal articles with figures in them that have not been referred to at all in the text. But usually, more than a reference is needed; some discussion or interpretation of the graph will aid the reader. You will note that we have not always provided extensive discussion of this type in the examples we present; our purpose here is to focus on the principles for good communication with graphs rather than to discuss the (interesting) content of our examples in detail.

We have modelled important principles for good graphical presentation in this chapter, and now elaborate five principles of good graphics and rules of thumb that follow from them. These principles are, of course, inter-related,

and the same rule of thumb can follow from more than one principle.

## 3.2 Five principles of good graphs

Table 1 summarises the plots described in these notes and the contexts in which you might use them. This table should not be seen as prescriptive; variations in use may be appropriate.

Table 1: Guide to the use of plots

Number of variables		Types of plot to consider
Numerical	Categorical	
1	0	dotplot, histogram, boxplot
0	1	dotchart, barchart
2	0	scatterplot, line plot
1	1	boxplot, individual value plot, dotplot
0	2	dotchart
3	0	scatterplot matrix
2	1	panel graph using scatterplots
1	2	panel graph using dotplot or histogram
0	3	panel graph using dotcharts

### 3.2.1 Show the data clearly

The first principle is to show the data clearly. This may seem self-evident, but there are many ways in which poor choices in the design of a graph can obscure the message you wish to communicate. The purpose of the graphic should influence its construction, so you need to consider your audience and the message.

Showing the data clearly means more than simply plotting the data points clearly. When you produce a graph in software, the defaults are produced automatically. In the labelling and title, for example, the software may take the variable names from the data worksheet. This is useful, but not necessarily optimal: you might want to make the labelling more explanatory. Rarely will the software include all the information that you need to include on a graph (the source of the data, for example). The source of the data can be included in the title or footnote of a graph, or described in the related text. Showing the data clearly means the interpretation of the data represented on the plot is unambiguous.

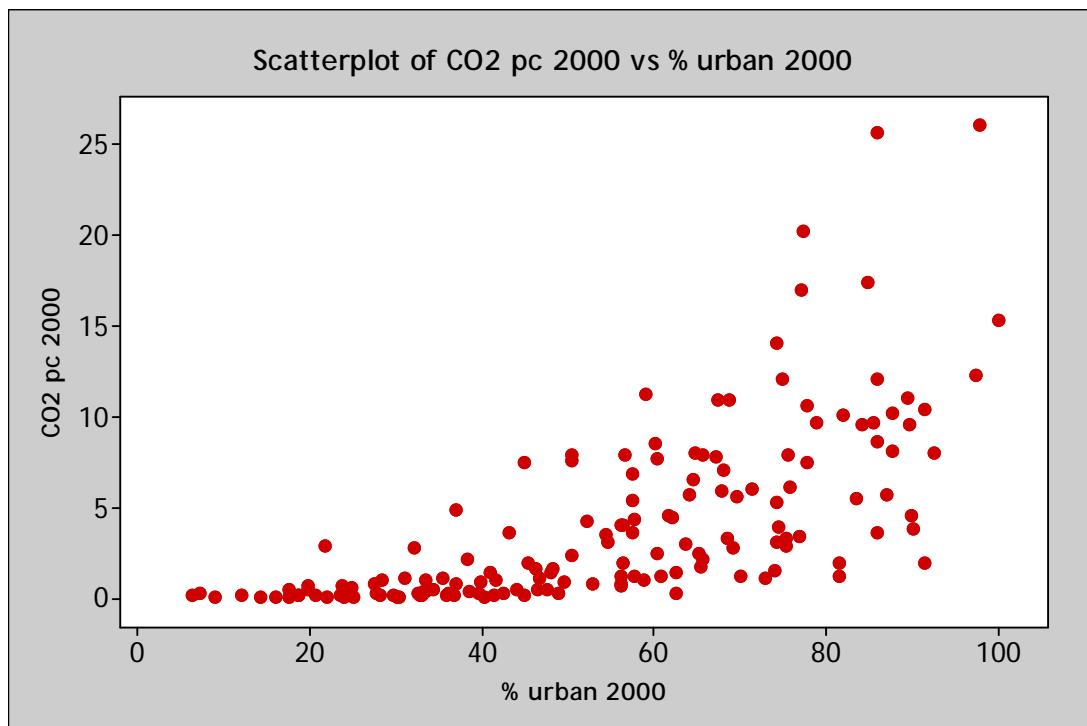
## Rules of thumb

### Good graphs

- identify the source of the data (as part of the graph, or in accompanying text);
- have accurate and informative titles;
- have well-labelled axes with the measurement units clearly defined on the graph (or elsewhere, if this is not possible);
- maintain constant measurement scales;
- avoid distractions and distortions.

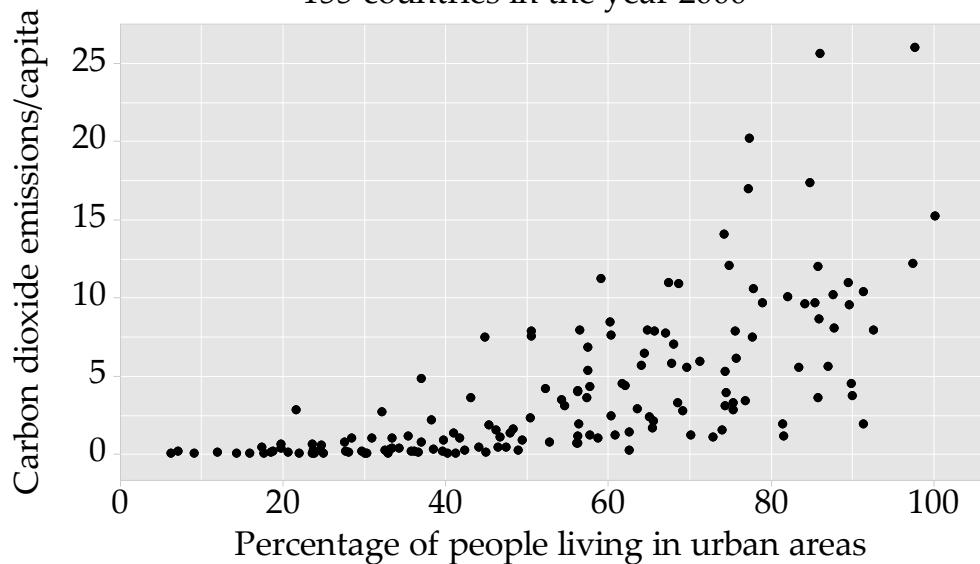
Figure 30 shows two plots of carbon dioxide per capita versus percentage of the population living in urban areas in 2000.

The plot on the top is automatically generated by MINITAB; this is a reasonable first plot. The plot on the bottom, however, shows the improvements that can be made by following the rules of thumb above.



Carbon dioxide emissions per capita (metric tons)  
versus percentage of people living in urban areas

155 countries in the year 2000



Source: *EarthTrends, World Resources Institute*

Figure 30: The figure on the top is automatically generated; the figure on the bottom shows the data more clearly.

In Figure 30 the data points are presented clearly. Distortions of the data are discussed next.

### 3.2.2 Use simplicity in design

Graphic designers are often tempted to embellish figures with decorative adornments, such as 3-D effects, shadowing, and perspective. Software often provides options to use shapes (two or three dimensional) to represent data points.

Representing data in a way that is not proportional to their actual sizes is inaccurate and ineffective. One way that this can occur is by attempting to representing data by using three dimensions, with shading, perspective, or volumes. This is a bad idea. It almost always leads to lower clarity than can be achieved using a two-dimensional plot. A classic example is Excel's "ribbon plot". The use of pictures or complex icons as symbols for data is also a dangerous practice; again, the pictures will not be proportional to the data.

Any attempt at adornment that is unrelated to the data *per se* is likely to be distracting; it is most important to make the data stand out, and to avoid unhelpful distractions. Any of these can block effective communication. Additions in or around the figure should also be avoided. It is almost always better to represent the data in the simplest way.

#### Rules of thumb

- Keep it simple, statistically.
- Omit extra elements in a statistical graph, such as pictures or icons.
- Represent data in a way that is proportional to their actual sizes.
- Avoid representing data with shading, perspective or volume.
- Avoid representing data using pictograms.

#### ▷ EXAMPLE. Greenhouse gas emission (MINITAB worksheet: emission.mwx)

Here are data describing the contributions of various sources to Australia's greenhouse gas emissions. These data were published in February 2008 by the Australian Government's Department of Climate Change: "National Inventory Report 2005 (Revised): Australia's National Greenhouse Accounts." A graph based on them appeared in the *Royal Auto* (March, 2008, p.12). Royal Auto is the monthly publication of the Royal Automobile Club of Victoria (RACV).

Stationary energy	50%	Land use, forestry	6%
Agriculture	16%	Fugitive emissions	6%
Passenger cars	8%	Industrial processes	5%
All other transport	6%	Waste	3%

Figure 31 shows a graphic representing the percentage of greenhouse gas emissions from various sources in Australia. This graph was published in Royal Auto in March 2008. Figure 32 shows the same data, plotted following the rules of thumb above.

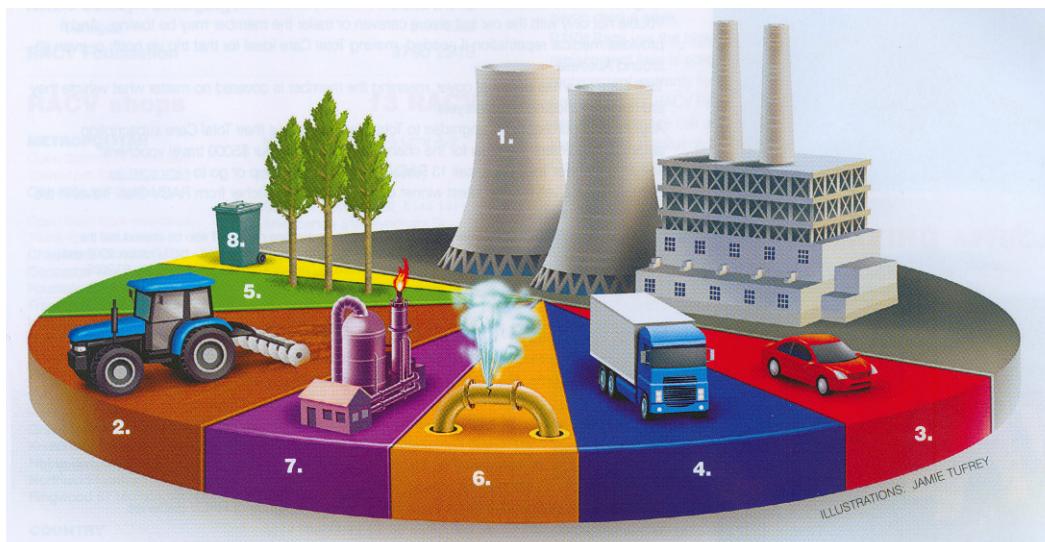


Figure 31: *Greenhouse gas emissions by source, Australia* (Source: Royal Auto, March 2008, p.12.)

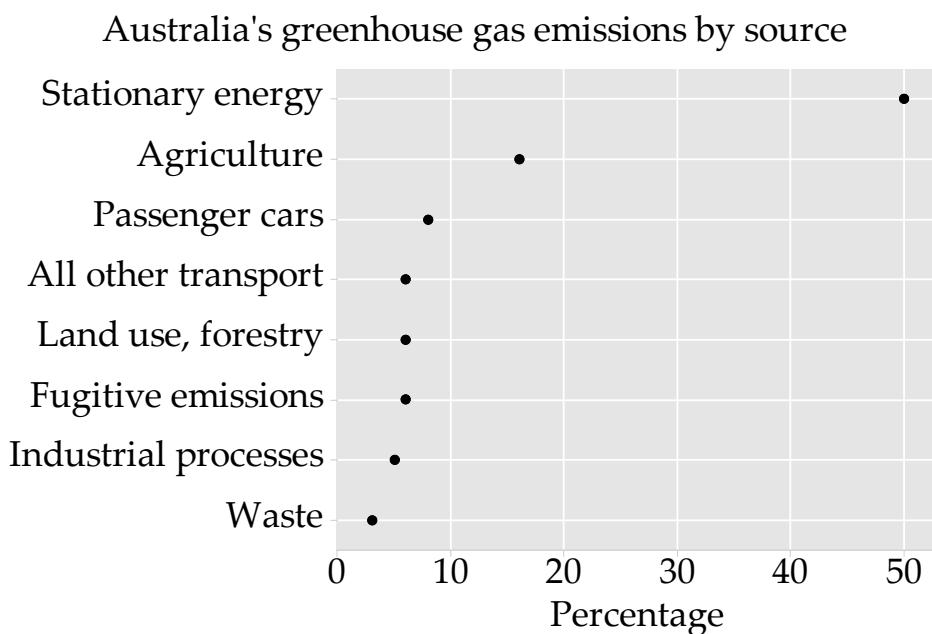


Figure 32: *Greenhouse gas emissions by source, Australia* (Source: National Inventory Report 2005 (Revised): Australia's National Greenhouse Accounts (2008).)

### 3.2.3 Use good alignment on a common scale for comparison

In this chapter there are many examples of graphics with a common scale for comparison. Often it is useful to use the horizontal axis of a graph for the measurement scale. Light gridlines can help guide the comparisons, and can assist with accurate estimation of the values represented by the data points.

Some often-used types of charts do not conform to this principle. Pie charts do not promote accurate comparisons, because the human eye is bad at comparing angles, not lined up along a common scale. Use a plot like Figure 32. Compare Figure 31 with Figure 32 — they show the same data.

For similar reasons, stacked bar charts are ineffective. Unstacked bar charts are lined up along a common linear scale: a good thing. However, they can be tricky to read if there are many adjacent groups. Sometimes, a panel plot will be better.

#### ▷ EXAMPLE. Alligator food preferences

In Figure 26 the percentages of different types of food eaten by alligators of different sizes and genders were presented. In Figure 33 shows percentages as a stacked bar chart. In the stacked version of the graph, only the first category (bird) is aligned at the same starting point; the other categories of food are not aligned, making comparisons challenging.

We have already seen a better graph — Figure 26. Another alternative is shown in Figure 34 which uses panels and plots the percentages as points. However, it is not always easy to make comparisons with this kind of display when there are many levels of the outcome to consider.

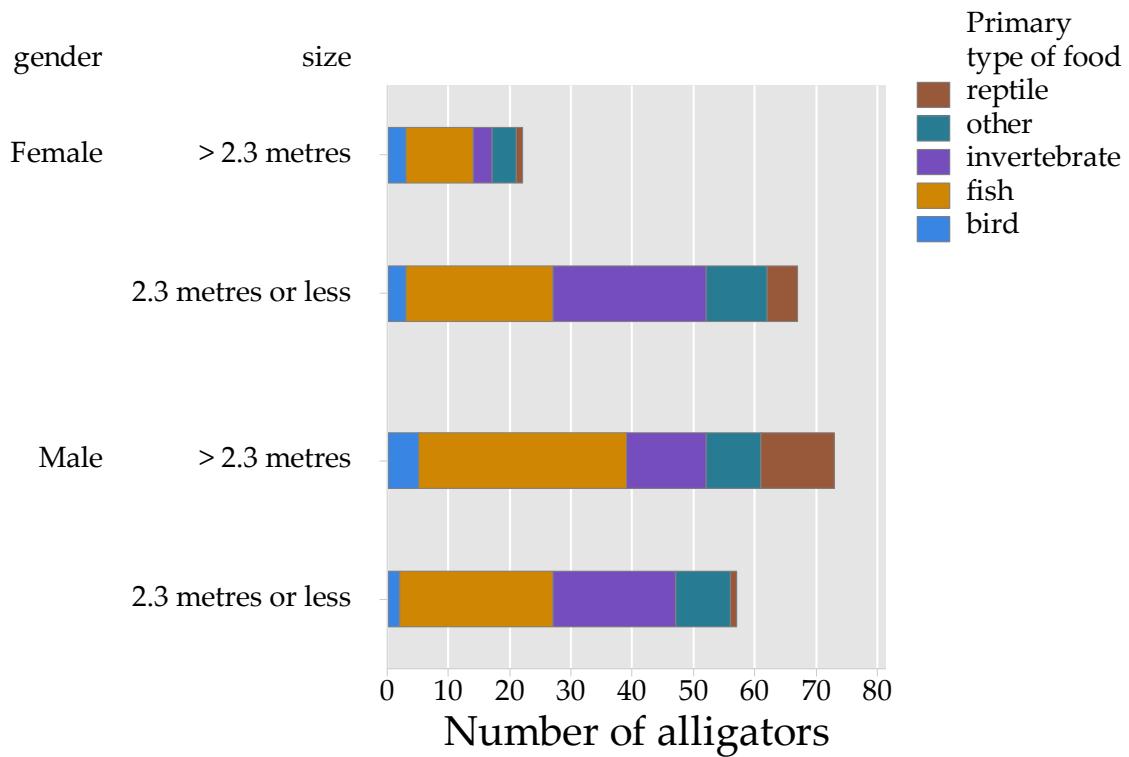


Figure 33: Stacked bar chart of alligator food preference by gender and size

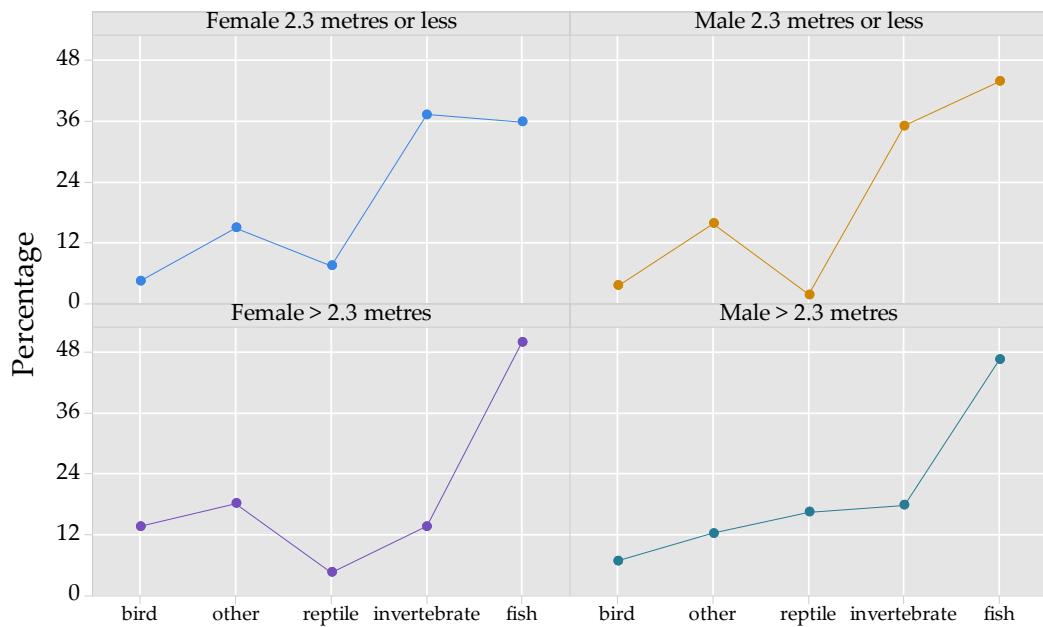


Figure 34: Panelled dot chart of alligator food preference by gender and size

## Rules of thumb

- Consider transposing a figure to find the easiest way to display the common scale.
- Add light gridlines on the common scale to assist with accurate comparison and estimation.
- Never use a pie chart.
- Never use a stacked bar chart; consider panel plots instead.
- If there are many adjacent groups or categories to plot, consider a panel plot.

### 3.2.4 Keep the visual encoding transparent

Any graph involves the encoding of data which then requires the viewer to decode the image. The task of the creator of the graph is to make the visual decoding as simple as possible. If possible, the decoding necessary should be transparent: the viewer should be barely aware of doing it.

There are many ways in which you can make it difficult for the reader to decode your figure. Again, simplicity in design and clear labelling of your figure contribute to easier decoding. Here are some other aspects.

Consider the use of your graph: colour is good and useful if the reader will see the graph in colour. But what if the graph is likely to be photocopied in black and white, for example: will the graph still work then? Some form of shading is an alternative to colour, or, sometimes, different shaped symbols. Avoid patterns such as cross-hatching if possible. Some of these patterns can lead to the “Moiré effect”, which is the phenomenon of an image appearing to shake or shimmer as you look at it.

Transparent visual encoding means defining the information that is represented on your graph. If you use bars to represent something in a graph, such as a standard deviation or a standard error, or a confidence interval, you should explain what the bars are, since the reader won’t know.

Time series plots can be difficult to decode if they are not well-designed. The plot of goat exports in Figure 35 illustrates how poor design can make the patterns over time difficult to decode.



Figure 35: Goat exports from Australia by destination (Source: [www.livecorp.com.au](http://www.livecorp.com.au))

Consider one alternative in Figure 36, which shows the data for countries in the Middle East and Asia ('other' is not shown). Another alternative is shown in Figure 37 in which all countries are in separate panels. In this plot, the exports to Netherlands Antilles have been added to 'other' as they only occurred in 1999. It is not appropriate to produce a time series plot with two variables on it and different scales for the two variables, especially if the two variables can be sensibly compared on the same scale.

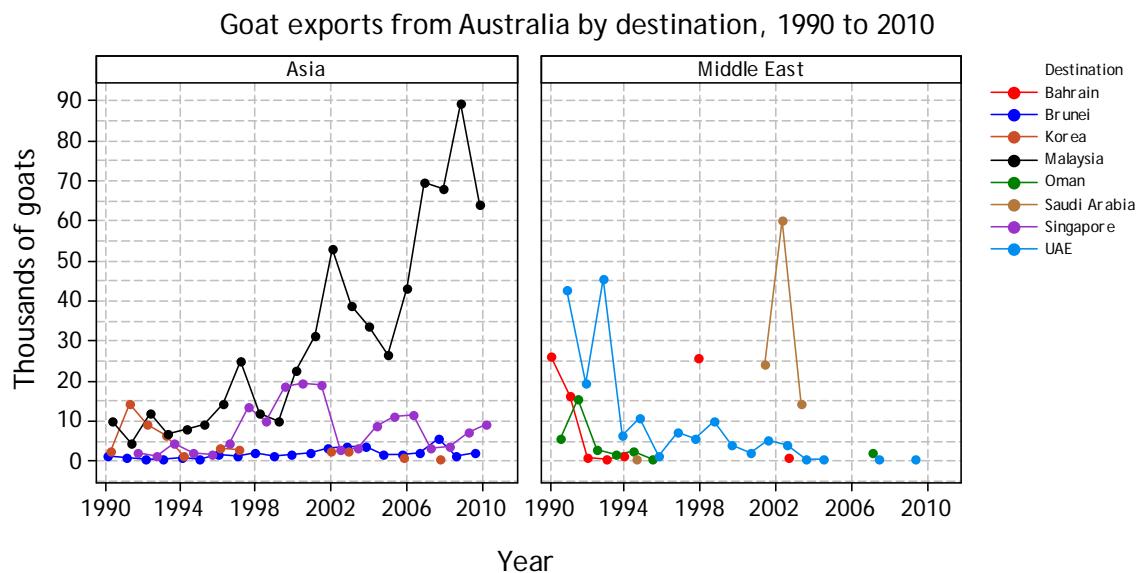


Figure 36: *Goat exports from Australia by destination* (Source: [www.livecorp.com.au](http://www.livecorp.com.au))



Figure 37: *Goat exports from Australia by destination* (Source: [www.livecorp.com.au](http://www.livecorp.com.au))

### Rules of thumb

- Stick to simplicity in design.
- Consider who will use your graph and how it might be reproduced and distributed.
- Avoid cross-hatching.
- Define additional features of the figure, such as bars around points.
- Avoid the use of more than one scale.
- Be careful about the use of colour.

#### 3.2.5 Prefer standard forms demonstrated to be effective

The imperative to be creative in producing a good figure should not drive you to use non-standard graphical forms. Don't invent your own form of a particular plot; you should follow standard conventions for the construction of a plot. The most important example of this is the boxplot, which has accepted traditions for the way the box and whiskers are defined. There are six standard and effective graphical forms that can cover the majority of graphing needs, noted in bold face below. Consider the type of plot that is most appropriate for your data, and if you need to add complexity, consider using a panel display of the plot you have chosen.

The following rules of thumb relate to the summary in Table 1 (page 68).

### Rules of thumb

- The distribution of a single numerical variable in a large sample is best shown using a **histogram**.
- **Dotplots** are appropriate for the distribution of a numerical variable in a small sample.
- To display the main features of distribution of a numerical variable, especially by a categorical variable, use **boxplots**.
- **Dotplots** can also be used to show the distribution of a numerical variable by a categorical variable in large sample(s).
- To show the relationship of a variable with time, use a **line plot** or more particularly a **time series plot**.
- To show the relationship of numerical variable with categorical variable, use a **bar or dot chart**.
- To show the relationship between two numerical variables, use a **scatterplot**.
- Follow standard conventions for the construction of a plot.
- Extend standard forms using panel graphs, where possible and appropriate.

### 3.3 Quality tables

Tables are an effective way of presenting various forms of data, but care is needed to ensure that the tables are easy to read. Tables may present raw data, summary statistics or the results of statistical inference. Tables of statistical output from software packages invariably can be improved.

There are a number of rules of thumb for effective use of tables. Most of these rules apply irrespective of the information in the table; some are more specific. They reflect the same kinds of principles for effective communication that were discussed for graphs.

- Keep the look of the table clean and simple.
- Be judicious in the use of lines; some academic journals insist on no vertical lines, for example.
- Start with no lines on your table; add lines to reflect important distinctions in the structure.
- Use headings to clearly define the data in the table.
- Use accurate and complete labels for columns and rows, as appropriate.
- Clearly label the statistics reported in rows or columns.
- Use accurate labels for elements in the table, rather than abbreviations that might be used in your data file.
- Consider how many significant figures are needed for the measurements you are reporting; two is usually adequate unless the purpose of the table is to record the 'original' data.
- Consider adding row and column means or totals, if they will aid interpretation.

- Consider your choice of row and column variables; tables where the numbers in the columns are approximately equal are easier to interpret than tables in which the numbers in the rows are approximately equal.
- Consider changing the default ordering used by the software unless there is a natural ordering, such as time. It may be useful to arrange the rows and columns in order of their (mean) size.
- Lightly shade alternating rows if there are many rows in your table.
- Numbers in columns should have consistent decimal alignment.

### 3.3.1 Tables of data

Here is a simple example that develops the presentation of a table of data, following the rules of thumb above.

▷ **EXAMPLE. Sales (in thousands of dollars) by area and period**

Table 2: The raw data, poorly presented

	1	2	3	4
A	97.63	92.24	100.90	90.39
B	48.29	42.31	49.98	39.09
C	75.23	75.16	100.11	74.23
D	49.69	57.21	80.19	51.09

In Table 2 there are several problems. The absence of lines of any sort makes the table harder to read. Two decimal places, when the data range from 39 to 101, is useless accuracy and distracting from the actual variation. The row and column headings are vague and therefore unhelpful. It's hard to get an immediate idea of which areas and quarters were generally high and which were generally low.

Table 3: Appropriate headings, two significant figures, row and column means

Area	Quarters (2009)				Mean
	I	II	III	IV	
North	98	92	101	90	95
South	48	42	50	39	45
East	75	75	100	74	81
West	50	57	80	51	60
Mean	68	67	83	64	70

Table 3 is a big improvement. The year to which the quarters belong is now clearly identified. The areas are identified by name. Row, column and total means are worked out, to invite comparison.

Table 4: Interchanging rows and columns and re-ordering of columns

2009	Area				Mean
	North	East	West	South	
Quarter I	98	75	50	48	68
Quarter II	92	75	57	42	67
Quarter III	101	100	80	50	83
Quarter IV	90	74	51	39	64
Mean	95	81	60	45	70

Table 4 is only a slight variation on Table 3; the rows and columns have been interchanged and the areas (now in columns) have been ordered from largest to smallest (according to the average) to aid the comparisons the reader of the table will inevitably make. Whether the table should be oriented as in Table 3, or as in Table 4, is debatable, although it is easy to see that data like these could be extended in time, to include further years, and if that were done, it would be easier to add more rows to Table 4 than columns to Table 3.

### 3.3.2 Tables of summary statistics

▷ **EXAMPLE. Life expectancy** (MINITAB worksheet: countries.mwx)

The output below gives summary statistics provided by MINITAB when we request: Stat > Basic Statistics ▶ Display Descriptive Statistics by region for the column containing life expectancy in 2000-2005 for all countries:

#### Descriptive Statistics: Life expectancy at birth 00-05

##### Statistics

Variable	Region	N	N*	Mean	SE Mean	StDev	Minimum
Life expectancy at birth 00-05	North America	2	0	78.20	1.10	1.56	77.10
	Middle East and North Africa	18	0	69.39	1.90	8.05	43.10
	Oceania	5	0	70.82	3.90	8.73	57.60
	Asia	28	0	67.55	1.22	6.46	54.50
	South America	12	0	70.77	1.18	4.09	63.20
	Central America	13	0	70.18	2.01	7.25	49.50
	Europe	36	0	75.147	0.607	3.640	66.800
	Sub-Saharan Africa	41	0	45.47	1.02	6.53	32.40
Variable	Region	Q1	Median	Q3	Maximum		
Life expectancy at birth 00-05	North America	*	78.20	*	79.30		
	Middle East and North Africa	68.77	71.45	72.97	79.20		
	Oceania	63.40	69.80	78.75	79.20		
	Asia	63.13	68.70	72.35	81.60		
	South America	68.52	71.00	74.08	76.10		
	Central America	67.80	71.30	75.20	78.10		
	Europe	72.100	75.800	78.450	80.100		
	Sub-Saharan Africa	40.50	45.70	50.15	57.90		

We provide two tables of summary statistics derived from the output above;

Table 5 shows the mean and standard deviation by region; Table 6 gives the five number summaries by region.

Table 5: An example of a table of summary statistics, ordered by the mean

Mean and standard deviation for life expectancy at birth in 2000-2005 by region

Region	Number of countries	Mean	Standard deviation
North America	2	78.2	1.6
Europe	36	75.1	3.6
Oceania	5	70.8	8.7
South America	12	70.8	4.1
Central America	13	70.2	7.3
Middle East and North Africa	18	69.4	8.1
Asia	28	67.6	6.5
Sub-Saharan Africa	41	45.5	6.5

Table 6: An example of a table of summary statistics, ordered by the median

Five number summary for life expectancy at birth in 2000-2005 by region

Region	Number of countries	Minimum	First quartile	Median	Third quartile	Maximum
North America	2	77.1	78.2			79.3
Europe	36	66.8	72.1	75.8	78.5	80.1
Middle East and North Africa	18	43.1	68.8	71.5	73.0	79.2
Central America	13	49.5	67.8	71.3	75.2	78.1
South America	12	63.2	68.5	71.0	74.1	76.1
Oceania	5	57.6	63.4	69.8	78.8	79.2
Asia	28	54.5	63.1	68.7	72.4	81.6
Sub-Saharan Africa	41	32.4	40.5	45.7	50.2	57.9

Note the different ordering in Tables 5 and 6. In practice, you would use one ordering only. If multiple tables by region were required, you would fix on a more general ordering such as alphabetical or (here) possibly geographical and maintain the same order for all tables.

### 3.3.3 Tables of inferential statistics

Although we have not discussed inferential statistics yet, for completeness we provide some output and appropriate summary tables for this type of output, following the rules of thumb described earlier. You should not expect to fully understand the tables presented here, but can refer back to this section later in the course.

#### ▷ EXAMPLE. White pine (MINITAB worksheet: pine.mwx)

Here is the MINITAB output from a General linear model comparing the mean moisture content in white pine under three different storage conditions. The analysis of these data is discussed in detail in Chapter 10.

## Descriptive Statistics: moisture

### Statistics

Variable	condition	N	N*	Mean	StDev
moisture	1	5	0	8.020	0.545
	2	3	0	6.633	1.079
	3	3	0	9.333	0.764

## General Linear Model: moisture versus condition

### Method

Factor coding (1, 0)

### Factor Information

Factor	Type	Levels	Values
condition	Fixed	3	1, 2, 3

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
condition	2	10.939	5.4693	9.35	0.008
Error	8	4.681	0.5852		
Total	10	15.620			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.764962	70.03%	62.54%	37.80%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.020	0.342	23.44	0.000	
condition					
2	-1.387	0.559	-2.48	0.038	1.16
3	1.313	0.559	2.35	0.047	1.16

### Regression Equation

$$\text{moisture} = 8.020 + 0.0 \text{ condition\_1} - 1.387 \text{ condition\_2} + 1.313 \text{ condition\_3}$$

## General Linear Model: moisture versus condition

### Comparisons for moisture

#### Tukey Pairwise Comparisons: condition

#### Tukey Simultaneous Tests for Differences of Means

Difference of condition Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	Adjusted T-Value	Adjusted P-Value
2 - 1	-1.387	0.559	(-2.983, 0.209)	-2.48	0.087
3 - 1	1.313	0.559	(-0.283, 2.909)	2.35	0.105
3 - 2	2.700	0.625	(0.916, 4.484)	4.32	0.006

Individual confidence level = 97.87%

The following table presents the main results in a suitable manner.

Table 7: An example of an arrangement of results from software output.

	<i>Storage condition</i>		
	Condition 1	Condition 2	Condition 3
Mean	8.0	6.6	9.3
Standard deviation	0.6	1.1	0.8
<i>n</i>	5	3	3
<i>Analysis of variance</i>	<i>df</i>	F	P-value
Explanatory variable – Condition	2, 8	9.4	0.008
<i>Tukey's post hoc comparisons</i>			
Comparison	Mean difference	95% CI for mean difference	
Condition 2 – Condition 1	-1.4	-3.0, 0.2	
Condition 3 – Condition 1	1.3	-0.3, 2.9	
Condition 3 – Condition 2	2.7	0.9, 4.5	

#### ▷ EXAMPLE. Young people and the environment

A study was carried out on Year 10 students to investigate their attitudes towards environmental issues. This is discussed in some detail in Chapter 13. The MINITAB worksheet `environ.mwx`<sup>8</sup> contains scores relating to three aspects of the study:

- degree of support for an environmental paradigm as opposed to a technological paradigm (`support`);
- level of environmental knowledge (`know`);
- degree of past involvement in improving the environment (`involve`).

For all three aspects above, the higher the score, the higher the degree or level. The MINITAB output from a general linear model predicting (`support`) from (`know`) and (`involve`) is given below.

---

<sup>8</sup>Permission to use the data was kindly granted by Professor David Yencken, Faculty of Architecture, Planning and Building, University of Melbourne.

## Regression Analysis: support versus know, involve

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	418.7	25.75%	418.68	209.34	8.50	0.001
know	1	166.6	10.24%	40.17	40.17	1.63	0.208
involve	1	252.1	15.50%	252.09	252.09	10.23	0.002
Error	49	1207.4	74.25%	1207.40	24.64		
Lack-of-Fit	39	1015.4	62.44%	1015.40	26.04	1.36	0.315
Pure Error	10	192.0	11.81%	192.00	19.20		
Total	51	1626.1	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
4.96395	25.75%	22.72%	1347.54	17.13%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2.96	3.20	(-9.39, 3.47)	-0.93	0.359	
know	0.400	0.313	(-0.230, 1.030)	1.28	0.208	1.15
involve	0.773	0.242	(0.287, 1.259)	3.20	0.002	1.15

### Regression Equation

$$\text{support} = -2.96 + 0.400 \text{ know} + 0.773 \text{ involve}$$

Table 8 presents some of the inferences from the output in an appropriate layout.

Table 8: Presentation of some regression results from software output.

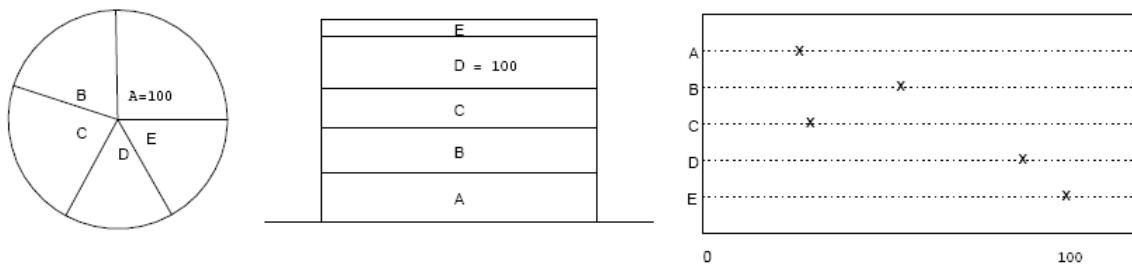
<i>Multiple linear regression</i>		df	F	P-value	Adjusted R <sup>2</sup>
		2, 49	8.5	0.001	22.7%
Coefficient					
Explanatory variable		Estimate	95% confidence interval		P-value
Constant		-2.97	-9.40, 3.47		0.359
Level of environmental knowledge		0.40	-0.23, 1.03		0.208
Degree of past involvement		0.77	0.29, 1.26		0.002

### 3.4 Exercises

3.1 Return to analyses you carried out in exercise 2.1 from Chapter 2 using the data file `lead_smelter.mwx`.

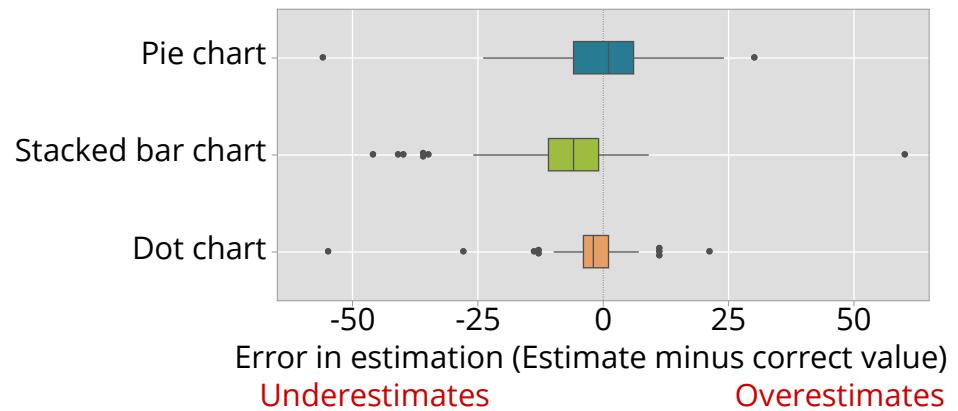
- (a) Review the graphs you produced in Chapter 2 and identify improvements you can make by applying the principles of good graphs discussed in this chapter.
- (b) Based on the material on structuring tables in this chapter, devise a table that presents the means and standard deviations by month, in a form suitable for presentation.

3.2 (a) Given below are three graphs: a pie chart, a bar chart and a dotchart. Each of them represents five numerical quantities (a different five numbers for each graph). On each graph, the longest line or greatest area is marked indicating its magnitude is equal to 100. Your task is to estimate the other lengths or areas on each graph. You should make a quick visual judgment and not try to make precise measurements, either mentally or with a physical object.



In each plot, the maximum is shown, and has value 100. Estimate by eye the values of the other levels. You should obtain four values per plot: twelve in all (and all between 0 and 100!) Write your guesses in the diagram above.

- (b) The figure below shows boxplots of the errors made by students in a recent SRW class in estimating quantities in (a).



- (a) Based on the boxplots, which is the best graphical representation (pie chart, dot chart or bar chart) to use? Give reasons for your choice.
- (b) Based on the boxplots, what problems arise with the other types of graphs?
- (c) Based on the boxplots, which is the least preferred type of graph?  
Explain why.

### 3.5 Answers

- 3.1 (a) Compare the graphs you have produced with those in the answers for Chapter 2, where the principles of good graphs have been applied. Edit your own graphs or apply suitable global defaults to improve them.
- (b) A suitable table is shown below. The table and columns are clearly labelled. Unnecessary lines have been removed. Decimal alignment has been used in the second and third column. In general, it is useful to present the sample sizes; in this case, it indicates to the reader that every day in every month was included.

Daily airborne lead emissions (kg) by month

	Mean	Standard deviation	Sample size
January	57.0	35.1	31
February	76.5	50.5	28
March	87.3	47.2	31
April	94.1	39.7	30
May	97.1	62.0	31
June	107.5	50.1	30
July	124.9	56.6	31
August	124.2	64.7	31
September	118.6	54.1	30
October	100.5	46.4	31
November	107.3	54.5	30
December	81.6	35.9	31
Entire year	98.1	53.5	365

- 3.2 (a) Please record all your guesses on the diagrams before looking at the answers.

Then once you are done, record your answers in the table below, and work out the difference between your estimate and the correct answer.

	Your guess	Correct value	Your guess – Correct value
<b>Pie Chart</b>			
B		78	
C		87	
D		64	
E		66	
<b>Dot Chart</b>			
A		26	
B		54	
C		29	
D		88	
<b>Bar Chart</b>			
A		96	
B		85	
C		76	
E		35	

- (b) (a) The dot chart is the best choice of the three representations. The median is close to zero, and the interquartile range is relatively small. However, a few people do make large errors with the dot chart.
- (b) Overestimation is a common problem with the pie chart. The median is positive. Underestimation is a common problem for the bar chart; the median is negative and 75% of the errors are negative.
- (c) Although the spread of the errors is greater for the pie chart than for the bar chart, the median (and mean) error is greater for the bar chart.

## 4 Foundations for inference: more advanced language

In research we seek to understand the world in a systematic way. We want to gain insight into underlying mechanisms or laws of nature; we hope to be able to predict what will happen in a given situation; we aim to make generalized conclusions about the subject of our research.

These intentions are broadly the same whether our research is classically academic, involving the advancement of knowledge in a particular discipline, or more functional, in a business setting for example, where we may want general commercial insights. We are thinking of “research” in a broad sense. It includes any systematic inquiry of a pattern of phenomena.

This sounds rather esoteric; it is described here only to distinguish this activity from other forms of inquiry, such as “do I like anchovies?”. I or my family may be interested in the answer to this question (when ordering pizza, for example), but it is not of real research interest. It might be, however, if it arises in an organized investigation of the percentage of people who like anchovies.

We do sometimes ask questions about particular events using research approaches. But this is only because they are embedded in a broader whole. The prediction of weather is a classic example. The reason that meteorologists can do a good job of predicting tomorrow’s weather — a particular question — is that they have developed elaborate models, based on past history, of *patterns* of weather in a given location. These patterns are both spatial and temporal. They involve complex understandings of the behaviour of weather, and they rely on a very extensive network of data.

Note, therefore, that from the research perspective the anchovies question and weather prediction both entail a broad, general context. We think of this as the **population** of interest. It is easy to think about this when we have in mind a population in the usual sense: a population of people. Sometimes the “population” is more vague and ill-defined, such as “Melbourne’s weather”. In practice, in such cases, a focussed research question will be about one aspect of the broad context, so that the population can be effectively refined. For Melbourne weather, this might be “daily rainfall in March”, or “maximum wind gusts, annual”. We make a forecast about tomorrow by seeing “today” in terms of a population of days like today, or a population model that includes today’s pattern as a particular case. The implied population may also be very broad in scope. A lot of medical research, for example, has the population of “all humans” in mind, more or less implicitly.

## 4.1 Populations versus samples

We begin our thinking about statistical inference using populations and samples. The reason we use this approach is that it is more tangible than many other approaches; it is easier to think about, and sometimes, directly relevant.

▷ **EXAMPLE. ASIC and The Cash Store** The Australian Securities and Investments Commission (ASIC) brought a case against a “payday” lender called The Cash Store Pty Ltd. ASIC alleged that The Cash Store had breached a number of their obligations under the Credit Act, such as ‘failure to make reasonable inquiries about the customer’s financial situation’.

Random samples of their loans were obtained and examined in detail. On the basis of these, the statistician Ian Gordon gave evidence that drew an inference from the sample to the whole population of loans, using estimates and 95% confidence intervals for the percentages of loans that had breaches. These inferences are exactly the type of inferences we learn about in this and subsequent chapters.

ASIC argued that the court, when setting the penalties, should take into account the inferences from the sample. In her judgment, Justice Davies quoted the 95% confidence intervals for the *overall* numbers of loans with breaches in the population and stated that:

“I consider that it is appropriate in setting the penalty to take into account the analysis conducted by Professor Gordon and the statistical likelihood of similar contraventions in respect of all contracts entered into over the period”.<sup>9</sup>

The award by the Federal Court of nearly \$19 million in penalties was a record civil penalty for ASIC.

An interesting aspect of this example is that the legal system — for good reason — is famously wary of relying on evidence that has not been *directly* examined. So ASIC did not seek findings about the loan contracts not specifically assessed in the sample. Despite this cautious approach, Justice Davies concluded that an inference to the whole population was warranted, because of the random sampling.

This is a simple example of an inference from a sample to a population.

Statistical problems may involve inferences about a property of a population based on a sample from the population.

We might be interested in the proportion of people in a population who are left-handed, or who do their income tax on time, or we might be inter-

---

<sup>9</sup>Australian Securities and Investments Commission (ASIC) v The Cash Store Pty Ltd (in liq) (No 2) [2015] FCA 93;BC201500806.

ested in the mean and standard deviation of the daily rainfall in March in Melbourne, or the slope and intercept of the line which relates the expected blood pressure of an individual to their age.

A comparison might be the goal of our inquiry, so we may be interested in the difference between the proportion of loan defaults for personal loans from two banks.

In several of these cases it is either not possible or impractical to try to determine the exact value of the quantity of interest in the population and the best that can be done is to obtain an estimate based on a **sample**.

It may not be possible to look at the whole population, because it extends into the future. If we want to make an inference about the amount of harvestable timber in a mature mountain ash tree, for example, we would have in mind the population “all mature mountain ash trees”. But . . . which ones? The ones alive today? The ones in a particular plantation or forest? We probably envisage the population as including future mountain ash trees, not even grown yet. The same is true of research on humans; we often envisage that the inference will be relevant to humans in the future. There is an important assumption here, about the stability of the world and its patterns: we cannot sample the future now.

A numerical summary characteristic of a population, such as those mentioned above, is termed a (population) **parameter**. It is vital to have a clear understanding of the distinction between a **population** and a **sample** and between a **parameter** and a (sample) **estimate**.

A sample is a subset of a population. One type of sample which is of special interest in statistical investigations is the so-called **random sample**, or **simple random sample**, which means that all possible samples, of a given size, are equally likely to be chosen. For example if we choose a sample of 5 people from a population of 100 people then there are  ${}^{100}C_5 = 75\,287\,520$  ways this can be done; the sample is random if each of these possibilities is equally likely. What makes a sample random is how it is chosen, not what it consists of.

There are many words used in statistics which have applications and meanings in other contexts. “Sample” is one of these. In statistics the word refers to a subset of a population, and *not* to a single specimen or instance of a physical or biological quantity, such as a sample of soil.

In many research situations, someone has obtained a set of data somehow. They may not have done it with the principles of sampling in mind, but they still wish to make statistical inferences in the standard way. In these situations, it is usually profitable to ask: “What population has this sample come from?” Addressing this question will usually sharpen the research. Similarly, we should get into the habit of asking: “How was this sample collected? What was the method used for choosing the observations? Can

we model it by random sampling?"

For many situations there is an obvious estimator, such as the sample proportion, sample mean and sample standard deviation as estimators of the corresponding population quantities. However, in other situations there is no immediately obvious estimator, as in the case of estimating the slope and intercept of a line.

Corresponding to the distinction between a population and a sample is the other important distinction, between a population **parameter** and a sample **estimate**. For example we might be interested in the proportion of people in a population who are left-handed. The actual proportion of people in the population who are left-handed is referred to as a population parameter. If we take a sample from the population and determine the proportion of left-handed people in the sample, then what we have is a sample estimate of the population parameter. In general it is most unlikely that a sample estimate will be exactly equal to the population parameter and it is most unlikely that different samples from the same population will give the same estimate: rather, they will vary.

It is instructive to use the example of political polling to illustrate these ideas. In such polls, we sample a small number of electors; but it would be possible, in principle, to poll the entire population of electors, and this is essentially what is done on election day. Imagine a poll in which a random sample of 1000 electors is taken *just after they voted*. (Such polls are, in fact, sometimes carried out, although not on a large scale, usually; they are known as "exit polls".) Imagine, too, that the respondents answer the exit poll questions honestly, and consistently with how they vote, so that the only difference between the survey result and the election result is sampling variation. Then the percentage voting ALP, say, from the **sample** is an estimate of the percentage of the **population** voting ALP. But it would not be surprising if the survey estimate was 44.3% and the election result was 42.9%. (Would you be surprised?) If 10 separate surveys using the same method were carried out (since the sampling is at random the chance of any voter being chosen in more than one survey is virtually zero), the survey percentages might look like this:

44.3 47.1 43.9 42.9 43.3 40.3 43.8 44.5 42.6 41.3

Three of these values are below the true value, and six above. One of them is exactly right! But you wouldn't know that at the time: that's what sample values are like. They're estimates of the truth, and the election example, while it illustrates the idea nicely, is actually atypical; when a researcher carries out a survey he or she will usually not find out the true value in the population.

sample	population
estimate, $\hat{\lambda}$	parameter, $\lambda$
mean, $\bar{x}$	mean, $\mu$
proportion, $\hat{\theta}$	proportion, $\theta$

We use the “hat” notation,  $\hat{\cdot}$ , to denote an estimate or estimator; these are estimates of the (hatless) parameters. We could use  $\hat{\mu}$  for the sample mean, but  $\bar{x}$  is the common notation.

In SRW we use Greek letters for parameters. This is a traditional convention of notation, and a useful one. To the extent that it challenges us because the symbols are unfamiliar, it serves as a reminder that population parameters are generally unknowable: fixed but unknown constants. Some of the Greek letters are used for more than one purpose. The table below gives the Greek letters used in SRW, how to pronounce them, and their applications.

Letter	Pronounced	Application
$\alpha$ , alpha	alfa	level of significance in hypothesis testing, block effect in model, intercept in simple linear regression
$\beta$ , beta	beeta	$\Pr(\text{type II error})$ in hypothesis testing, treatment effect in model, slope parameter in regression
$\mu$ , mu	mew	population mean
$\sigma$ , sigma	sigma	population standard deviation
$\theta$ , theta	theeta	population proportion
$\rho$ , rho	roe	population correlation coefficient
$\gamma$ , gamma	gamma	interaction term in analysis of variance for factorial designs
$\lambda$ , lambda	lamda	arbitrary parameter
$\phi$ , phi	fie	another arbitrary parameter
$\chi$ , chi	kie	$\chi^2$ distribution, also $\chi^2$ test
$\Sigma$ , Sigma	sigma	sum of (this is “capital sigma”)
$\nu$ , nu	new	arbitrary degrees of freedom

**Estimates** come from the sample; **parameters** apply to the population of interest. An estimate is a number we can calculate from the sample; the parameter will generally be fixed, but unknown. Estimates vary from sample to sample; the parameter is a single value. Paradoxically, we don’t care at all about the sample and the estimate *per se*, but only in so far as they tell us something about the population.

Keeping in mind this essential idea of the distinction between **samples** and **populations**, and **estimates** and **parameters**, the aim of formal statistical inference is to *infer* something about the parameters of a population, from the sample estimates. This is an **inference**, because we remain uncertain about the parameters; we cannot find them out for sure. It is an inference made with associated uncertainty, and it is the job of statistical inference to set up a structure so that an indication can be given for the associated uncertainty.

## 4.2 Understanding variation

To make inferences from a sample to a population, it is necessary to have an understanding of how samples vary. This involves **probability theory**.

To see why, consider again the political polling example, and again imagine we have a number of different polls, each of size  $n = 1000$ , each randomly chosen.

Consider the following possible results from 20 polls of 1000. Three sets have been created. The data are made up, so they do not necessarily show the kind of patterns you would see in reality. In fact, one of the sets *does* show the variation you'd expect, and the other two do not. The total number of people reporting an ALP vote is plotted.

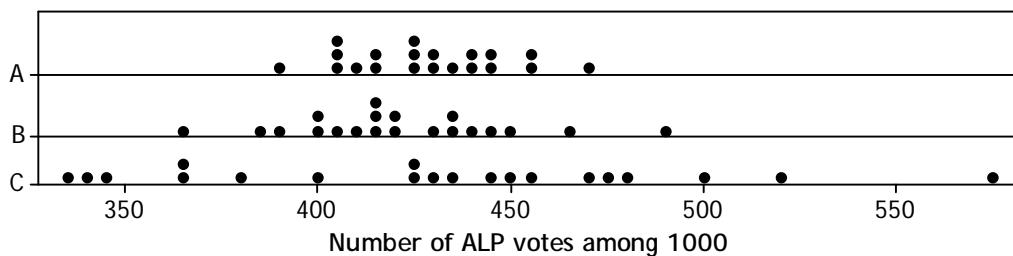


Figure 38: Sets of 20 hypothetical exit poll results from samples of  $n = 1000$

Since the number of people polled is assumed to be 1000 for each poll, a number of ALP votes of, say, 441 is readily converted to the corresponding percentage, 44.1%.

All three sets have around about the same average percentage voting ALP: the averages are 42.8%, 42.1% and 43.1% respectively, and these averages are close to the true population parameter, which was 42.9%.

But the sets are quite different in how the individual polls vary.

In set C, the largest percentage is 57.5%. Is it plausible that such a percentage could be obtained from a random sample of 1000, when the true percentage is 42.9%? How likely is this? If the 20 samples in set A show the kind of variation to expect in hypothetical repetitions of the same sampling approach, we would be more confident in claiming that the true percentage is close to 42.9%, than for set B or set C.

This line of reasoning introduces a fundamental idea of inference, which directly links it to probability. In order to make an inference about a population parameter from a single, specific sample — which is all we ever have, in practice — we must have insight into how similar samples vary, in a long run, hypothetical sequence of repeated samples. “Similar” means sampled from the same population, using the same sampling method, and with the

same sample size. This is a thought experiment: an abstraction. But it is a very important one. An understanding of the basis of statistical science involves the use of abstractions, as discussed below in Section 4.4.6.

We may have an intuitive idea about the amount of variation you would typically see. But our intuition only gets us so far. We would find it difficult to say whether set A, B or C represents the expected pattern, probably. We need this knowledge, in order to make sensible claims about our uncertainty.

In order to obtain an objective answer to this question of variation, we need a formal structure. This is the structure that forms the basis for statistical inference; it involves probability theory, random variables and distributions.

### 4.3 Probability

Ideas about probability began when gamblers wanted to understand better their chances in various games played in the 18th century in Europe. There was much development of the subject subsequently, but it was not until the 20th century that the basic formal rules were clearly established.

The *meaning* of a probability statement is quite subtle, and there are important philosophical debates about this. If we roll a six-sided die, most people would be happy to agree with the assertion that “the probability of obtaining a 4 is one in six ( $\frac{1}{6}$ )”. There are two questions worth asking about this:

- What is the meaning of this statement?
- On what basis do we assert the actual probability value  $\frac{1}{6}$ ?

Probabilities are about things that may happen, or “events”. We may label an event:  $A$  = “the 5-year survival rate in the treated group is at least 10% higher than in the control group”. We use the notation  $\Pr(A)$  to mean “the probability that  $A$  occurs”, or, more succinctly, “the probability of  $A$ ”. So for the die-rolling example, defining  $B$  to be “a 4 is obtained”, we write  $\Pr(B) = \frac{1}{6}$ .

There are three fundamental axioms of probability: rules that are assumed for the whole of the probability structure to work.

- (a) For any event  $A$ ,  $0 \leq \Pr(A) \leq 1$ . This is a convention really, arrived at by analogy with proportions, which must be between 0 and 1. And just as proportions are sometimes expressed on a percentage scale, we may do the same thing with probabilities: “there’s a 20% chance that it will rain tomorrow”. Provided it is clear from the context which convention is being used, there is no great harm in this, but it can lead to confusion, particularly if the probability is less than 0.01, and hence can look legitimate on either scale. If the percentage scale is used the percentage sign should always be retained.

If  $\Pr(A) = 0$  then  $A$  cannot occur, and if  $\Pr(B) = 1$  it must occur. When rolling a fair six sided die, define  $A$  = “a 7 is obtained” and  $B$  = “the number obtained is greater than 0 and less than 7”. Then  $\Pr(A) = 0$  and  $\Pr(B) = 1$ .

- (b) If we consider every possible distinct outcome that can occur, the total of their probabilities equals 1. This is saying that, among the complete scope of possible outcomes we are considering, something has to happen.
- (c) If two events are mutually exclusive — they can't both occur — then the probability of either of them occurring is the sum of the two individual probabilities. If  $A$  and  $B$  are mutually exclusive then  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ .

These rules are really just formal versions of propositions that are intuitively clear for proportions.

All of formal probability theory flows from these three axioms. There are two useful rules, in particular, that can be derived:

- 4. The probability that an event does not occur equals 1 minus the probability that it does occur. In notation:  $\Pr(\text{not } A) = 1 - \Pr(A)$ . This obvious rule is deceptively useful. For example, if we want to work out the chance of at least one “6” in ten rolls of a fair die, working it out directly looks like an overwhelming task. However, the chance of at least one “6” equals 1 minus the chance of no sixes at all, and the latter probability is easily found (see below).
- 5. If  $A$  and  $B$  are independent, the chance that they both occur is equal to the product of their individual probabilities. For independent events, knowing whether or not one occurs does not change our assessment of the probability of the other event. In notation: if  $A$  and  $B$  are independent,  $\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$ . We usually think about independence from first principles. Independence between outcomes is frequently assumed in data sets. For example, in a sample of humans or animals, we are likely to assume that the result of a measurement taken from one of the subjects in the sample is independent of the measurements on the other subjects. A good example of dependent events is the weather. If today's maximum temperature in Melbourne is 43 degrees, that changes the assessment of the probability of tomorrow's temperature being, say, over 40 degrees. In statistical terms, today's and tomorrow's maximum temperatures are not independent.

In simple die rolling and coin tossing experiments, it is generally reasonable to assume that the outcomes from successive rolls or tosses are independent. For that reason, we calculate the chance of no sixes

in 10 successive rolls of a die as  $\frac{5}{6} \times \frac{5}{6} \times \dots \times \frac{5}{6} = (\frac{5}{6})^{10} = 0.1615$ . Hence (by the previous rule) the probability of at least one six in 10 rolls is  $1 - 0.1615 = 0.8385$ .

## 4.4 Random variables and distributions

A **random variable** is a numerical outcome of a random procedure. “Random”, in this usage, does not mean haphazard or chaotic, but simply uncertain: before we make an observation of the random phenomenon or process, we do not know what its value will be. A random variable might be a count, or a measure on a continuous scale, or a binary (zero-one) variable, or a proportion, or a mean, or .... In a research project anything we measure in the “sample” (in the experiment or the survey we are carrying out, for example) is a random variable because we choose the sample using a random mechanism; in this sense it is a *random variable*. If we took a different sample the thing we measure would turn out to have a different value; in this sense what we measure is a *random variable*.

In a long run of repeated samples, the value of the random variable is thought to follow some rule of probability, which may be described by some mathematical relationship. This defines the **distribution** of the random variable. The notion of a distribution is quite a deep one. We have already seen many distributions of data in Chapter 2; these are empirical distributions, constructed from observed data. What we are now considering are theoretical distributions for random variables. The connection between the two is a reminder of the reciprocal nature of probability and inference and is captured well in the following diagram.

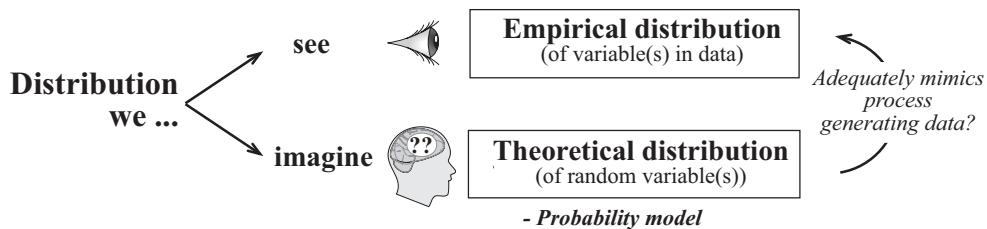


Figure 39: *Empirical versus theoretical distributions. Figure 3 in Wild C. "The concept of distribution."* Statistics Education Research Journal 2006; 5:10-25.

There are two types of random variables, **discrete** and **continuous**.

### 4.4.1 Discrete random variables

**Discrete** random variables are ones which can only take some values; al-

most always, they are based on counts of some sort. The word “discrete” is used here to mean “separate, distinct”. The number of children in a family is an example of a discrete random variable. The distribution of a discrete random variable can be defined by specifying somehow the probabilities corresponding to each possible value that the random variable may take.

The probabilities in the distribution of a discrete random variable must be all non-negative, and they must add to 1.

A specific example of the distribution of a discrete random variable is shown in Figure 40. The height of the spike at an  $x$  value shows the probability of observing that value. For example, we see that the probability that this random variable takes the value 10 is about 0.15.

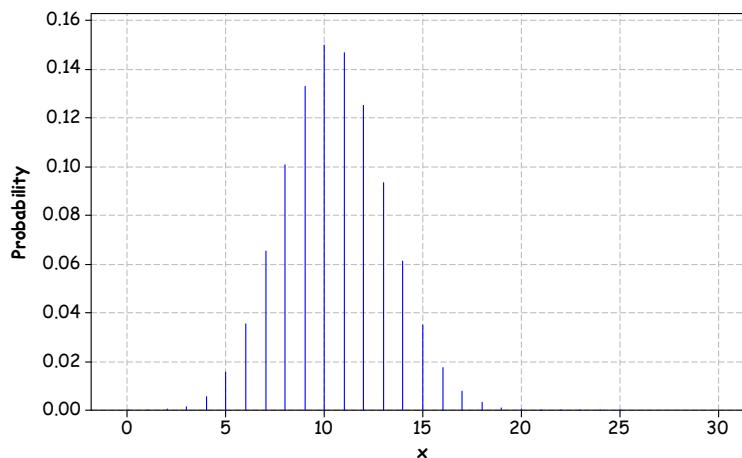


Figure 40: A discrete distribution, showing the probabilities for each outcome.

There are quite a few discrete distributions, corresponding to different sorts of discrete random variables. Here we look at two discrete distributions, the integer distribution and the binomial distribution.

The integer distribution has a simple, clear form, which helps to consolidate the essential features of a discrete distribution. It is also used to introduce some of the more general features of discrete random variables, and how you examine them in MINITAB.

#### ▷ EXAMPLE. (1) Integer

The number obtained when throwing a fair die is a discrete random variable. If the die is fair,

$$\Pr(\text{the number is } x) = \frac{1}{6}, \text{ for } x = 1, 2, 3, 4, 5, 6.$$

Here  $\Pr(\text{the number is } x)$  denotes the probability that the number is  $x$ , and  $x$  can be 1, 2, 3, 4, 5 or 6.

In general, capital letters near the end of the alphabet are used for random variables, and if only one is considered,  $X$  is the letter commonly used.

So we could define, here,  $X$  = number obtained when throwing a fair die. Using this notation, we define  $X = x$  to mean “the random variable  $X$  takes the value  $x$ ”. In this formulation,  $x$  is just a number, such as 4, or 2.3, whereas  $X$  is a random variable. So “ $X = 4$ ” means “ $X$  takes the value 4”, or, equivalently, “4 is the result when  $X$  is observed”.

If you are strongly mathematical, you may wonder about the nature of a random variable. Formally, a random variable is a set function that maps sets in the sample space to the real line.

Using this notation, we can write the distribution for this random variable  $X$  in the standard form:

$$\Pr(X = x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6.$$

This simple distribution is called the **integer distribution** (by MINITAB).

Remember that the probability distribution of a discrete random variable tells us the probabilities for each distinct value that the random variable can take. This may be done by a table, a list, a formula, or even — approximately — by a graph. Can you draw the integer distribution?

To get the complete probability distribution for a discrete random variable from MINITAB you first need to create a column which contains all the values that the random variable can take. For the die, the possible values are 1, 2, ..., 6. With these values in a MINITAB column, C1, say:

- Use Calc > Probability distributions ▶ Integer,
- Check the button labelled Probability,
- Enter 1 for the Minimum and 6 for the Maximum,
- Check the button labelled Input column and select C1 as the column: this specifies the values for the random variable whose probabilities you are seeking.

This will seem like an awful lot of trouble to get six probabilities which are all the same and whose value you knew since the time you first used a die. However, the procedure generalizes to other types of random variables.

The dialogue box for all **discrete** distributions, including the integer, has three choices:

- **Probability:** This gives the probability distribution. For each value of  $x$  input, either in a column or as a single constant, the output is  $\Pr(X = x)$ , or the probability that  $X$  equals  $x$ .
- **Cumulative probability:** This gives the cumulative probability distribution. For each value of  $x$  input, either in a column or as a single constant, the output is  $\Pr(X \leq x)$ , or the probability that  $X$  is less than or equal to  $x$ .
- **Inverse cumulative probability:** This is tricky for discrete random variables and is not considered in this course.

### 4.4.2 Binomial distribution

#### ▷ EXAMPLE. (2) Binomial

The number of heads obtained in tossing a fair coin a number of times is another example of a discrete random variable. The distribution of this random variable can be described by the **binomial distribution**. The binomial distribution is useful for modelling data with a binary outcome, where each unit is independent and can only take one of two values. The classic example is coin tossing, but it is useful in a wide variety of contexts.

Suppose that we are studying a population in which the proportion of units with an attribute of interest is  $\theta$ . Each unit is independent of others and either has the attribute or not. When we sample at random, the probability of selecting an unit with the attribute is  $\theta$ . Suppose we have a random sample of  $n$  units and  $n$  is much smaller than the population size; the effect of this is that we can ignore any impact of removing units from the population as we sample. Let  $X$  denote the number of units in the sample with the attribute. Then  $X$  is a random variable with a binomial distribution with parameters  $n$  and  $\theta$ , and we write  $X \stackrel{d}{=} \text{Bi}(n, \theta)$ .

Note the use of the symbol  $\stackrel{d}{=}$ . This symbol means: “is distributed as” or “has the following distribution”. The distribution indicated immediately to the right of the symbol is that of the random variable immediately to the left of the symbol.

Cork taint in wines was a significant problem<sup>10</sup>. Suppose we have a large population (many thousands) of bottles of wine, which we wish to test for cork taint. We take a random sample of 30 bottles and ask a panel of experts to classify the wines according to whether or not the wine has a taint. The protocol forces them to agree on a yes/no outcome for each bottle. Then the number of tainted wines among the 30 sampled has a binomial distribution, which depends on the true proportion tainted in the whole population. The distribution of the number of cork-tainted wines in the sample, if the true population proportion is 0.35, is shown in Figure 40. Figure 41 shows a variety of possible distributions for this case, as the population proportion varies.

---

<sup>10</sup>... avoided altogether by the screwtop cap, of course, which is one of the main reasons for its introduction.

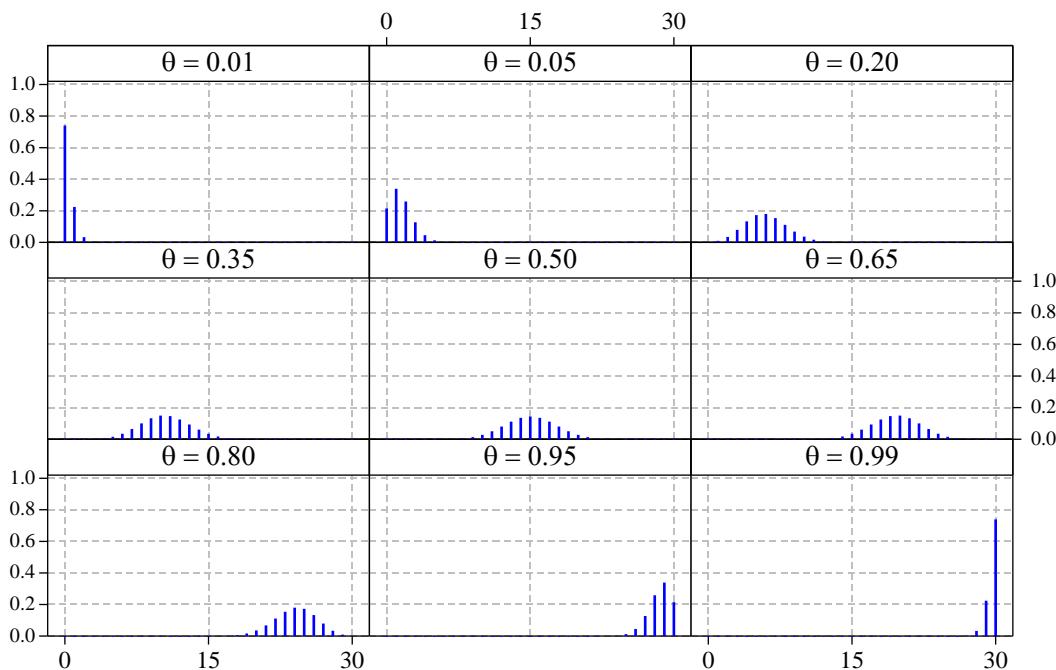


Figure 41: Binomial distributions for  $n = 30$  and a variety of values of  $\theta$ , the population proportion

It is not too hard to derive the binomial distribution from first principles. We want to determine the probability that  $X$  takes some particular value  $x$ , which we write as  $\Pr(X = x)$ , read as “the probability that the random variable  $X$  takes the value  $x$ ”. For this event to occur, there needs to be exactly  $x$  units in the sample of size  $n$  with the attribute of interest, and  $n - x$  without the attribute. One way for that to occur is for the first  $x$  units sampled to have the attribute, and for the other  $n - x$  units not to have the attribute. The chance that the first unit has the attribute is  $\theta$ ; the chance that the second one has it is also  $\theta$ , and so on. The chance that a unit does not have the attribute is  $1 - \theta$ .

Putting all this together, and remembering that the probability of independent events occurring together is the product of the individual probabilities, we get that the probability of the first  $x$  units having the attribute, and the others not, is

$$\underbrace{(\theta \times \theta \times \cdots \times \theta)}_{x \text{ of these}} \times \underbrace{(1-\theta) \times (1-\theta) \times \cdots \times (1-\theta)}_{n-x \text{ of these}} = \theta^x (1-\theta)^{n-x}.$$

But this is only one of the ways that we could end up with  $x$  units in the sample having the attribute, and  $n - x$  not having it. Each of these ways has the same probability,  $\theta^x (1-\theta)^{n-x}$ , so it's just a matter counting all the ways it can happen. This number is equal to the number of ways of choosing  $x$  objects from among  $n$ , and is given by the formula

$${}^n C_x = \frac{n!}{x!(n-x)!} = \frac{n \times (n-1) \times (n-2) \times \cdots \times (n-x+2) \times (n-x+1)}{x \times (x-1) \times (x-2) \times \cdots \times 2 \times 1}.$$

Finally, we therefore have

$$\Pr(X = x) = {}^n C_x \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \dots, n.$$

The Binomial distribution can be found using MINITAB:

Calc > Probability distributions ► Binomial.

Again, you need to put the possible values of the random variable into a column, if you want the complete probability distribution.

- **Probability:** This gives the probability distribution. For each value of  $x$  input, the output is  $\Pr(X = x)$ , or the probability that  $X$  equals  $x$ .
- **Cumulative probability:** This gives the cumulative probability distribution. For each value of  $x$  input, the output is  $\Pr(X \leq x)$ , or the probability that  $X$  is less than or equal to  $x$ .

A crucial point about cumulative probabilities for discrete random variables is that you need to be careful about the difference between  $\leq x$  and  $< x$ , and also  $\geq x$  and  $> x$ . If the discrete random variable  $X$  takes integer values, then  $X < 14$  is the same as the event  $X \leq 13$ . This is not a trivial point: it can make a large difference to the probabilities. It means that questions concerning probabilities for discrete random variables need careful reading of the words used.

If we are asked: “What is the chance that the outcome of the die roll is more than 3?”, then that corresponds to observing a 4, a 5 or a 6. That is,  $X$  more than 3 is equivalent to  $X \geq 4$ . On the other hand, the chance of getting an outcome of at least 3 is equivalent to getting a 3, 4, 5 or 6, or  $X \geq 3$ .

We are all familiar with the occasional importance of this in ordinary life. “The fine must be paid within 14 days.” What does that mean? We want to be clear whether paying it in exactly 14 days’ time is OK, or not.

#### 4.4.3 Mean, variance and standard deviation

The **mean** of a random variable  $X$ , which we denote by  $\mu$  or  $E(X)$ , is the weighted average of values that  $X$  can take, where the weights are provided by the distribution of  $X$ . It is at the “centre of gravity” of the distribution. Sometimes the term the “expectation of  $X$ ” is used, which is where the notation  $E(X)$  originates ( $E$  for Expectation).

However, the use of this term does not imply that we “expect”  $X$  to take this value; in fact  $E(X)$  is often an impossible value for  $X$  to take. The mean result for a roll of a die is 3.5; this is not a value that we can observe. Usually though, we expect  $X$  to be “around” the mean. The idea of a mean in this sense was first developed in gambling contexts in the 18th century, as a way of assessing whether a game was fair or not. A commercial gambling operation (horse betting, casino games, Tattslotto etc.) aims to have a slightly negative mean winning amount, that is, a small average losing

amount, from the gambler's perspective; this is how the company ensures a profit.

The **variance** of a random variable  $X$ , which we denote by  $\sigma^2$  or  $\text{var}(X)$ , is the weighted average of squared deviations from the mean of  $X$ , where the weights are provided by the distribution of  $X$ .

The mean and the variance for a random variable have parallels with the sample mean and variance that we have already met. The difference is that for the sample mean and variance each observation is given the same weight in the averaging. On the other hand, for the mean and variance of a random variable we need to take into account the likelihood of each value in the averaging. So you can think of these as weighted averages, with more weight given to more likely outcomes and less weight to outcomes with small probability.

The **standard deviation** of a random variable  $X$  is the square root of the variance and is denoted by  $\text{sd}(X)$ : i.e.  $\text{sd}(X) = \sqrt{\text{var}(X)}$ .

Both types of random variables, discrete and continuous, have means, variances and standard deviations.

We now consider the means, variances and standard deviations of the two types of discrete random variables we have so far considered.

#### ▷ EXAMPLE. (1) Integer

Consider, again, the number obtained in throwing a die is a discrete random variable. It takes the values  $1, 2, \dots, 6$  with equal probabilities.

The mean of this random variable is 3.5 (obviously, because of symmetry) and the variance is 2.92 (not obviously). Hence the standard deviation is 1.71 ( $= \sqrt{2.92}$ ).

#### ▷ EXAMPLE. (2) Binomial

If  $X \stackrel{d}{=} \text{Bi}(n, \theta)$ , then the mean of  $X$  is  $n\theta$  and the variance is  $n\theta(1-\theta)$ . The standard deviation is equal to  $\sqrt{n\theta(1-\theta)}$ .

The value for the mean is intuitively compelling. If a proportion  $\theta$  of the population have a characteristic of interest, how many will have the characteristic, on average, in a sample of size  $n$ ? Intuitively, we just apply the population proportion to the sample and get the answer  $n\theta$ .

The formula for the variance of  $X$ ,  $n\theta(1-\theta)$ , is not obvious at all, but is worth thinking about. It implies that, for fixed  $n$ , binomial distributions are more spread out when  $\theta$  is close to  $\frac{1}{2}$  than when  $\theta$  is close to zero or one. This has direct implications for survey work, because it means that the estimation of a population proportion is least precise when  $\theta$  is near  $\frac{1}{2}$ . This is somewhat unfortunate for political polling in two-party systems, for example, when proportions of crucial interest are often near  $\frac{1}{2}$ .

You can see the effect of the formula, roughly, in Figure 41. The distribution

has the greatest spread when  $\theta = 0.5$ , and the least spread when  $\theta$  is close to zero or one. Further, there is a symmetry about the distributions for  $\theta$  and  $(1-\theta)$ , for example,  $\theta = 0.35$  and  $\theta = 0.65$ . This makes sense: we are modelling a binary characteristic, so we might just have well considered *not* having the characteristic. If we had, the underlying structure would be the same, just a reflection of the model for having the characteristic.

We are now able to answer the question raised by Figure 38. The binomial distribution is the appropriate model for a simple random sample, which is what is often assumed for political polling. In Figure 38, for a simple random sample of size 1000, set A shows the kind of variation that you would see in different samples of the same size; it is based on binomial variation. Sets B and C have unrealistically large amounts of variation, for samples of size 1000.

#### 4.4.4 Continuous random variables

A continuous random variable can take any value within the range of possible values. The distribution of a continuous random variable is defined by specifying a curve which relates the height of the curve at any particular value to the chance of an observation close to that value. This curve is called the **probability density function**.

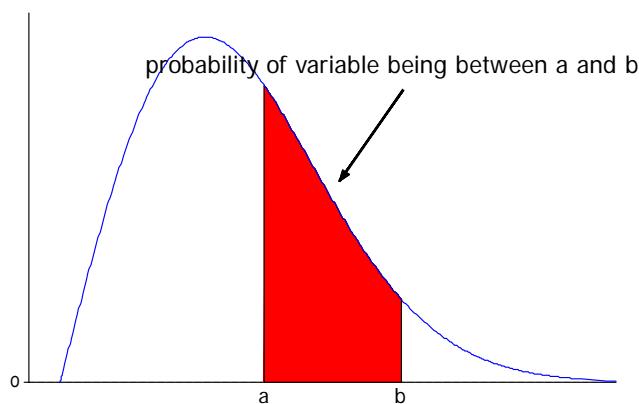


Figure 42: An arbitrary probability density function; the probability of the corresponding random variable being between  $a$  and  $b$  is equal to the size of the area shaded.

Formally, the chance that a continuous random variable takes a value in an interval between two points  $a$  and  $b$  is the **area under the curve** between  $a$  and  $b$ , as shown in Figure 42. Example 3 below is a continuous random variable.

Why can't we use the discrete random variable approach for a continuous random variable? We may ask about the probability that a continuous random variable takes the value 12. But ... what do we mean by that? Remem-

ber that it can take any value in a given range, so it can be 11.9, or 12.2 etc. A reasonable way of giving an answer to the probability required is to suggest that what is meant by “12” in this case is “12, to the nearest whole number”. This means a number between 11.5 and 12.5; and now we are talking about an interval again: quite a narrow interval, perhaps, but an interval all the same. If we insist that we want the probability that a continuous random variable takes the value 12 *exactly*, that is, 12.00000..., then this is equal to zero.

The probability density function must be non-negative (or else we could get some negative probabilities), and the total area under its graph must be 1 (because the corresponding random variable must be somewhere between  $-\infty$  and  $\infty$ ).

#### 4.4.5 Normal distribution

There are many continuous distributions, corresponding to different types of continuous random variables. But the most important continuous distribution is the so-called Normal distribution.

Later on we will meet a number of other continuous distributions, including the  $\chi^2$ ,  $t$  and  $F$  distributions. It is one reflection of the importance of the Normal distribution that each of these distributions depends on the Normal distribution.

The Normal distribution is the third specific distribution we have looked at so far.

##### ▷ EXAMPLE. (3) Normal

The height of individuals in a population give a continuous random variable. For a particular population, it may be that height can be described by the **Normal distribution**, which has a shape as given in Figure 43. A Normal distribution is symmetric about its mean  $\mu$  and its variance is usually denoted by  $\sigma^2$ . A way to interpret  $\sigma$  in the Normal distribution’s shape is that it is equal to the distance between  $\mu$  and the points (one on either side of  $\mu$ ) where the curve changes from convex to concave.

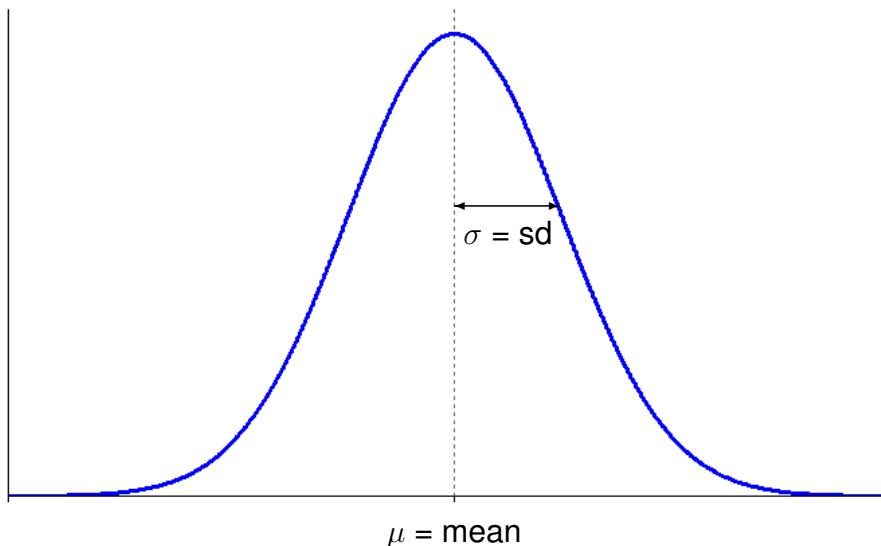


Figure 43: A Normal distribution with mean =  $\mu$  and standard deviation (sd) =  $\sigma$ .

In carrying out calculations based on the Normal distribution it is always useful to sketch the distribution and write in the actual mean and standard deviation.

If  $X$  has a Normal distribution with mean  $\mu$  and variance  $\sigma^2$  then we write  $X \stackrel{d}{=} N(\mu, \sigma^2)$ . This distribution is useful in this course not only because some variables are approximately Normally distributed but also, and more importantly, because estimators of parameters are often approximately Normally distributed, a remarkable fact used extensively in inference. We look at the archetypal case of this, the Central Limit Theorem, in Chapter 4.

Probabilities for the Normal distribution can be found using MINITAB: Calc > Probability distributions ▶ Normal.

The dialogue box for a continuous distribution, including the Normal, has three choices:

- **Probability density:** This is the height of the probability density curve, for a given input value. This is not generally important from the point of view of statistical inference. It is not a probability.
- **Cumulative probability:** This gives the cumulative probability distribution. For each value of  $x$  input, either in a column or as a single constant, the output is  $\Pr(X \leq x)$ , or the probability that  $X$  is less than or equal to  $x$ . It is the area under the curve to the left of  $x$ .
- **Inverse cumulative probability:** This is the inverse, or reverse calculation from the previous one. For each value of  $p$  input, the output is the value of  $x$  that satisfies  $\Pr(X \leq x) = p$ .

For any continuous random variable  $X$ , the **cumulative probability** for a value  $x$  is the probability that  $X$  is less than or equal to  $x$ ; i.e.,  $\Pr(X \leq x)$ .

We can use this to find the probability that  $X$  lies in any interval for a continuous random variable. For example, suppose  $X \stackrel{d}{=} N(20, 4^2)$ , and we want to know the chance that  $X$  is between 22 and 24. Then we can work this out by subtraction: the required probability is

$$\begin{aligned}\Pr(22 \leq X \leq 24) &= \Pr(X \leq 24) - \Pr(X \leq 22) \\ &= 0.8413 - 0.6915 \\ &= 0.1498\end{aligned}$$

This is also shown graphically in Figure 44. MINITAB has a useful graphical facility to get these probabilities for the Normal distribution (and for many other distribution); see Graph > Probability Distribution Plot ▶ View Probability.

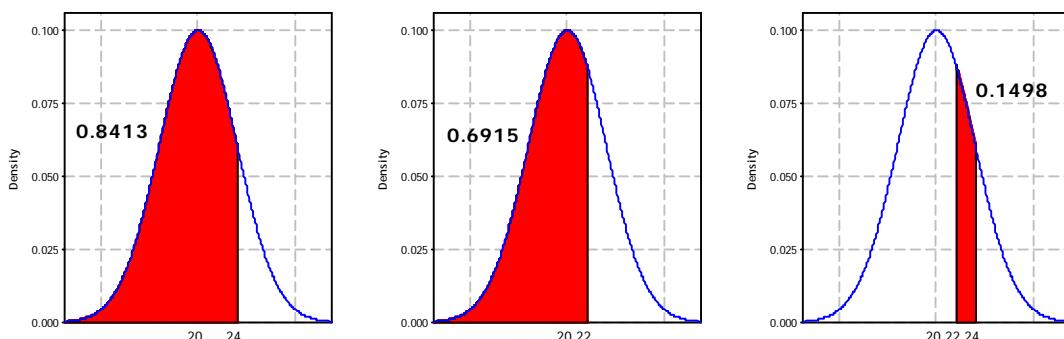


Figure 44: *Obtaining a probability for a Normal random variable, example in text.*

If we were to use this approach to obtain the probability that a continuous random variable  $X$  takes the value 22 exactly,  $\Pr(X = 22)$  is equal to  $\Pr(22 \leq X \leq 22) = \text{area under the curve between } 22 \text{ and } 22$ , i.e. zero, confirming the point made previously.

If we were considering a discrete random variable we would need to be careful about the inequalities. In fact, for a discrete random variable which only takes integer values:

$$\Pr(22 \leq X \leq 24) = \Pr(X = 22, 23 \text{ or } 24) = \Pr(X \leq 24) - \Pr(X \leq 21).$$

This is another reminder that continuous and discrete random variables need different treatment.

The **inverse cumulative probability**, as the name suggests, does the calculation in reverse. One calculation goes in one direction (like feet to metres) and the other goes back in the opposite direction (like metres to feet), but the underlying calculation is the same.

Look at the left-hand panel in Figure 44. This shows the Normal distribution with mean 20 and standard deviation 4. The shaded area is the *cumulative probability* for 24, and is equal to 0.8413. That is, the probability that this random variable is less than or equal to 24 is 0.8413. In symbols:  $\Pr(X \leq 24) = 0.8413$ . It can occur that we are interested in looking at the same probability “in reverse”: this is the *inverse cumulative probability*. Hence we can ask: for what value of  $x$  is the cumulative probability equal to 0.8413? The answer is “ $x = 24$ ”.

For any probability  $p$ , the inverse cumulative probability is the number  $x$  such that the random variable has a probability  $p$  of being less than or equal to  $x$ . So if  $X \stackrel{d}{=} N(20, 4^2)$ , and we want the inverse cumulative probability for 0.10, we find that the answer is 14.8738. That is,  $\Pr(X \leq 14.8738) = 0.1$ .

The **standard Normal distribution** (so-called) has mean 0 and variance 1, and usually is denoted by  $Z$ :  $Z \stackrel{d}{=} N(0, 1)$ , and it is special, in the following sense.

Any Normal distribution can be “standardized”: if  $X \stackrel{d}{=} N(\mu, \sigma^2)$  then it turns out that  $Z = \frac{X-\mu}{\sigma} \stackrel{d}{=} N(0, 1)$ . It is quite easy to show that the mean of  $Z$  is equal to 0 and that the variance of  $Z$  is equal to 1; it is less obvious, but true, that the distribution of  $Z$  is Normal.

In the example,

$$\Pr(X \leq 24) = \Pr\left(\frac{X - 20}{4} \leq \frac{24 - 20}{4}\right) = \Pr(Z \leq 1) = 0.8413.$$

This means that, in an important sense, there is really only one underlying Normal distribution: probabilities for any Normal distribution can be readily converted into a problem about probabilities for the standard Normal distribution. This is not the case for any distribution; it is not the case for the binomial distribution, for example. This explains why, if you have looked at a book of statistical tables, you find tables for the standard Normal distribution only; by contrast, there are different tables provided for various  $n$  and  $\theta$  for the binomial distribution.

Look at Figure 45 and guess the probabilities:

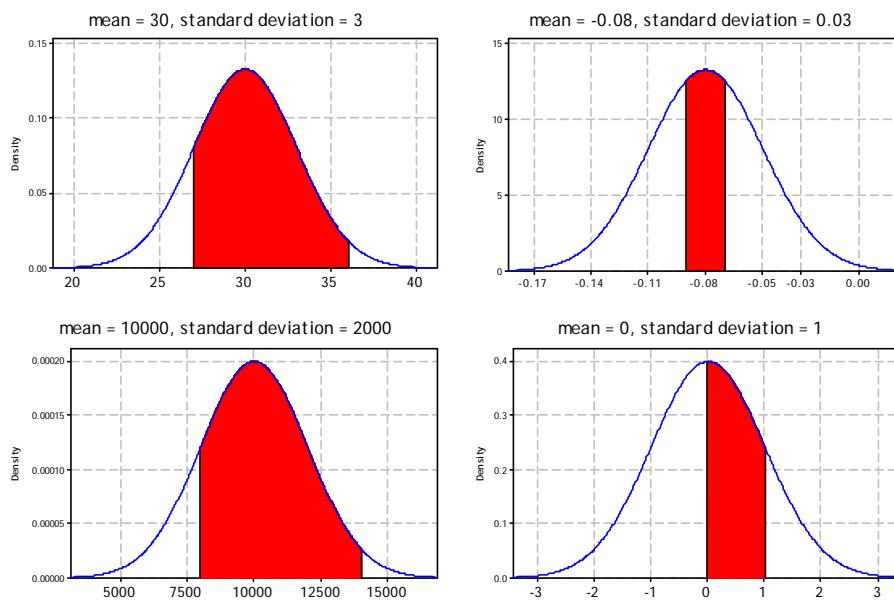


Figure 45: Four different Normal distributions: guess the probabilities.

Remember that the total area under each probability density function is equal to one (the random variable has to be somewhere). You should be able to guess the probabilities, roughly, even though the means and standard deviations vary a lot; the answers are below<sup>11</sup>. Note that the two left hand probabilities are the same, because in each case we are considering the interval between a point one standard deviation below the mean, and another point two standard deviations above the mean.

There is one more important feature of the Normal distribution to note. If two or more Normally distributed random variables are added together, the resulting random variable is itself Normally distributed. Again, this is not a property that applies in general, to any type of distribution.

#### 4.4.6 Abstractions

Notice the word “imagine” in Figure 39. In understanding the theory and practice of statistics, it is necessary to deal with abstractions of various kinds. Ironically, often these abstractions represent what we believe or hope is reality; but we cannot observe it directly. There are many words and phrases used in these notes that entail this notion of abstraction. We have already met population parameters. We think of these as fixed but unknown constants. To the extent that they are unknown, they are abstract; we usually can’t identify them. But we are vitally interested in their values: we make inferences about them.

<sup>11</sup>Reading across the rows, the probabilities are 0.82, 0.26, 0.82 and 0.34.

Models and distributions are abstract. A problem might ask you to assume that the random variable  $X$  has a particular distribution. This is because inference is only possible in a framework that has some understanding of what random process generated the data. If we want to make an inference about an unknown population proportion, then we know how to quantify the uncertainty if the sample has been generated from a binomial model. Of course models, and abstractions more generally, may or may not be true. So for a particular data set, we always need to ask ourselves, at least implicitly: how reasonable is the model? and, more subtly: how wrong will my inference be if the model is not reasonable? But we can get nowhere without assuming something abstract about the underlying probability structure.

Another example of a substantial abstraction is the hypothetical endless repetition of the same study, under identical conditions. We indulge in this thought experiment when we interpret the meaning of a probability, and specifically the meaning of the “95%” in a 95% confidence interval.

So it is useful to consciously allow your mind to entertain various abstract concepts and scenarios: it genuinely helps with understanding. These abstractions are introduced in many ways, sometimes by a very simple word or phrase. You may be asked to “assume that”, or “suppose”, or “model the data as ...”. Perhaps the simple word “if” may be used, often in an “if ... then” construction, e.g. “If the data are binomially distributed with parameters  $n$  and  $\theta$ , then ...”. To remind you of this process, we occasionally use the symbol  $(\circlearrowright)$ , to indicate this thought process: it’s all in the mind!

#### 4.4.7 An important notational convention

There is a key notational feature to which we (attempt to) adhere, because it is a useful reminder of the structure we adopt. We use capital letters ( $X, Y, \dots$ ) to denote random variables, and lower case letters ( $x, y \dots$ ) to denote the **observed values**, or **realizations** of the random variable. In this way we can say that  $X$  has a distribution; but  $x$ , being an observed value in a particular case, is just a number and therefore cannot have a distribution.

This means that when we speak of a random sample “on”  $X$ , we mean that the observed values in the sample come from  $X$ ’s distribution. If we know what the distribution of  $X$  is, we can say something about the kinds of samples that we expect to see. A lot of statistical ideas in inference are based on this idea.

Another useful notational convention that we use in these notes is that a Greek letter is generally used for a population parameter (e.g.  $\mu, \sigma, \theta, \dots$ ).

## 4.5 Some important results

There are some basic elements of statistical theory which we use repeatedly. It is possible to use statistics a lot without understanding these results, but in a serious introductory course they need to be tackled.

A core notion in statistical science is that of *averaging*. It is rather startling how much inference reduces to asking questions about averages. There are deep reasons why this is so, but at a simple level, the idea is that the general position of a distribution of a random variable can be captured by an average. In more subtle contexts, effects of interest can be seen as depending on averaging process.

This means that understanding the properties of averages of random variables is fundamental to doing inference. When we begin to think about this, there are two operations on random variables that we need to consider:

- Rescaling a random variable;
- Summing random variables.

We therefore now look at these operations and learn about the distributional properties.

### 4.5.1 Rescaling a random variable

If we multiply a random variable by a factor  $k$ , then:

- the mean changes by the multiplicative factor of  $k$ ;
- the standard deviation changes by the multiplicative factor of  $|k|$ ;
- the variance changes by the multiplicative factor of  $k^2$ .

In notation: if  $E(X) = \mu$ ,  $\text{sd}(X) = \sigma$  and  $\text{var}(X) = \sigma^2$ , then:

$$\begin{aligned} E(kX) &= k\mu, \\ \text{sd}(kX) &= |k|\sigma, \\ \text{var}(kX) &= k^2\sigma^2. \end{aligned}$$

Note that the natural and important way to think about the standard deviation result is for positive  $k$ : if  $k > 0$  then  $\text{sd}(kX) = k\sigma$ .

The application of these results is *re-scaling* a random variable. For example, we may measure  $X$  in days. If we wish to convert to hours, then we are multiplying by 24, and the above results apply with  $k = 24$ .

These results are common sense. The mean and the standard deviation are in the same units as  $X$ . The result is saying if we multiply the random variable by  $k > 0$ , the resulting random variable's mean and standard deviation are  $k$  times greater than those of  $X$ .

#### 4.5.2 Sums and differences of random variables

Suppose the random variables  $X$  and  $Y$  have means  $\mu_X$  and  $\mu_Y$ , and variances  $\sigma_X^2$  and  $\sigma_Y^2$ . Define  $T = X + Y$  and  $D = X - Y$ . Then

$$\mathbb{E}(T) = \mu_T = \mu_X + \mu_Y$$

$$\mathbb{E}(D) = \mu_D = \mu_X - \mu_Y$$

This is another common sense result: it says that when you add or subtract random variables, their means behave in the same way. It extends to any number of random variables.

The result is impressive in that there are no conditions to worry about; the result applies regardless of whether the random variables have the same distribution or not, and they do not have to be independent. The result holds completely generally.

The corresponding result for the variance of a sum depends on the notion of the independence of two (or more) random variables. We say that the random variables  $X$  and  $Y$  are independent if the distribution of  $X$ , given  $Y$ , is the same as the distribution of  $X$  without knowledge of  $Y$  (and vice versa).

*If  $X$  and  $Y$  are independent, then*

$$\text{var}(T) = \sigma_T^2 = \sigma_X^2 + \sigma_Y^2$$

$$\text{var}(D) = \sigma_D^2 = \sigma_X^2 + \sigma_Y^2$$

Note that the variances are “additive” and not the standard deviations. In fact, when  $X$  and  $Y$  are independent,  $\text{sd}(T) = \sqrt{\sigma_X^2 + \sigma_Y^2}$  and not  $\sigma_X + \sigma_Y$ .

When  $X$  and  $Y$  are independent, why do variances *add* when we consider the variance of the *difference* between them?

One way to see this is by analogy with measurement error. If you measured the height and width of a standard sheet of A4 paper, you'd expect there to be some measurement error in both measurements. If you add the two measurements together, the measurement error in the total will, on average, be worse (larger): sometimes you'll make two errors in the same direction and they will combine.

However, if you subtract the width measurement from the height measurement, the same thing will apply.

We can also derive the result formally. If  $X$  and  $Y$  are independent, then

$$\begin{aligned}\text{var}(D) &= \text{var}(X - Y) = \text{var}(X + (-Y)) \\ &= \text{var}(X) + \text{var}(-Y) \quad \text{because } X \text{ and } -Y \text{ are independent} \\ &= \text{var}(X) + \text{var}(Y) \quad \text{because } \text{var}(-Y) = \text{var}(Y).\end{aligned}$$

The above results for the means apply regardless of whether  $X$  and  $Y$  are independent, but the results for the variances apply if  $X$  and  $Y$  are independent.

These results apply to sums of random variables, which is quite a different thing to the rescaling considered in section 4.5.1.

### Sums and differences of Normal random variables

Recall that the addition (or subtraction) of Normally distributed random variables results in another Normally distributed random variable. So given the above, this means that if  $X \stackrel{d}{=} N(\mu_X, \sigma_X^2)$ , and  $Y \stackrel{d}{=} N(\mu_Y, \sigma_Y^2)$ , and  $X$  and  $Y$  are independent, then

$$\begin{aligned}X + Y &\stackrel{d}{=} N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2), \text{ and} \\ X - Y &\stackrel{d}{=} N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2).\end{aligned}$$

When we add and subtract Normally distributed random variables, the result is also Normally distributed. For many other distributions, this is not true.

#### 4.5.3 Mean and variance of $\bar{X}$ , the sample mean

We have been working towards considering the distribution of the sample mean. In order to do this, we need to define, formally, a random sample.

A random sample “on  $X$ ” of size  $n$  can be regarded as  $n$  random variables  $X_1, X_2, \dots, X_n$ , which

- are independent;
- have the same distribution as  $X$ .

The idea here is that there is an underlying parent distribution, the distribution of  $X$ . Each member of the random sample,  $X_i$ , comes from this distribution: the observation we get is governed by the pattern of the distribution of  $X$ .

Each individual member of the random sample has the same distribution as  $X$ . This means that if  $E(X) = \mu$  then  $E(X_i) = \mu$ ; and if  $\text{var}(X) = \sigma^2$  then  $\text{var}(X_i) = \sigma^2$ .

Consider the sample mean,  $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ .

The expectation of  $\bar{X}$  is given by:

$$\begin{aligned}\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n} \mathrm{E}(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n}(\mu + \mu + \dots + \mu) \\ &= \frac{1}{n}(n\mu) \\ &= \mu\end{aligned}$$

In words: the mean of the sample mean is equal to the mean of the distribution from which the random sample came.

In a repeated sampling sense, if we take many many samples of size  $n$ , the histogram of the sample means is centred around the mean of the population from which we have sampled. (◊)

This is a good thing: it says that the sample mean is “unbiased” as an estimator of  $\mu$ . When we have a particular random sample and an observed  $\bar{x}$ , sometimes it will above  $\mu$  and sometimes below. But the long run average is  $\mu$ .

Now consider the variance of  $\bar{X}$ .

$$\begin{aligned}\mathrm{var}(\bar{X}) &= \mathrm{var}\left(\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right) \\ &= \frac{1}{n^2} \mathrm{var}(X_1 + X_2 + \dots + X_n) \\ &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) \\ &\quad (\text{because the } X_i\text{'s are independent}) \\ &= \frac{1}{n^2}(n\sigma^2) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

In words: the variance of the sample mean is equal to the variance of the distribution from which the random sample came, divided by the sample size.

This means that in larger random samples,  $\bar{X}$  will tend to be closer to  $\mu$  than in smaller samples; the spread of the distribution of  $\bar{X}$  is inversely proportional to the sample size,  $n$ .

It follows that the standard deviation of  $\bar{X}$  is  $\frac{\sigma}{\sqrt{n}}$ .

The key results arising from this theory are that for a random sample of size  $n$  on  $X$ , that is,  $X_1, X_2, \dots, X_n$ , independent, with  $E(X_i) = E(X) = \mu$  and  $\text{var}(X_i) = \text{var}(X) = \sigma^2$ ,

- $E(\bar{X}) = \mu$ : the distribution of  $\bar{X}$  is centred around  $\mu$ ;
- $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$ : the variance of the distribution of  $\bar{X}$  is proportional to  $\text{var}(X) = \sigma^2$ , and inversely proportional to  $n$ , the sample size.

These two results apply regardless of the shape of the distribution of  $X$ .

There are two further points to make in this context.

### Standard error of the mean

The standard deviation of  $\bar{X}$  is  $\frac{\sigma}{\sqrt{n}}$ .

When we don't know the value of  $\sigma$ , we estimate it by the sample standard deviation,  $s$ . Thus an estimate of  $\text{sd}(\bar{X})$  is  $\frac{s}{\sqrt{n}}$ .

This is called the **standard error** of the mean:

$$\text{se}(\bar{X}) = \frac{s}{\sqrt{n}}.$$

In words: the standard error of the mean is an estimate of the standard deviation of the sample mean.

This is shown in the standard output of the MINITAB descriptive summary.

Note that if we knew  $\sigma$  we would prefer to use  $\frac{\sigma}{\sqrt{n}}$  for the standard deviation of  $\bar{X}$ . But in general the standard deviation  $\sigma$  will be unknown.

### Distribution of $\bar{X}$ when $X$ is Normal

The results for the mean and variance of  $\bar{X}$  that we have derived did not specify any distribution in particular.

If we make the further and specific assumption that the distribution from which we are sampling is *itself* normally distributed, that is,  $X \stackrel{d}{=} N(\mu, \sigma^2)$ , then, in addition to the two results for the mean and the variance of  $\bar{X}$ , it follows that  $\bar{X}$  is also normally distributed. This result is true for any value of  $n$ .

In other words, for a random sample of size  $n$  on  $X \stackrel{d}{=} N(\mu, \sigma^2)$ ,  $\bar{X} \stackrel{d}{=} N(\mu, \frac{\sigma^2}{n})$ .

## 4.6 Other inference perspectives

Here we make some final observations to round out chapter 4.

We have introduced the thinking about inference in terms of samples and populations. This is because it is the easiest way to see the essential ideas of inference. However, not all scientific inferences fit neatly into that framework. Most obviously, we sometimes have situations in which it is more natural to think of an underlying mathematical model that is a data generating mechanism. The model has unknown parameters of interest. We seek to estimate these parameters using data. A good example of this which even attracted media interest during the COVID-19 pandemic is the quantity known as  $R_0$ , defined to be the average number of people who are infected by someone who is infected with COVID-19. This parameter plays a vital role in the mathematical model for transmission of an infectious disease.

Although this is a different perspective, it remains true that we need to use all of the elements of probability that we have looked at here, to make inferences. We (still) have random variables and their distributions, means, variances and standard deviations, unknown parameters that will be estimated with imprecision, and so on. Hence the material in chapter 4 is generally applicable, regardless of how we think about inference.

Finally: at a fundamental level, there is more than one approach to statistical inference. The most important types are what are called “frequentist” and “Bayesian”. It is beyond the scope of this subject to cover this in any detail; the simplest way to characterise the distinction is to say that in frequentist inference, parameters are fixed but unknown, whereas in Bayesian inference, parameters themselves have a distribution.

In this subject, a frequentist approach is used. If you continue with more education in statistics and encounter the Bayesian approach, you will find that there, too, we need to use the general terms and ‘advanced language’ that we have covered in this chapter.

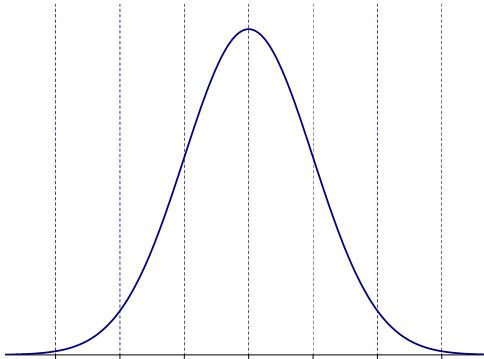
## 4.7 Exercises

4.1 *The point of this problem is to think about probability and probability distributions, in particular, the Binomial distribution. As you work through it, think carefully about the meaning of the probabilities.*

- (a) Suppose that the probability that it will rain tomorrow is 0.3. (°°)  
What is the probability that it will not rain tomorrow?
- (b) Suppose that 6.3% of families with four children consist of four boys. (°°)  
What is the probability that a family of four children has at least one girl?
- (c) Let  $Y$  be the number of students in a class who have an inadequate diet. Suppose that the probability that  $Y$  is less than or equal to 2 is 0.91. That is,  $\Pr(Y \leq 2) = 0.91$ . (°°) What is the probability that at least three students in the class have an inadequate diet? That is, what is  $\Pr(Y \geq 3)$ ?
- (d) Suppose that 60% of all tertiary students, if asked, would agree with the statement that "I like learning, but I don't like assessment". A random sample of 20 tertiary students were asked if they agreed with the statement. If  $X$  is the number of students among the 20 who agree with the statement, we can write  $X \stackrel{d}{=} \text{Bi}(20, 0.6)$ . That means that  $X$  has a Binomial distribution with parameters  $n = 20$  and  $\theta = 0.6$ .
- (a) What is the probability that exactly 12 of the 20 students agreed?  
i.e. find  $\Pr(X = 12)$ .  
[ Calc > Probability Distributions > Binomial; click Probability; enter 20 for the Number of trials; enter 0.6 for the Event probability; click Input constant; enter 12 in the box; click OK. ]
- (b) What is the probability that more than 15 of the students agreed?  
i.e. find  $\Pr(X \geq 16)$ .
- (c) What is the probability that more than 16 of the students agreed?

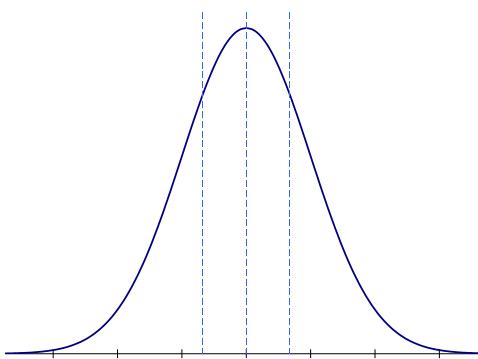
4.2 Assume that the time (in minutes) to complete a particular task is Normally distributed with a mean of 40 and a standard deviation of 5. (°)

- (a) Indicate the values on the horizontal axis of the graph of the completion time distribution shown below:



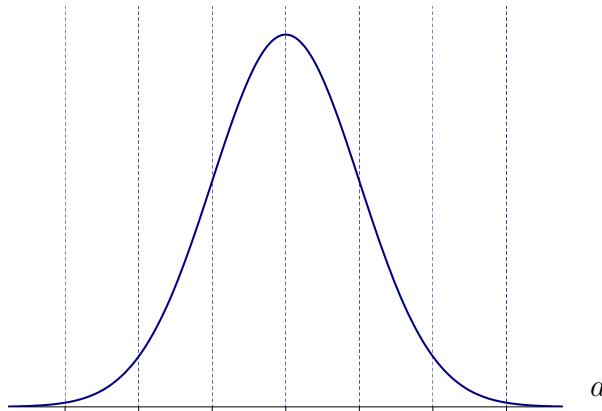
Calculate:

- (b) the probability that an individual completes the task in under half an hour; (Show this probability on the diagram above, and hence guess its value before doing the calculation.)
- (c) the proportion of individuals who are expected to complete the task between 30 minutes and 45 minutes;  
(Show this probability on the diagram.)
- (d) the time by which at least 90% of the individuals will have completed the task;  
(Indicate roughly where this value is on the diagram, before you compute it.)
- (e)\* the quartiles of the task-completion times.



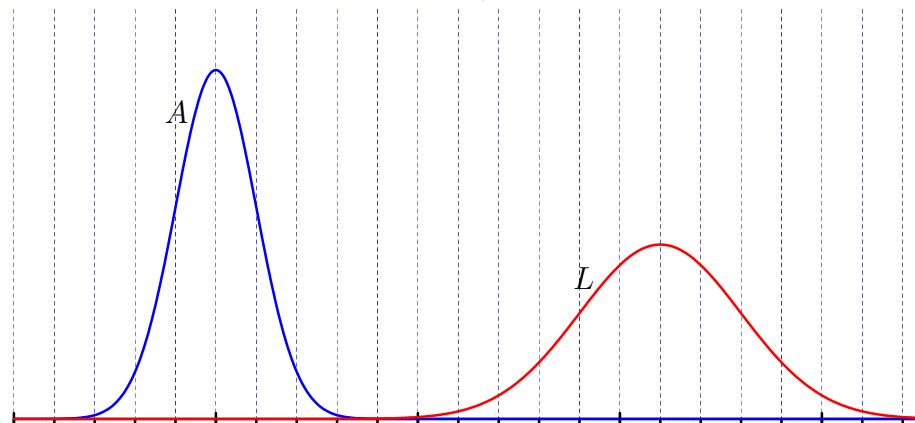
4.3 The amount of an anaesthetizing agent,  $A$ , required to cause surgical anaesthesia in patients has a Normal distribution with a mean of 50 mg and a standard deviation of 10 mg. (♦)

- (a) The diagram below represents the distribution of  $A$ . Label the ticks on the axis and, without any calculation, indicate roughly on the diagram the dose required to bring 99.9% of patients to surgical anaesthesia.



- (b) The “anaesthetic dose” is the dose required to bring 99.9% of patients to surgical anaesthesia. Find the anaesthetic dose in this case.  
 (c) The lethal dose,  $L$ , is also Normally distributed, but with a mean of 160 mg and a standard deviation of 20 mg. If the anaesthetic dose were used, i.e. the dose that brings 99.9% of patients to surgical anaesthesia, as found in (b), what percentage of patients would be killed?

(Indicate 0, 50, 100, 150 and 200 on the horizontal axis in the diagram below; and mark the anaesthetic dose.)



4.4 The electrical resistance of a coil is subject to an upper specification of 25 ohms and a lower specification of 24 ohms. Examination of a large number of coils indicates that the manufacturer is producing coils such that the resistances are Normally distributed with mean 24.62 ohms and standard deviation 0.22 ohms. (◦)

(a) What proportion of the coils would you expect to find outside each specification?

(b) Four of these coils are used in series in a production component. We are concerned with the resistance of a component, which is the sum of the resistances of the coils. Assuming that the coils are selected randomly, the resistances of the coils are independent and each is  $N(24.62, 0.22^2)$ . (◦)

(a) Show that the mean of the sum of the resistances is 98.48.

(b) Show that the variance of the sum of the resistances is 0.1936.

It follows that the distribution of the resistance of a component, i.e. the sum of the resistances of the four coils, is  $N(98.48, 0.44^2)$ .

Find the probability that the resistance of a component is more than 100 ohms.

(c) The final unit contains two of these components and it is important that the total resistance in the separate components should agree closely.

The distribution of the resistance of component 1 is  $N(98.48, 0.44^2)$ ; and

the distribution of the resistance of component 2 is  $N(98.48, 0.44^2)$ .

The resistances of the two components are independent.

(a) Show that the difference between these two resistances has mean 0.

(b) Show that the difference between these two resistances has variance 0.3872.

It follows that the distribution of the difference in resistances of two independent components is  $N(0, 0.6223^2)$ .

Within what limits, symmetrically arranged about zero, would you expect 95% of the differences between the resistances to lie?

4.5 As part of an experiment, a random sample of 25 mice are to be injected with a drug at a dose level of 0.004 mg per gram of body weight. For this strain of mice, body weight (in grams)  $\stackrel{d}{=} N(19, 4^2)$ . (°) On average, the total weight of 25 mice is  $25 \times 19 = 475$  g, which requires  $475 \times 0.004 = 1.9$  mg of the drug.

- (a) If the investigator has 2 mg of the drug, this may not be enough to give the required dose to all 25 mice, because the mice will not all weigh 19 g.

Further, although the mice are similar, they are not identical! It is reasonable to assume that their weights are independent observations from the population of mouse-weights.

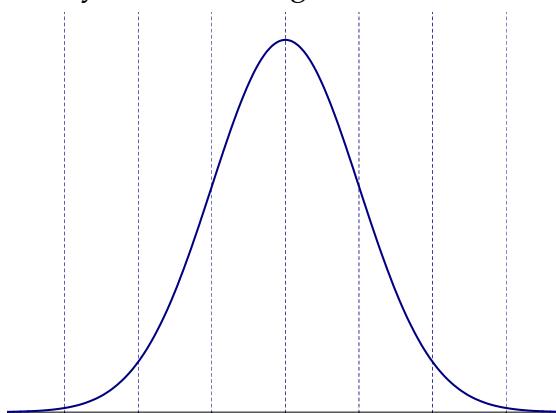
- (a) What is the distribution of the total weight of 25 mice?

- (b) Show that if the total weight is greater than 500 g then more than 2 mg of the drug would be required. Note: drug (mg) =  $0.004 \text{ (mg/g)} \times \text{total.weight (g)}$ .

- (c) Find the probability that the total weight is greater than 500 g. What is the probability that 2 mg of the drug will not be enough?

- (b) Find the 99th percentile of the total weight distribution.

How much of the drug should the investigator secure in order to run a risk of only 1% of running short?



*(Use this diagram for the distribution of the total weight of the 25 mice.)*

4.6 The Omega program is a diet and exercise-based weight loss program aimed at middle-aged women. The weight lost by women using this program for 6 months has a Normal distribution with a mean of 3 kg and a standard deviation of 2 kg. If  $X$  denotes the weight lost by a randomly chosen woman on the program, then  $X \stackrel{d}{=} N(3, 2^2)$ . (◦◦◦)

- (a) Consider first the weight lost by one woman on the program. What is the probability of a woman (selected at random from women enrolling in the program) losing between 2.5 and 3.5 kg? i.e. find  $\Pr(2.5 < X < 3.5)$ .

(Note: If  $X < 0$  then this means the woman has actually gained weight; a negative decrease = an increase. What proportion of women gain weight while on this program?)

- (b) Consider the average weight lost by a random sample of four women on the program. What is the probability that the average weight lost for these four women will be between 2.5 and 3.5 kg? i.e. find  $\Pr(2.5 < \bar{X} < 3.5)$ .

Recall that if  $X \stackrel{d}{=} N(3, 2^2)$ , then  $\bar{X} \stackrel{d}{=} N(3, \frac{2^2}{n})$  where  $n$  is the sample size.

- (c) Now consider the average weight lost for a different random sample of 16 women on the program. What is the probability that the average weight lost by these 16 women will be between 2.5 and 3.5 kg? i.e. find  $\Pr(2.5 < \bar{X} < 3.5)$ .

- (d) What conclusion follows from a comparison of the three answers above?

## 4.8 Answers

- 4.1 (a)  $\Pr(\text{no rain}) = 1 - \Pr(\text{rain}) = 0.7$ .  
 (b)  $\Pr(\text{at least one girl}) = 1 - \Pr(\text{all boys}) = 1 - 0.063 = 0.937$ .  
 (c) 0.09  
 (d) (a) 0.1797  
 (b)  $\Pr(X \geq 16) = 1 - \Pr(X \leq 15) = 1 - 0.9490 = 0.0510$ .  
 (c) 0.0160
- 4.2 As we noted in the previous question, for a discrete random variable that only takes whole number values,  $\Pr(X \geq 7) = 1 - \Pr(X \leq 6)$ , for example, because  $X$  cannot take any value between 6 and 7. A continuous random variable, on the other hand, can take any value within a particular interval, so  $\Pr(X \geq 14.1) = 1 - \Pr(X \leq 14.1)$ .
- (a) Label the mean and the values 1, 2 and 3 standard deviations either side of the mean on the x-axis: (25, 30, 35, 40, 45, 50, 55).  
 (b)  $\Pr(X \leq 30) = 0.0228$ .  
 (c)  $\Pr(30 \leq X \leq 45) = \Pr(X \leq 45) - \Pr(X \leq 30) = 0.841 - 0.023 = 0.819$   
 (d)  $\Pr(X \leq 46.4) = 0.9$   
 (e)\* First quartile:  $\Pr(X \leq Q_1) = 0.25$ , population  $Q_1 = 36.6$   
 Second quartile (median):  $\Pr(X \leq Q_2) = 0.50$ , population median = 40  
 Third quartile:  $\Pr(X \leq Q_3) = 0.75$ , population  $Q_3 = 43.4$
- 4.3 (a) Label the mean and the values 1, 2 and 3 standard deviations either side of the mean on the x-axis.  
 (b) Let  $A$  be the dose required for anaesthesia,  $A \stackrel{d}{=} N(50, 10^2)$   
 Let  $d$  = dose that brings 99.9% of patients to surgical anaesthesia, then  
 $\Pr(A \leq d) = 0.999$ . We find  $d = 80.90$ .  
 (c) Let  $L$  be the lethal dose.  $L \stackrel{d}{=} N(160, 20^2)$   
 Probability of being killed by a dose of 80.902 mg is given by  
 $\Pr(L \leq 80.902) = 0.0000383$  (about 4 in 100 000).
- 4.4 (a) Let  $R$  be the resistance of the coils:  $R \stackrel{d}{=} N(24.62, 0.22^2)$ .  
 Proportion of coils below the lower specification:  $\Pr(R \leq 24) = 0.002$   
 Proportion of coils above the upper specification:  $\Pr(R \geq 25) = 1 - \Pr(R \leq 25) = 1 - 0.958 = 0.042$ .

- (b) Let  $T$  be the total resistance of four coils: i.e.  $T = R_1 + R_2 + R_3 + R_4$ , where  $R_i \stackrel{d}{=} N(24.62, 0.22^2)$

The means add, and the variances add in this case because  $R_i$  are independent.

Hence  $T \stackrel{d}{=} N(98.48, 0.44^2)$  [ $0.1936 = 0.44^2$ ].

(Draw a sketch of the distribution of  $T$ .)

$$\Pr(T \geq 100) = 1 - \Pr(T \leq 100) = 1 - 0.9997 = 0.0003.$$

- (c) Let  $D$  be the difference in resistance of two components:  $D \stackrel{d}{=} N(0, 0.6223^2)$ .

We take the difference of the means, and add the variances in this case because  $T_i$  are independent.

$$\Pr(-d \leq D \leq d) = 0.95, \text{ so } d = 1.96 \times 0.6223 = 1.22$$

We expect 95% of the differences to be between  $-1.22$  and  $1.22$  ohms.

#### 4.5 (a) Consider the distribution of total weight.

- (a) Let  $T$  = total weight of the 25 mice. The mean of  $T$  is the sum of the means, hence  $25 \times 19 = 475$ . Since the weights are independent, the variance of  $T$  is equal to the sum of the variances, or  $25 \times 16 = 400 = 20^2$ . The sum of a number of normal random variables is itself normally distributed. So  $T \stackrel{d}{=} N(475, 20^2)$ .

- (b) The dose level is 0.004 mg of drug per gram of body weight, so 2 mg of drug would be enough for 500g of body weight. So if the total weight is more than 500g, 2 mg of drug will not be enough.

$$(c) \Pr(\text{investigator runs short of the drug}) = \Pr(T \geq 500) = 1 - 0.8944 = 0.1056.$$

- (b) Find  $t_0$  such that  $\Pr(T \geq t_0) = 0.01$ . Use Graph > Probability Distribution Plot > View Probability ...

$$\text{Hence amount of drug required} = 521.5 \times 0.004 = 2.086 \text{ mg.}$$

#### 4.6

- (a)  $\Pr(2.5 \leq X \leq 3.5) = \Pr(X \leq 3.5) - \Pr(X \leq 2.5) = 0.5987 - 0.4013 = 0.1974$ .

$\Pr(X \leq 0) = 0.0668$ , so about 7% of women gain weight on the program.

- (b) When  $n = 4$ ,  $\bar{X} \stackrel{d}{=} N(3, 1)$ .

$$\Pr(2.5 \leq \bar{X} \leq 3.5) = \Pr(\bar{X} \leq 3.5) - \Pr(\bar{X} \leq 2.5) = 0.383.$$

- (c) When  $n = 16$ ,  $\bar{X} \stackrel{d}{=} N(3, 0.25)$ .

$$\Pr(2.5 \leq \bar{X} \leq 3.5) = 0.683.$$

- (d) As the sample size increases, the probability of a sample mean falling between 2.5 and 3.5 kg increases. This is because the variability in sample means is decreasing as the sample size increases. So advertising for the weight loss program could be based on means estimated from large samples, but “individual cases may differ”.



## 5 Confidence intervals

### 5.1 Concept behind confidence intervals

The sample proportion, sample mean and sample standard deviation are examples of **point estimators**: they result in a single value. It is extremely unlikely that a point estimate will be exactly equal to the parameter being estimated and without some idea about the precision of the estimate, its usefulness is limited.

For example, suppose we are interested in the probability,  $\theta$  (say), that a drawing pin lands with its point up and suppose that two people carry out appropriate experiments, independently of each other, and come up with point estimates of 0.3 and 0.5. What can we conclude? Very little, unless we are given more information. If, however, the first estimate is likely to be within  $\pm 0.1$  of the true value of  $\theta$ , and the second within  $\pm 0.2$  of  $\theta$ , then the first estimate is more precise than the second.

It is common practice to associate intervals with estimates to indicate their precision; such intervals are referred to as confidence intervals.

A **confidence interval** is an interval (a set of values between two limits) within which we are *quite* confident that the true parameter value lies. ‘Quite confident’ here is specified by the **confidence coefficient** (or confidence level), typically 95%, associated with the confidence interval.

Consider the exit poll example again. We probably intuitively feel comfortable about saying: “The sample estimate was 44.3%, but the population figure might be 43% or 45%.” However, intuitively, with a random sample of 1000 and an estimate of 44.3%, we feel sure that the population value could not be 90%.

So it is useful to ask the question: can we work out an interval that is highly likely to contain the true value? The answer to this question is a confidence interval. For example, a 95% confidence interval is an interval within which we are 95% confident that the true value of the parameter lies.

Before we describe how to work out a confidence interval, it is important to understand the meaning of the confidence coefficient. Firstly note that any *actual* interval either contains or does not contain the true value of the parameter. So the confidence coefficient, 95% say, does not mean that the chance of a *particular* interval containing the true value of the parameter is 95%. Rather, it refers to the long term proportion of such intervals containing the true value.

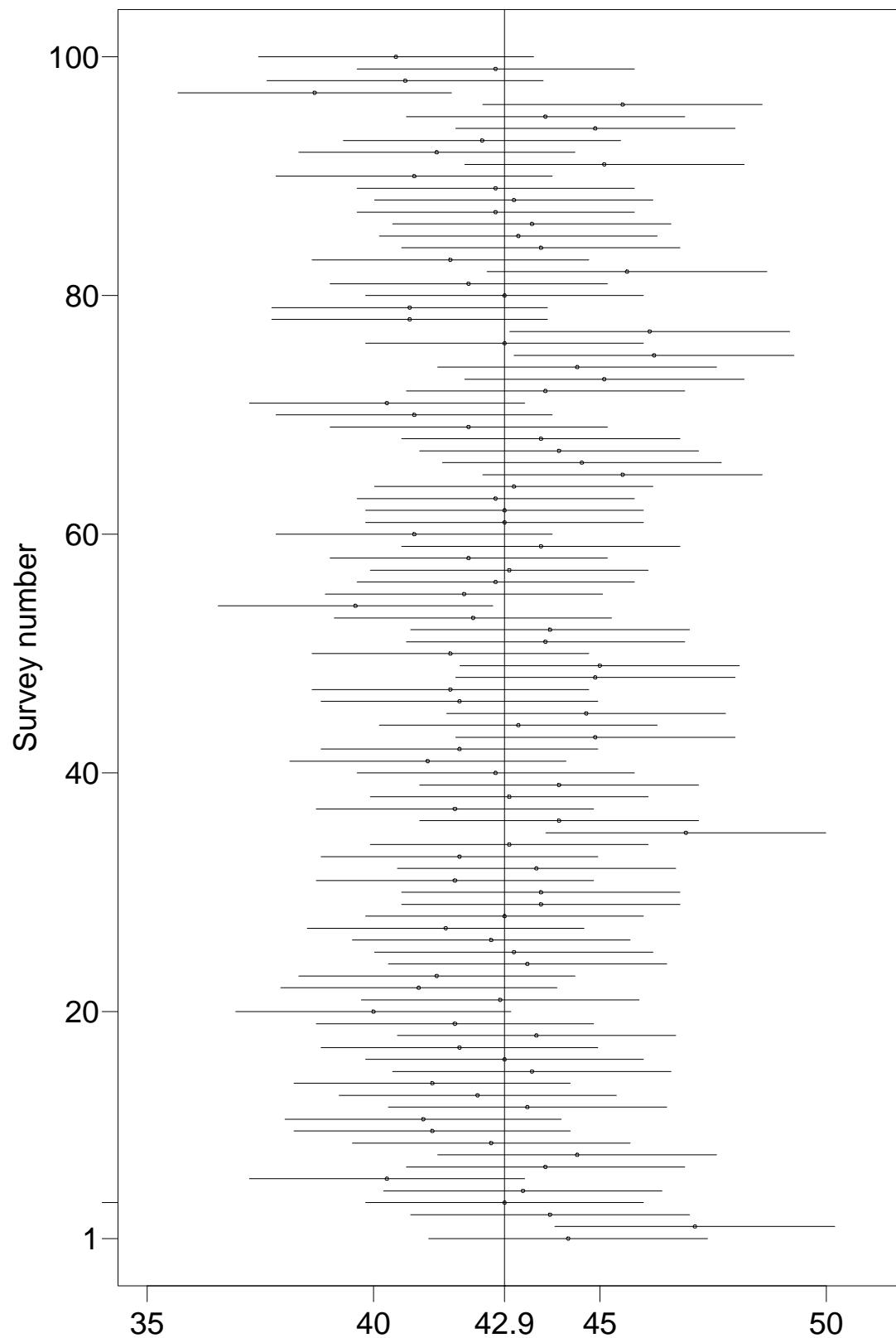


Figure 46: Estimates and 95% confidence intervals from 100 surveys of size 1000 when the true population percentage is 42.9%.

To understand this better, it is helpful to imagine an endless sequence of repetitions of the study you carried out. (°) For each of these repetitions, a specific interval around the estimate is obtained. We obtain these intervals in such a way that *in the long run* 95% of the intervals include the true, unknown value of the parameter. Figure 46 helps to illustrate this concept. It is a plot of estimates and 95% confidence intervals for the population parameter, the percentage of voters voting ALP, from a hypothetical set of 100 surveys, the first 10 of which were mentioned in Section 3.1. The parameter value, 42.9%, is shown on this plot. In general, we do not know this value! Not all confidence intervals include the true value: the second one does not, in fact. Sometimes the estimate is exactly right (see survey number 4), and the confidence interval is positioned symmetrically around the true value. In the long run, 95% of such intervals contain the true value. In practice, we have only one sample and hence only one interval. So we say that we are 95% confident that the interval we have contains the true value.

▷ **QUESTION:** There are 100 intervals represented in Figure 46. How many contain the true value? Are you surprised by the answer?

### 5.1.1 Choice of the confidence coefficient

▷ **QUESTION:** Why do we settle for 95% confidence? Why not 100%? Or, for that matter, why not 50%?

The choice of confidence coefficient is made *a priori* — before the confidence interval is calculated. The use of 95% is almost standard practice, and means that the precision of results from different studies, for example, can be compared on the same basis.

If a study uses a confidence coefficient smaller than 95%, you should look carefully at the justification for this. This is an unusual choice and potentially could mislead the reader to think that the study has particularly good precision.

▷ **EXAMPLE. Passive smoking and lung cancer**

A court case in the USA between tobacco companies and the Environmental Protection Agency (EPA) debated the choice of the confidence coefficient.<sup>12</sup> The EPA had chosen to report a 90% confidence interval in estimating the relative risk of lung cancer for those exposed to passive tobacco smoke, compared with those not exposed. The estimate of the risk was a 19% increase, if exposed, with the 90% confidence interval ranging from 4% to 35%. The EPA argued that a 95% confidence interval would be wider, with a lower bound that could allow for the possibility that passive smoking reduced

---

<sup>12</sup>Health: Statisticians occupy the front line in battle over passive smoking, The Wall Street Journal, July 28, 1993.

lung cancer; as this was ‘inconceivable’ they chose not to use a 95% confidence coefficient.

Statisticians on the tobacco company side argued that 90% was non-standard. The court agreed with the view that health research that involves the assessment of risk should not incorporate judgmental positions or be influenced by the ‘mission’ of those conducting the research.

### 5.1.2 Considering precision

Use of a standard confidence coefficient facilitates comparison of the relative precision of results from different studies.

Since the mid-1970s there has been a strong campaign in the medical sciences (in particular) to promote the use of confidence intervals. In the medical and epidemiological literature, this developed momentum with the publication in 1978 in the New England Journal of Medicine of the article “A show of confidence” by Ken Rothman.<sup>13</sup> The 6th edition of the American Psychological Association’s (APA) publication guidelines prescribed confidence intervals in the minimal expectations of all APA journals.

This has been partly a reaction to the most commonly found alternative mode of inference, namely, hypothesis testing. But the virtues of a confidence interval can be stated without reference to hypothesis testing. The basic idea is that estimation of a parameter is a process that has good research focus, and that a confidence interval is a natural and interpretable way to express the precision of a point estimate.

#### ▷ EXAMPLE. A confidence interval in advertising

A columnist in The Age newspaper, ‘Leaping Larry L’, discussed an advertisement for a hair product where the claim was that “70% of independent hairstylists would use … [the product] … on their own hair.” This was supported by the claim: “Survey verified by independent research agency. Plus or minus 20.7 % at the 95% confidence interval.” Leaping Larry L’s comment is “Hmm, I don’t know what a “confidence interval” is but I think I just had a big one. Plus or minus nearly 21% is not a margin of error I’d care to build a bridge by, frankly.”

Leaping Larry L clearly *does* know what a confidence interval is, and he is concerned about the lack of precision in the estimate reported. He uses the term ‘margin of error’ which is one way a confidence interval is sometimes reported. For symmetric confidence intervals, the margin of error is the half-width of the confidence interval; it specifies the distance from the point estimate to the endpoints of the interval. This is illustrated in Figure 47.

It seems likely that the data were 13 out of 18 people surveyed agreed they

---

<sup>13</sup>Rothman, KJ. (1978) A show of confidence. NEJM 299:1362-3.

would use the product, as this gives an approximate 95% confidence interval of  $72.2 \pm 20.7$ , assuming a random sample. Put another way, the confidence interval is 52% to 93% — as Leaping Larry L notes, very wide — reflecting the poor precision of the survey.

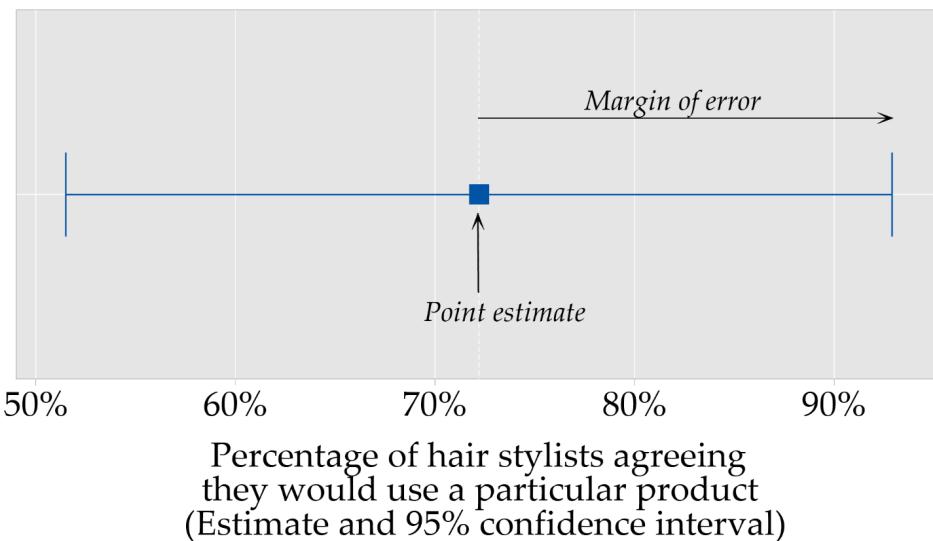


Figure 47: *Estimates and approximate 95% confidence interval for the likely data in the survey described by Leaping Larry L*

### 5.1.3 Reporting confidence intervals

There is a practical issue: how should you write down a confidence interval, having worked it out? An actual confidence interval is a range, including all numbers between two limits. We report the lower limit and the upper limit of this range to describe the confidence interval.

- (a) In mathematics, a common notation for an interval of this sort is  $(a, b)$ , where  $a$  and  $b$  are the ends of the interval; for example,  $(2.3, 4.5)$ . This representation is sometimes used for a confidence interval, but it is not always well-understood by non-mathematicians. The APA recommends this form, but with square brackets:  $[-5.6, -2.7]$ .
- (b) An alternative notation is  $(a - b)$ ; for example  $(2.3 - 4.5)$ . To some people, this more specifically suggests the interval between the two limits. However, one or both of the limits of a confidence interval may be negative, which can be confusing to read. If the limits are  $-5.6$  and  $-2.7$ , it is awkward to write the interval as  $(-5.6 - -2.7)$ , and the other notation is preferable:  $(-5.6, -2.7)$ .
- (c) A third alternative which uses a word to help with the possible confusion is to write the interval as  $(a \text{ to } b)$ , for example:  $(2.3 \text{ to } 4.5)$ , or

(−5.6 to −2.7).

- (d) A final form which can sometimes be useful is to write the interval as “estimate  $\pm$  the distance to the endpoints”. For example, the interval we’ve been discussing may have had a point estimate of 3.4; the 95% confidence interval can then be expressed as  $3.4 \pm 1.1$ . This has the virtue of directly representing the point estimate, which is hidden somewhat in the other representations. This method is not applicable when the confidence interval is not symmetric around the point estimate.

When the confidence interval can be expressed in these terms, the “distance to the endpoints” referred to above is, as we have just seen, often called the “margin of error” (see Figure 47). This is commonly used in survey science. The confidence coefficient will often not be stated; if so, it can usually be assumed to be 95%. You may notice this term used in discussion of political polling, especially near election time.

## 5.2 Confidence intervals in Normal populations

### 5.2.1 The mean of a Normal population, $\mu$

In this subsection, we derive the confidence interval for the mean of a Normal distribution, based on a single random sample. This derivation has essential elements of the structure of many different confidence intervals.

We are implicitly using a common idea here, that the *distribution* we consider reflects the *population* of a variable of interest. We learn about the population by considering the distribution.

We have already focussed on 95% confidence intervals; 95% is an arbitrary but reasonable and very common choice for the confidence coefficient. As part of the derivation, we need to find the interval around the mean of a Normal distribution that corresponds to a (central) probability of 95%. It makes sense for this interval to be symmetrically placed around the mean.

For the standard Normal distribution, which has mean 0 and standard deviation 1, i.e.  $Z \stackrel{d}{=} N(0, 1)$ , it turns out that

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

We can find this by using the inverse cumulative probability in MINITAB, for  $N(0, 1)$ .

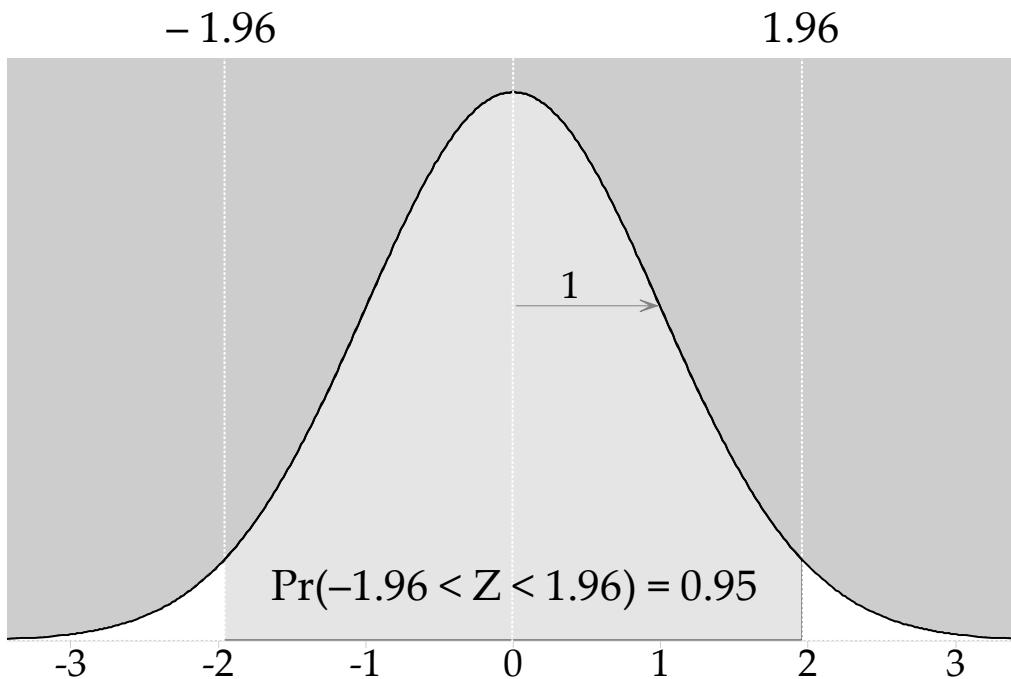


Figure 48: Standard Normal distribution  $Z$  with mean 0 and standard deviation 1, showing that  $\Pr(-1.96 < Z < 1.96) = 0.95$ .

In order to obtain this symmetrically placed central interval, we seek numbers  $-c$  and  $c$ , either side of zero, symmetrically placed, which are such that  $\Pr(-c < Z < c) = 0.95$ . This implies that we want a central probability of 0.95, and a probability of 0.025 above  $c$  and a probability of 0.025 below  $-c$ . So we can ask for the inverse cumulative probability for 0.025: we get  $-1.96$ . If we ask for the inverse cumulative probability of 0.975, we get  $1.96$ . Hence

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

From the properties of the Normal distribution, the probability that *any* Normal random variable is within 1.96 standard deviations<sup>14</sup> of its mean is 0.95. So for a Normal random variable  $U$  with mean  $\lambda$  and standard deviation  $\phi$ , we can write

$$\Pr(\lambda - 1.96\phi < U < \lambda + 1.96\phi) = 0.95.$$

You can see examples of this in Figure 49.

---

<sup>14</sup>It turns out that the 0.975 quantile is 1.95996 to 5 decimal places, so 1.96 is very close to the precise value.

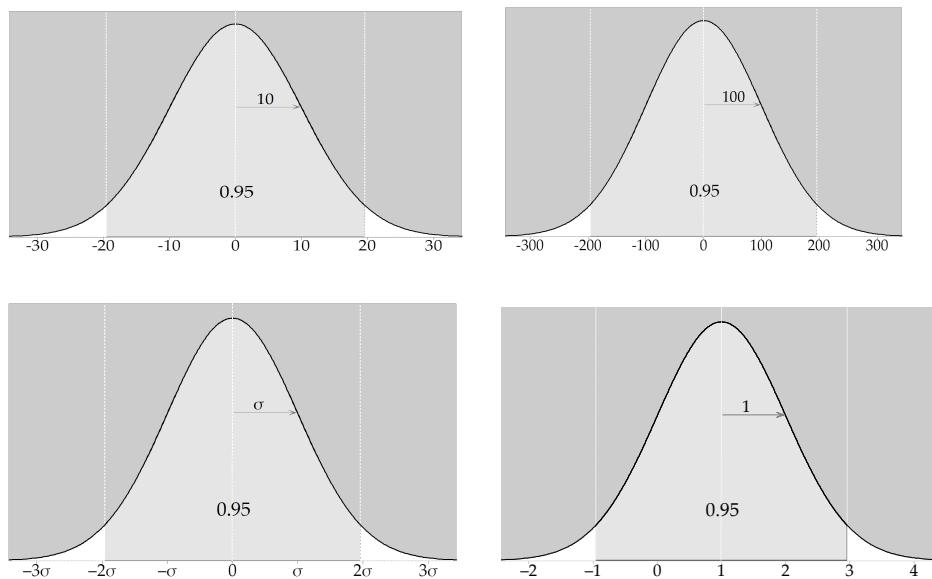


Figure 49: Various Normal distributions showing a central probability of 95%

Now we return to the task of deriving the confidence interval for  $\mu$  in this exemplar case.

Let  $X$  be a Normal random variable with mean  $\mu$  and variance  $\sigma^2$ . The parameters  $\mu$  and  $\sigma$  are generally unknown. Let  $X_1, X_2, \dots, X_n$  denote a random sample of size  $n$  from the population. Then  $\bar{x} = \frac{\sum x_i}{n}$  is an estimate of  $\mu$ ; and the estimator,  $\bar{X} = \frac{\sum X_i}{n}$ , follows a Normal distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ , i.e.  $\bar{X} \stackrel{d}{=} N(\mu, \frac{\sigma^2}{n})$ .

In order to simplify the discussion we assume for the moment that  $\sigma$  is known (◊). It is important to understand that this is quite artificial; it is not a realistic assumption in practice. We are only making it to learn about the derivation of a confidence interval. When it comes to inferences and practical contexts, we do not have to make the assumption that  $\sigma$  is known, and we don't.

For random samples from a Normal distribution, the mean of  $\bar{X}$  is  $\mu$  and the standard deviation of  $\bar{X}$  is  $\frac{\sigma}{\sqrt{n}}$ . So for a given value of  $\mu$ , there is a probability of 0.95 that  $\bar{X}$  will be within  $1.96 \frac{\sigma}{\sqrt{n}}$  of  $\mu$ . That is:

$$\Pr \left( \mu - 1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 \frac{\sigma}{\sqrt{n}} \right) = 0.95. \quad (1)$$

The sample mean,  $\bar{X}$ , is in the centre of this probability statement, which is saying that the chance is 0.95, or 95%, that  $\bar{X}$  is within a distance  $\pm 1.96 \frac{\sigma}{\sqrt{n}}$  of  $\mu$ , that is, within the interval  $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$ , as illustrated in Figure 50.

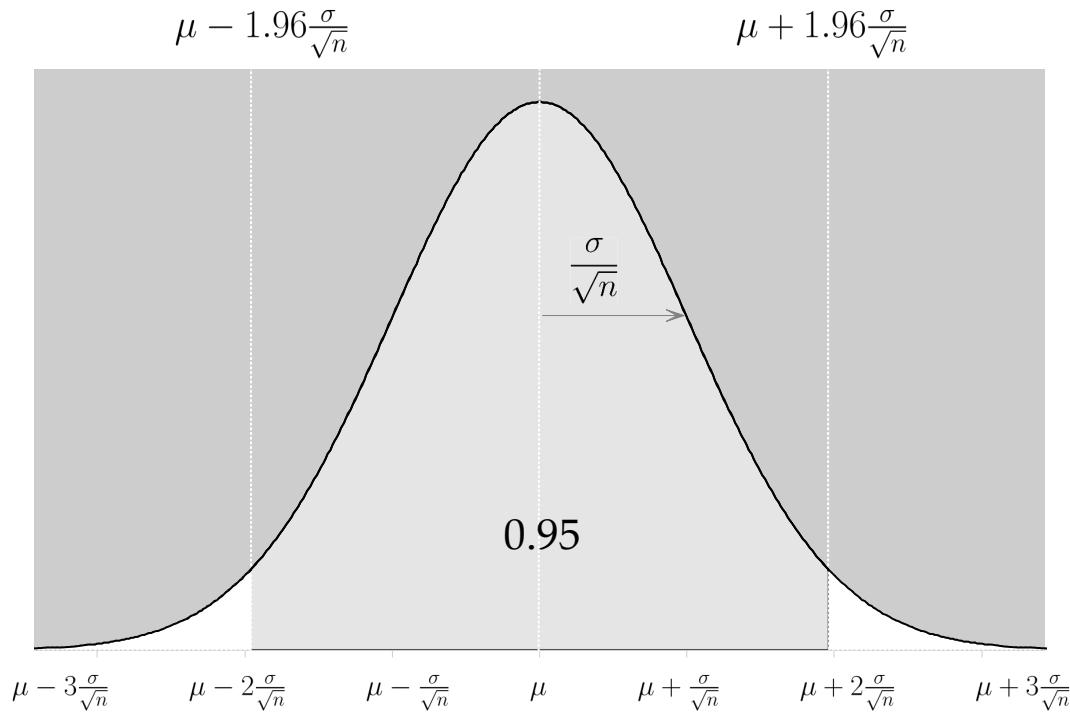


Figure 50: Distribution of  $\bar{X}$  based on a random sample from a Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

This implies that there is a chance of 0.95, or 95%, that  $\mu$  is within a distance  $\pm 1.96 \frac{\sigma}{\sqrt{n}}$  of  $\bar{X}$ , that is, within the interval  $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ .

In other words, we can write:

$$\Pr(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95. \quad (2)$$

There is no subterfuge here. The two results are saying exactly the same thing. Each can be derived algebraically from the other. They just have a different focus: in equation (1) the focus is on the fixed interval  $\mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$ , and the chance that the random variable  $\bar{X}$  falls within it. In equation (2), the focus is reversed: now we are considering a random interval  $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ , and asking: what is the chance that it includes the fixed value  $\mu$ ?

Equation (2) is the basis for a 95% confidence interval. It is a probability statement about the random interval  $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ . We are saying that there is a 95% chance that  $\mu$  will fall within the interval

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

When we have actual data, we have the observed  $\bar{x}$ , and we say that the 95% confidence interval is  $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ .

Now that we are talking about an actual observed value there is no randomness:  $\bar{x}$  is just a number, and so are  $\sigma$  and  $n$ . So, in fact, the interval

$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$  either does include  $\mu$ , or it doesn't. We don't know which of these is true. Look at Figure 46 again: each individual confidence interval either includes the true value, or not. Hence the 95% does not refer to the individual, reported confidence interval. It refers to the properties of the method.

It is always useful to think of the repeated sampling idea here, illustrated in Figure 46, and also seen in StatPlay. Imagine taking many random samples, and working out the confidence interval each time. Most of them will include the value of the mean,  $\mu$ ; some will not. In the long run, 95% of the intervals will include it.

Hence  $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$  is a 95% confidence interval for  $\mu$ . The number 1.96 is a multiplier which comes from the standard Normal distribution, and is a consequence of using a confidence coefficient of 0.95 (95%); if we change the confidence coefficient to some other value then a different multiplier is required.

▷ **EXAMPLE. Language ability in youth in detention**

The Clinical Evaluation of Language Fundamentals (CELF) is a test for assessing general language ability. It provides a standardised score (based on age) which has a mean of 100 and a standard deviation of 15 for the general population. As part of a study to develop an intervention for young males in detention in Victoria, Dr Nathaniel Swain assessed communication needs using the CELF.

For the sample of 27 youth, the mean CELF score was 79.4.

Assume that the distribution of CELF among youths in detention is Normal, with standard deviation  $\sigma$  equal to 15, and that we have a random sample.

A 95% confidence interval for  $\mu$ , the true population mean of male youth in detention, is therefore

$$79.4 \pm 1.96 \times \frac{15}{\sqrt{27}}, \text{ i.e. } (73.7, 85.1).$$

▷ **QUESTION:** Would you ever know if the general population distribution was Normal or not?

▷ **QUESTION:** How would you know that the true population standard deviation was any particular value, let alone 15?

▷ **QUESTION:** What would a 50% confidence interval be? What about a 0% confidence interval?

In practice, of course, it is unlikely that  $\sigma$  will be known. Under these circumstances the obvious thing to do is to replace  $\sigma$  by the sample estimate,  $s$ , which gives an approximate 95% confidence interval of the form

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

Provided  $n$  is large enough ( $> 100$  say),  $s$  is a ‘good’ estimate of  $\sigma$  and the above interval will be quite satisfactory. If  $n$  is small, it is usual to account for the uncertainty associated with  $\sigma$  by replacing the constant, 1.96, by a value from the ‘ $t$ -distribution’.

The  $t$ -distribution is a continuous distribution based on the Normal distribution and the  $\chi^2$  distribution (we define the  $\chi^2$  distribution later on). Three  $t$  distributions with different “degrees of freedom” are shown in Figure 51. The concept of degrees of freedom (df) is discussed below (page 138).

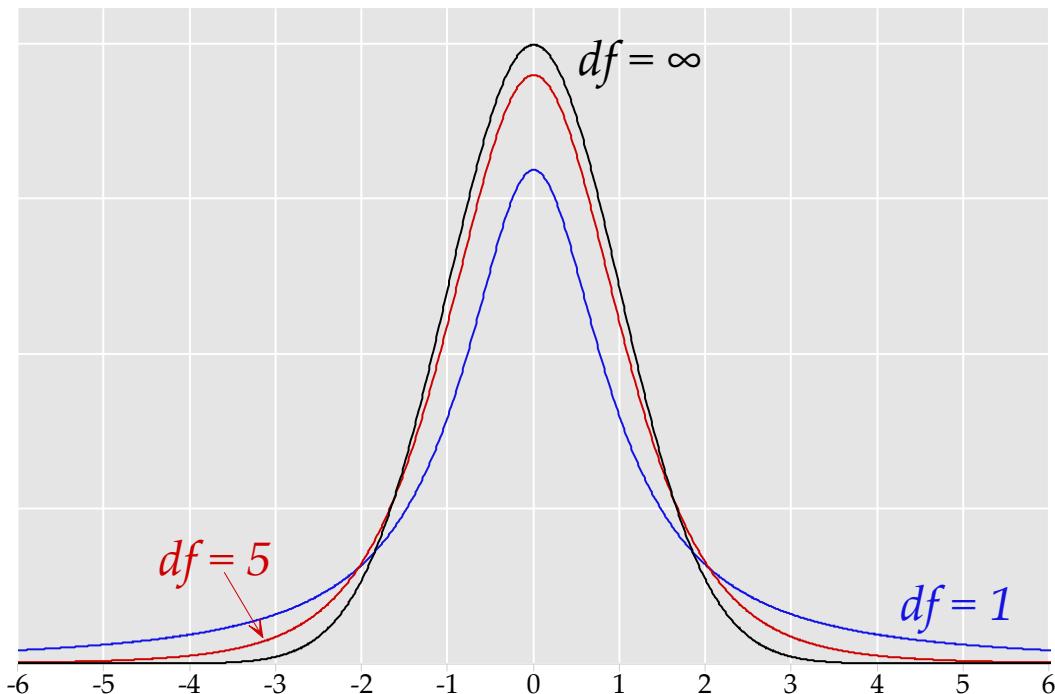


Figure 51: Probability density functions for  $t_1$ ,  $t_5$  and  $t_\infty$ ; the  $t_\infty$  distribution is the same as the standard Normal distribution.

This gives a 95% confidence interval of the form

$$\bar{x} \pm t_{(n-1)}(0.975) \frac{s}{\sqrt{n}},$$

where  $t_{(n-1)}(0.975)$  is the 97.5 percentile of the  $t$ -distribution with  $(n - 1)$  degrees of freedom. Note that  $t_{(n-1)}(0.975) > 1.96$  but it tends to 1.96 as  $n$  increases.

Recall that in Chapter 4 we learned that the quantity  $\frac{s}{\sqrt{n}}$  is called the “standard error” of  $\bar{X}$ . It is an estimate of the standard deviation of  $\bar{X}$ . The confidence interval we have just obtained is the first and iconic example of a general form of confidence interval, which we see repeatedly in statistics.

This form is

$$\text{estimate} \pm (k \times \text{standard error}).$$

In this general form,  $k$  is a number that comes from an appropriate distribution, which will be either the standard Normal distribution or the  $t$  distribution. The estimate, in any specific context, is thought of as an observation on an estimator, a random variable with a sampling distribution. This distribution has a standard deviation. The standard error, in any specific context, is the estimate of the standard deviation of the estimator. Many confidence intervals — though not all — are of this general form.

So in this context, the estimate is  $\bar{x}$ , the estimator is  $\bar{X}$ , its distribution is  $N(\mu, \frac{\sigma^2}{n})$ , which has a standard deviation  $\frac{\sigma}{\sqrt{n}}$ . The estimate of this standard deviation, the standard error, is  $\frac{s}{\sqrt{n}}$ .

▷ **EXAMPLE. Language ability in youth in detention**

In the study of language ability (CELF) of young males in detention, suppose we don't know the value of  $\sigma$  (or do not want to assume that it is the same as for the general population) and we have to estimate it. We find that  $s = 20.8$  and  $t_{26}(0.975) = 2.056$ ; so (still assuming Normality for the distribution of CELF) a 95% confidence interval for  $\mu$  is now

$$79.4 \pm 2.056 \times \frac{20.8}{\sqrt{27}}, \text{ i.e. } (71.2, 87.6).$$

The outcome in this example is typical of what we expect: the interval for  $\mu$  is a bit wider when we don't assume that we know the value of  $\sigma$ . We are paying a precision price — an appropriate one — by not assuming that  $\sigma$  is known.

In MINITAB: Stat > Basic Statistics ▶ 1-sample t ...

### 5.2.2 Degrees of freedom

The “degrees of freedom” used here was  $n - 1$ , where  $n$  is the sample size. This comes from the sample variance  $s^2$ , which has  $n - 1$  degrees of freedom associated with it. What does this strange, anthropomorphic term mean? The **degrees of freedom** of a set of numbers is the number of elements in the set which are free to vary. For a random sample of size  $n$ , the degrees of freedom is simply  $n$ . But for the set of deviations from the sample mean,  $(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$ , the degrees of freedom is not  $n$  but  $n - 1$ . This is because the sum of these deviations is zero, so if we know  $n - 1$  of the deviations, the remaining one is not “free” but determined.

If you have a sample of size 3, and you tell me that the first deviation from the sample mean is  $-4$ , and the third is  $+1$ , I can tell you what the second one is: it has to be  $+3$ , in order that the sum of them equals zero. They have to balance out. So the last deviation is implied, or determined.

So the set of  $n$  deviations from the sample mean has  $n - 1$  degrees of freedom. The sample variance,  $s^2$ , is based on these  $n - 1$  deviations, so it has the

same  $n - 1$  degrees of freedom associated with it, and this also determines the degrees of freedom for the  $t$  distribution in this case.

### 5.2.3 The difference between the means of two Normal populations — paired samples

In Chapter 1, we introduced paired samples designs where two treatments or interventions are compared based on pairs, where the two members of a pair are very similar and a different treatment is given to each member of the pair.

Here we consider finding a confidence interval for the mean of a numerical variable using a paired design.

When a paired design is used, we expect the two variables to be positively correlated. When an observation on one of the pair members is high (or low), the other one tends be high (or low). It is this correlation that we aim to exploit in a paired design: it is a way of removing sources of variation that are not due to the intervention.

Let  $(X_{11}, X_{21}), \dots, (X_{1n}, X_{2n})$  be a random sample of  $n$  pairs of observations from two **positively correlated** Normally distributed random variables,  $X_1$  and  $X_2$ , with means  $\mu_1$  and  $\mu_2$  respectively. Then if we are interested in making inferences on  $\mu_1 - \mu_2$  it is statistically efficient to consider the random variable  $D = X_1 - X_2$  which has mean  $\mu_D = \mu_1 - \mu_2$  and variance  $\sigma_D^2$ .

Note an important feature of means that we are using here, which we saw in section 4.5. The mean difference is equal to the difference of the means, so if we obtain an inference about  $\mu_D$  we are getting an inference about  $\mu_1 - \mu_2$ .

A 95% confidence interval for  $\mu_1 - \mu_2$  is given by

$$\bar{d} \pm t_{n-1}(0.975) \frac{s_D}{\sqrt{n}}$$

where  $\bar{d}$  is the average of the differences,  $d_i = x_{1i} - x_{2i}$ ,  
 $s_D$  is the sample standard deviation of the differences, and  
 $t_{n-1}(0.975)$  is the 0.975 quantile of the  $t_{n-1}$  distribution.

Unless  $n$  is very small,  $t_{n-1}(0.975)$  is a number which is around 2: thus,  $t_6(0.975) = 2.447$ ,  $t_{60}(0.975) = 2.000$  and  $t_{600}(0.975) = 1.964$ .

Note that the main requirement for the above results is that  $D$  should be Normally distributed. One situation where this requirement is satisfied is when  $X_1$  and  $X_2$  are themselves Normally distributed.

#### ▷ EXAMPLE. Fuel economy, engine tuning study (oz.cars.1.6.litres.mwx)

We return to the data in Exercise 1, from the study of the pollution of cars in use in Australia, in the early 1990s. In addition to examining possible reductions in emissions due to tuning the engine, the researchers also looked at fuel economy. Fuel economy is something that varies a lot with engine

size. For the present purpose, we will examine the cars with an engine size of 1.6 litres in the study. There were 22 such cars: models such as Holden Camira and Gemini, Toyota Corolla, Ford Laser and Mitsubishi Colt. They were all manufactured between 1980 and 1985 inclusive.

The cars were tested, tuned and then re-tested. The following table gives the fuel consumption, pre- and post-tuning, in litres per 100 kilometres, a measure for which a low number is good.

Car	1	2	3	4	5	6	7	8	9	10	11
Pre tuning	11.51	10.75	11.00	10.77	11.32	11.00	10.68	9.47	10.90	9.50	9.98
Post tuning	11.04	10.27	11.27	9.42	11.20	10.79	10.46	10.02	10.61	9.56	10.47
Car	12	13	14	15	16	17	18	19	20	21	22
Pre tuning	9.72	8.04	10.00	10.47	11.01	9.39	9.45	9.04	10.78	10.47	9.93
Post tuning	9.55	7.71	9.41	10.40	9.32	8.10	9.62	8.94	10.64	10.03	10.19

As usual, it is hard to perceive patterns in the data from a listing of it: we need a graph. One way to graph the data is to produce two dotplots, lined up, one for pre-tuning and one for post-tuning.

This is shown in Figure 52. If you just examined this plot, you would not be left with the impression that tuning affects fuel economy very much. It appears that the average fuel economy might be slightly better post-tuning than pre-tuning in these data, and this is confirmed from the descriptive statistics: the mean for pre-tuning is 10.24 litres per 100 km, and post-tuning it is 9.96. The difference between these two means is 0.28 litres per 100 km.

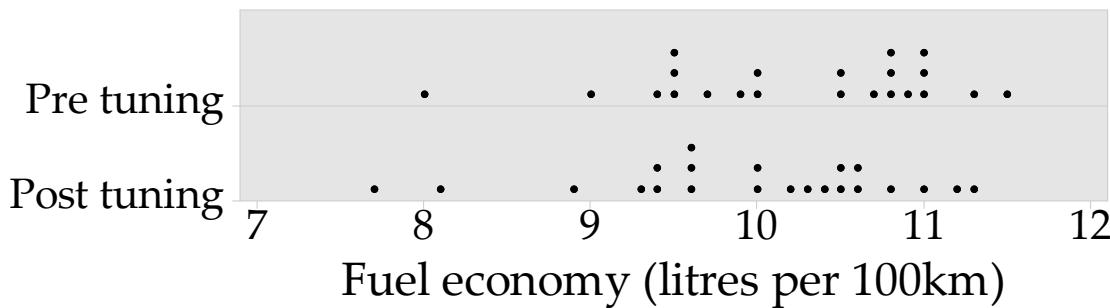


Figure 52: Dotplots for pre-tuning and post-tuning fuel economy (litres per 100 km) on the same set of 22 cars with 1.6 litre engine capacity

But these separate dotplots do not represent the data in the best way. Even within a set of relatively similar cars (all 1.6 litre engine capacity), there will be variation in fuel economy: some will have poor economy, and others good. We would expect there to be quite a strong association between the pre- and post-tuning data, considering the two observations on one car as a pair. This is borne out by simple inspection: look at cars numbered 1 (poor) and 13 (good).

Another perspective is given by the correlation between the pre- and post-tuning data; we find that  $r = 0.80$ , with a 95% CI of (0.57, 0.91). As expected for paired data, there is a positive correlation.

A logical way to consider the tuning effect, allowing for these differences between cars, is to calculate the *difference* between the pre- and post-tuning economy, for each car. We then make inferences about the true mean difference.

So with paired data, the natural focus is on the differences. A simple piece of arithmetic confirms that the difference between the pre-tuning mean and the post-tuning mean is exactly the same as the average of the differences. So we get the same estimated effect, either way; in the example, 0.28 litres per km. But by considering the differences we eliminate unhelpful random variation: in this case, the variation between cars. We are left with only the variation within cars, and the tuning effect.

When differences are used, the direction of the difference is arbitrary, but it is always important to be clear about it; in this example, we take the difference as pre-tuning minus post-tuning, for which a positive difference means a benefit of tuning.

Figure 53 shows the dotplot of the differences.

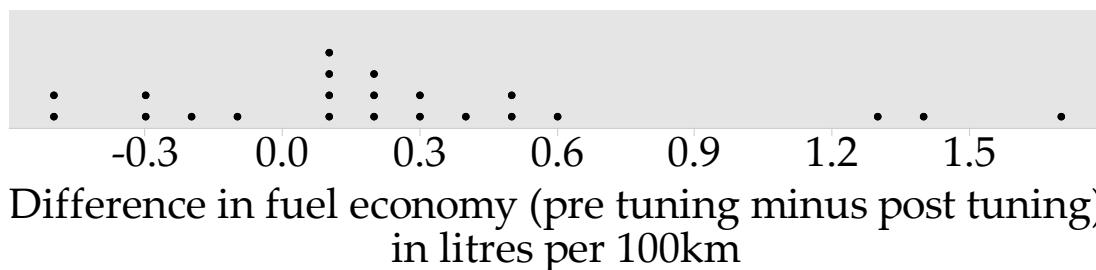


Figure 53: Dotplot of the differences in fuel economy (litres per 100 km), pre-tuning minus post-tuning, for 22 cars with 1.6 litre engine capacity

Some of these differences are positive (an improvement) and some are negative. Note the position of zero in the dotplot. The average difference is small: 0.28 litres per 100 km. But that might be worthwhile if it represented a population effect, both for environmental and economic reasons.

So we seek an inference about the average difference in fuel economy, comparing pre- and post-tuning. We can estimate this average difference and find a confidence interval.

Because of the pairing, we consider the differences within each pair, and make inferences about the distribution of the differences.

This is a before and after study, and therefore, in principle, has the concern

mentioned about such studies in Chapter 1. But in this case there was a very short time between the two measurements, and it is therefore reasonable to suppose that changes associated with the passing of time, in itself, are not very likely to be important.

Let  $\mu_{\text{pre}}$  be the mean fuel economy pre-tuning, and  $\mu_{\text{post}}$  be the mean fuel economy post-tuning. Define  $\mu_D = \mu_{\text{pre}} - \mu_{\text{post}}$ .

$$\bar{d} = 0.28, s_D = 0.564.$$

A 95% confidence interval for  $\mu_D$  is (0.03, 0.53).

This can be done in MINITAB in one of two ways.

- If the differences are in a column, you can use Stat > Basic Statistics ► 1-sample t ..., with the sample in question being the differences.
- The confidence interval can be found without storing the differences in a column explicitly, by using Stat > Basic Statistics ► Paired t ...

A graphical representation of this confidence interval is shown in Figure 54. This illustrates the effectiveness of analysing the differences: the two separate confidence intervals at the top might suggest that the difference between the after and before data is not very pronounced, but when the differences are analysed the confidence interval for the mean difference is quite narrow. (Note that Figure 8 uses scales with the same interval, 0.25, in both parts.)

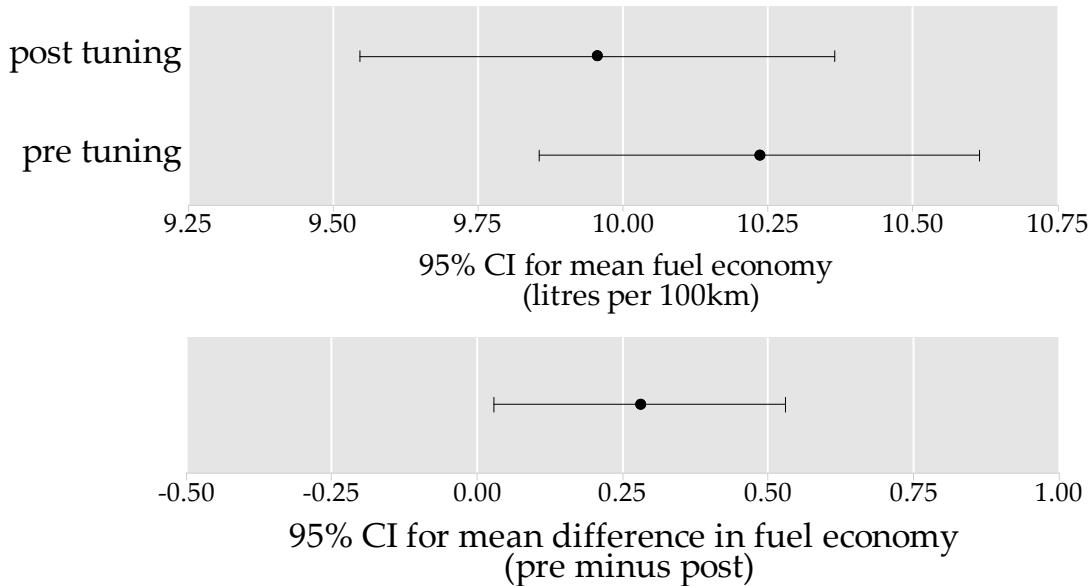


Figure 54: 95% confidence interval for the mean difference in fuel economy, pre-tuning minus post-tuning. Also shown, at the top, are separate 95% CIs for the mean for the pre-tuning setting and the mean for the post-tuning setting.

### 5.2.4 The difference between the means of two Normal populations — independent samples

Now we consider the second kind of experimental design introduced in Chapter 1 — an independent samples design with two groups. Recall that in this design, subjects in the two groups are independent of each other, because they have no structural link with each other. Our interest is again in making an inference about the difference between two treatments or interventions, and when randomisation is used in assigning participants to groups, claims about causality can be considered.

Here we consider finding a confidence interval for the difference of two means when the data arise from an independent samples design.

Let  $X_{11}, \dots, X_{1n_1}$  be a random sample from a random variable  $X_1$  with mean  $\mu_1$  and variance  $\sigma_1^2$ , and let  $X_{21}, \dots, X_{2n_2}$  be a random sample from a random variable  $X_2$  with mean  $\mu_2$  and variance  $\sigma_2^2$ ;  $X_1$  and  $X_2$  are assumed to be independent. Then, if  $X_1$  and  $X_2$  are Normally distributed,

$$\bar{X}_1 - \bar{X}_2 \stackrel{d}{=} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Here we are applying several of the results from Chapter 4:

- The mean of  $\bar{X}$  is  $\mu$  and the variance of  $\bar{X}$  is  $\frac{\sigma^2}{n}$  (we have two sample means here);
- The sum or difference of Normally distributed random variables is itself Normally distributed;
- The mean of a difference between two random variables is the difference of the means;
- The variance of the difference between two *independent* random variables is the sum of the variances.

This result allows us to say that

$$\Pr\left(\bar{X}_1 - \bar{X}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 0.95.$$

This looks very complicated, but is essentially in exactly the same form as the first confidence interval we derived, the CI for  $\mu$  based on a single random sample from a Normal population.

If we knew  $\sigma_1$  and  $\sigma_2$  (most unrealistic) a 95% confidence interval for  $\mu_1 - \mu_2$  would be given by

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 &\pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= \bar{x}_1 - \bar{x}_2 \pm 1.96\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{if } \sigma_1^2 = \sigma_2^2 (= \sigma^2) \end{aligned}$$

When  $\sigma_1^2$  and  $\sigma_2^2$  are unknown (the typical case), the cases of equal and unequal variances need to be considered separately. Here we need to consider what we are prepared to assume about the true underlying population variances.

### Assuming equal population variances

If  $\sigma_1^2 = \sigma_2^2 (= \sigma^2)$ , a 95% confidence interval is given by

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2}(0.975)s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where  $s^2$  is an estimate of  $\sigma^2$  with  $n_1 + n_2 - 2$  degrees of freedom:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The idea of this “pooled” estimate of  $\sigma^2$  is that it is an optimally weighted average of the two sample variances, both of which estimate the common  $\sigma^2$ , by assumption. More weight is given to the sample variance which came from the larger sample.

In MINITAB Stat > Basic Statistics ► 2-sample t . . . , and check the box labelled Assume equal variances.

▷ **EXAMPLE.** iBobbly intervention for suicide prevention (iBobbly.mwx)

The Black Dog Institute carried out a pilot study of a self-help mobile app (iBobbly) for “at risk” Indigenous Australians aged 18-35 years in remote Australia.<sup>15</sup> The app was designed to provide a form of therapy to target suicidal ideation, depression, psychological distress, and impulsivity. The 61 participants were randomised either to receive and use the app (iBobbly) for 6 weeks ( $n = 31$ ) or to be waitlisted for 6 weeks ( $n = 30$ ). (The waitlist group then received the app for 6 weeks.)

Here we consider one outcome, the Kessler Psychological Distress Scale (K10); the K10 was designed to provide an overall measure of distress based on questions about anxiety and depressive symptoms over the past four weeks. Higher scores indicate higher distress. Summaries of the data are shown under Figure 55.

Figure 55 provides boxplots of the K10 total scores at 6 weeks.

---

<sup>15</sup>Tighe, Joseph et al. (2016), Data from: Ibobbly mobile health intervention for suicide prevention in Australian Indigenous youth: a pilot randomised controlled trial, Dryad, Dataset, <https://doi.org/10.5061/dryad.860kn>

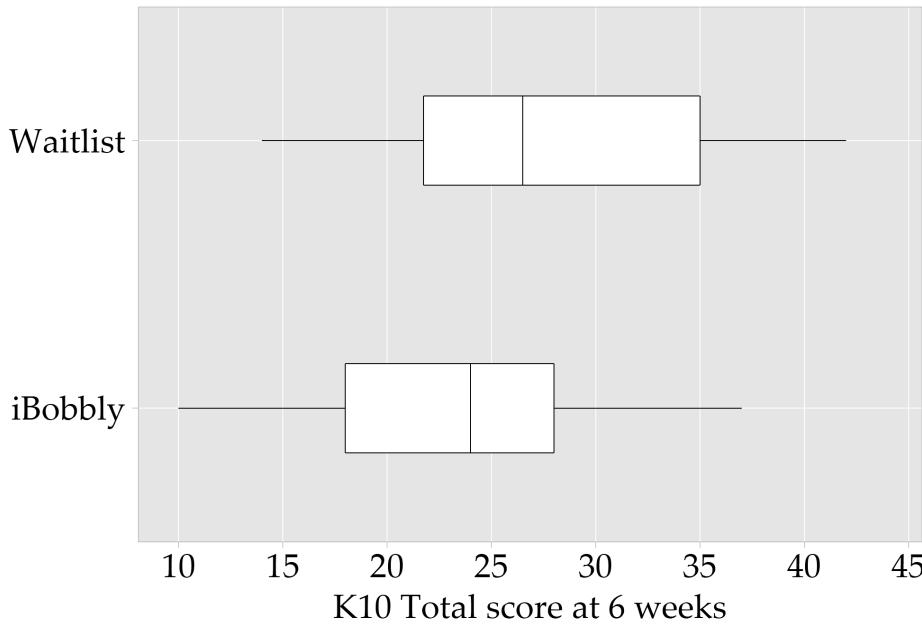


Figure 55: Boxplots of K10 total scores after 6 weeks for the waitlist and iBobbly groups

Group	$n$	$\bar{x}$	$s$	$s^2$
Waitlist (W)	30	27.83	8.04	64.63
iBobbly (B)	31	23.55	7.76	60.26
Pooled		7.90	62.41	

The difference between the two sample means is 4.28; this is the point estimate of  $\mu_W - \mu_B$ .

A 95% confidence interval for  $\mu_W - \mu_B$ , assuming equal population variances, is:

$$27.83 - 23.55 \pm t_{59}(0.975) \times 7.90 \times \sqrt{\frac{1}{30} + \frac{1}{31}}, \quad \text{or } 4.28 \pm 4.05.$$

The 95% confidence interval is therefore (0.24, 8.33).

### Not assuming equal population variances

If we are not prepared to make the assumption that the true population variances are equal, the standard error is based on the separate estimates of variance from each group.

If  $\sigma_1^2 \neq \sigma_2^2$  an approximate 95% confidence interval is given by

$$\bar{x}_1 - \bar{x}_2 \pm t_m(0.975) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where  $s_1^2$  and  $s_2^2$  are the usual estimates of  $\sigma_1^2$  and  $\sigma_2^2$ , and  $m$  is given by a

complicated formula;  $m$  will always be between the smaller of  $n_1 - 1$  and  $n_2 - 1$ , and  $n_1 + n_2 - 2$ .

In MINITAB: Stat > Basic Statistics ▶ 2-sample t, and *do not* check the box labelled Assume equal variances.

▷ **EXAMPLE.** For the iBobbly trial, a 95% confidence interval for  $\mu_W - \mu_B$ , without assuming equal population variances, is (0.23, 8.34).

### Considering the assumption of equal population variances

As this example illustrates, if the spread in the two samples is similar, the two approaches give correspondingly similar results.

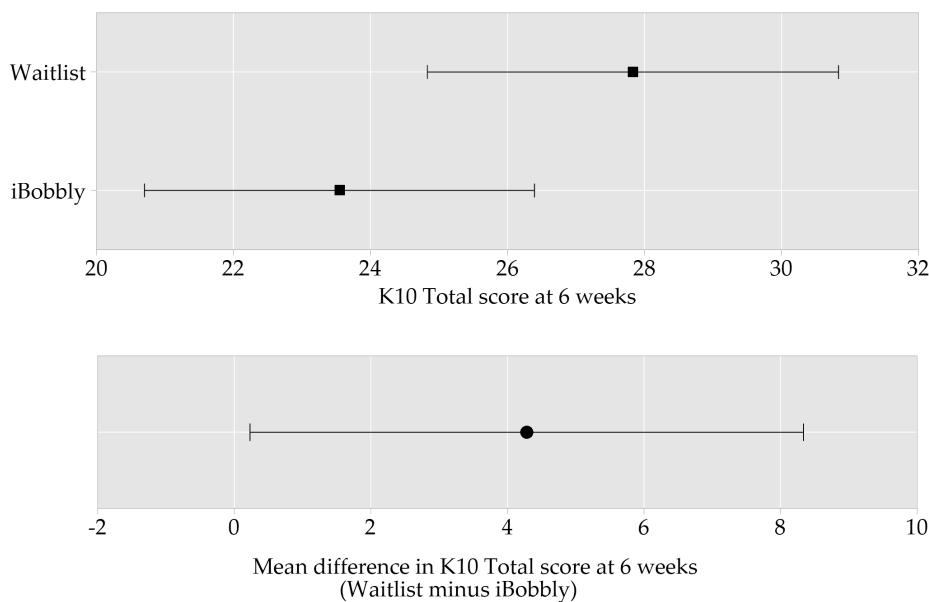


Figure 56: Graph showing the confidence interval for the difference between the two means in the iBobbly pilot study. Also shown are separate confidence intervals for the mean of each group.

Figure 56 shows the confidence interval for the difference between the two means. The two separate confidence intervals at the top of the figure are valid for their purpose, but they do not provide directly an inference about the difference between the two means.

If in doubt it is safer to assume that  $\sigma_1^2 \neq \sigma_2^2$ . It is therefore reasonable to ask: why bother with the method that assumes equal variances at all? The answer is that the context of two independent samples is a very simple and basic case, and for many extensions of this case, the equivalent of the unequal variances approach does not exist. This will be discussed further in the sections on linear models and analysis of variance.

Further, many other methods for inferences on differences in location between two groups, including non-parametric (distribution-free) approaches,

make the assumption that the only difference between the two groups is a shift in location, and, in particular, that the shape in the two groups is the same. This assumption of a location shift (only) implies that the variances in the two groups are the same.

## 5.3 The amazing Central Limit Theorem

### 5.3.1 The Theorem

We have already mentioned that many estimators have approximate Normal distributions. This means that in repeated sampling, a histogram of the values of an estimate looks like a Normal distribution.

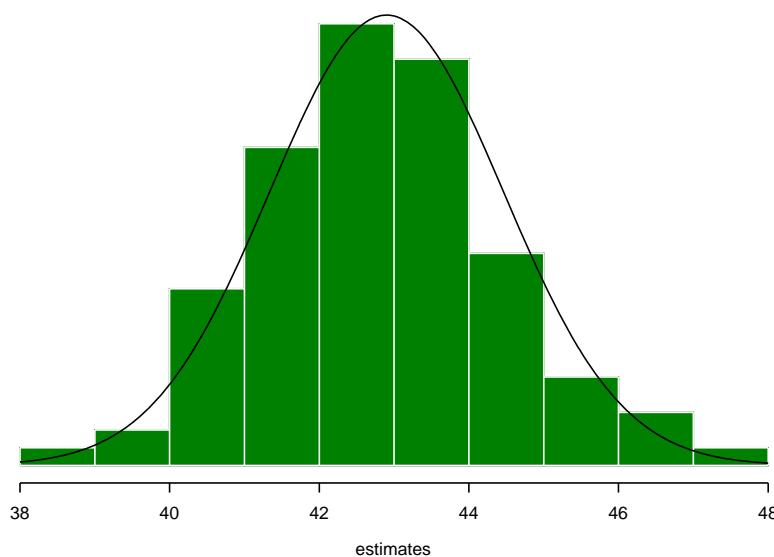


Figure 57: Histogram of the 100 sample estimates from Figure 46, with the limiting Normal distribution super-imposed.

For example, consider the hypothetical set of 100 samples in Figure 46. In Figure 57 a histogram of the 100 point estimates is shown. Note that its shape is approximately that of the Normal distribution.

The theoretical basis for this result is a truly remarkable theorem which is used throughout statistics. It is called the Central Limit Theorem, and it says that for large samples, the distribution of the sample mean is approximately Normal. In mathematical notation: if we have a random sample of size  $n$  from a population with mean  $\mu$  and variance  $\sigma^2$ , then as  $n$  grows large the distribution of  $\bar{X}$ , the sample mean, tends to a Normal distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

Before explaining this further, we need to be absolutely clear on what is meant by the “distribution of the sample mean”. The way to think about this distribution is to imagine an endless sequence of samples taken under identical conditions, in a single population. (°°) So Figure 46 is a bit like

this, except that in that Figure there is only a sequence of 100 samples, not an endless repetition. From this imagined sequence we could work out each sample mean, and then look at the distribution of those sample means. So this is like Figure 57, except that, again, this histogram is based on just 100 samples, not an endless repetition. If we went on for ever the histogram would become smoother and smoother and more and more bell-shaped until eventually it would approximate the shape of the Normal curve shown in Figure 57.

There are three important features of the distribution of the sample mean. The first is obvious. The first two are intuitively attractive, and are true no matter what the value of  $n$ , the sample size is. We derived them formally in section 4.5. The third is not even remotely intuitive, and it is the crux of the Central Limit Theorem.

- (a) The mean of the distribution of the sample mean is  $\mu$ . That is, the long-term average of the sample means, in the imagined sequence of samples, is  $\mu$ . In other words, the histogram of the sample means is positioned around  $\mu$ : i.e.  $E(X) = \mu$ .
- (b) The variance of the sample mean is  $\frac{\sigma^2}{n}$ :  $\text{var}(\bar{X}) = \frac{\sigma^2}{n}$ . While it is not obvious that the variance should be of this form, the form fits with the intuitive idea that we get more precise estimates from averages based on large samples than from small samples. So the spread of sample means in a long-run sequence based on samples of 1000 each time will be smaller than the spread of sample means in a long-run sequence of samples based on samples of size 50 each time.
- (c) Note that these first two points only give us the mean and variance of the distribution of the sample mean. In fact, we have already encountered these results in section 4.5. This is not enough to specify the *shape* of the distribution. But the Theorem says that, as  $n$  grows large, the shape of the distribution of the sample mean becomes closer and closer to a Normal distribution.

Why is this “amazing”? Because we have not made any restrictions on the shape of the distribution in the population, of the variable which we are averaging. It might be Normal, but it might not: it might be flat, triangular, badly skew, U-shaped, binary .... It doesn’t matter! This is the remarkable feature: that averages *from any shape of distribution* tend to have a Normal distribution. This is an extremely powerful result, because it means that we can use the Theorem to make inferences about a population mean *without even knowing the form of the underlying distribution of the data*.<sup>16</sup>

---

<sup>16</sup>There is a requirement that the population mean and variance are finite. But this is not likely to be a serious limitation in any practical context.

Even if this result is seen as amazing, its application may be thought to be somewhat limited. However, many estimation problems which do not involve means in the direct sense, turn out to have an underlying averaging process, and the Central Limit Theorem and its many cousins are applied throughout statistics. This relates to the second reason given for the importance of the Normal distribution in Chapter 4: estimators of parameters are often approximately Normally distributed:

$$\hat{\lambda} \stackrel{d}{\approx} N(\lambda, \sigma^2).$$

Some applications of the Theorem are shown below.

### 5.3.2 Confidence intervals for means

Consider the three situations mentioned in section 5.2. If the sample sizes are large, then similar working as in Section 5.2 can be used to find the confidence intervals concerned even if the population distributions are not Normal because of the Central Limit Theorem. The details are as follows.

(a) Confidence interval for the mean of a population

Suppose a random sample of size  $n$  is taken from a random variable  $X$  with mean  $\mu$ . If  $n$  is large, then

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

is an approximate 95% confidence interval for  $\mu$ .

(b) Confidence interval for the difference between two population means (paired samples)

Suppose  $n$  pairs of observations are taken from two random variables  $X_1$  and  $X_2$  with means  $\mu_1$  and  $\mu_2$  respectively. If  $n$  is large, then an approximate 95% confidence interval for  $\mu_1 - \mu_2$  is given by

$$\bar{d} \pm 1.96 \frac{s_D}{\sqrt{n}},$$

where  $\bar{d}$  and  $s_D$  are respectively the sample mean and sample standard deviation of the pairwise differences.

(c) Confidence interval for the difference between two population means (independent samples)

Suppose a random sample of size  $n_1$  is taken from a random variable  $X_1$  with mean  $\mu_1$  and variance  $\sigma_1^2$ , and a random sample of size  $n_2$  is taken from another random variable  $X_2$  with mean  $\mu_2$  and variance  $\sigma_2^2$ . Further suppose that  $X_1$  and  $X_2$  are independent, and that  $n_1$  and  $n_2$  are large.

(a) If  $\sigma_1^2 = \sigma_2^2$ , an approximate 95% confidence interval for  $\mu_1 - \mu_2$  is given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where  $s$  is the pooled standard deviation defined in Section 4.1.3.

- (b) If  $\sigma_1^2 \neq \sigma_2^2$  an approximate 95% confidence interval is given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

### 5.3.3 Confidence interval for a population proportion, $\theta$

This context is dealt with easily once we appreciate a basic point: *a proportion can be thought of as a mean of a binary outcome*. If we are interested in the proportion of subjects in a survey testing positive for hepatitis B, we can record the data as zeroes and ones, 1: positive, 0: negative. Then  $\hat{\theta}$ , the sample proportion, is just the mean of this variable: the numerator is the total of the 1s (i.e., the number of subjects who are positive) and the denominator, as for any sample mean, is  $n$ . So we need to know the population standard deviation of this variable, and then the above theory applies. If we define the random variable  $Y$  to take the value 1 if an individual is positive and 0 otherwise, then  $Y$  has a Binomial distribution with parameters 1 and  $\theta$ , in which case its variance is  $\theta(1 - \theta)$ . So the working above would suggest that an approximate 95% confidence interval for  $\theta$  would be  $\hat{\theta} \pm 1.96 \sqrt{\frac{\theta(1-\theta)}{n}}$ .

▷ **QUESTION:** What is the problem with implementing this formula for the confidence interval?

There are several ways to deal with this. The simplest is to substitute  $\hat{\theta}$  for  $\theta$  to obtain the standard error.

Expressing things formally:

Let  $X \stackrel{d}{=} \text{Bi}(n, \theta)$ . If  $n$  is large, then  $\frac{X}{n} \stackrel{d}{\approx} N(\theta, \frac{\theta(1-\theta)}{n})$ , by the Central Limit Theorem, and an approximate 95% confidence interval for  $\theta$  is given by

$$\frac{x}{n} \pm 1.96 \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}},$$

or, more succinctly, since  $\hat{\theta} = \frac{x}{n}$ ,

$$\hat{\theta} \pm 1.96 \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}.$$

This formula is so commonly required that it is worth remembering, since estimating a proportion is a very common research goal.

This confidence interval is one of the examples of the general form

$$\text{estimate} \pm (k \times \text{standard error}).$$

It is reasonable to ask: how large is “large”? For this case, a guideline often given is that both  $x$  and  $n - x$  should be greater than 10.

▷ **EXAMPLE.** Consider the political poll data used to discuss the fundamentals of inference. 443 out of 1000 polled said that they voted ALP. The sample estimated proportion is therefore 0.443, and an approximate 95% confidence interval for the true proportion is

$$0.443 \pm 1.96 \times \sqrt{\frac{0.443 \times 0.557}{1000}}, \text{ i.e. } 0.443 \pm 0.031, \text{ or } (0.412, 0.474).$$

This would often be expressed on a percentage scale, as an estimate of 44.3% and a 95% confidence interval of (41.2%, 47.4%).

Look at Figure 46. This formula was used to obtain the confidence intervals shown there.

This confidence interval can be obtained from MINITAB using: Stat → Basic Statistics → 1-Proportion ....

If the data for which you want to make an inference are stored as a column (of ones and zeroes, for example) then click on Samples in columns and identify the column. This will only work out a confidence interval if there are just two distinct values in the columns nominated (e.g. ones and zeroes). The dialogue box also allows you to enter Summarized data which means the number of trials (the denominator of the sample proportion) and the number of successes (the numerator).

The interval we have described above is based on a Normal approximation. MINITAB also provides an “exact” confidence interval; in fact, it provides the exact interval by default. We defer discussion of the method behind this interval until the hypothesis testing section of the notes. To get the Normal approximation, then, we need to click on Options ... and then tick the box: Use test and interval based on Normal distribution.

In the example, if we enter the summarized data and use the Normal approximation we get:

## Test and CI for One Proportion Method

p: event proportion

Normal approximation method is used for this analysis.

### Descriptive Statistics

N	Event	Sample p	95% CI for p
1000	443	0.443000	(0.412212, 0.473788)

### Test

Null hypothesis  $H_0: p = 0.5$

Alternative hypothesis  $H_1: p \neq 0.5$

On the other hand, if we do not use the Normal approximation, we get:

## Test and CI for One Proportion Method

p: event proportion

Exact method is used for this analysis.

### Descriptive Statistics

N	Event	Sample p	95% CI for p
1000	443	0.443000	(0.411919, 0.474418)

### Test

Null hypothesis  $H_0: p = 0.5$

Alternative hypothesis  $H_1: p \neq 0.5$

With such large numbers, the two methods agree for practical purposes. In both cases, in percentage terms, the 95 confidence interval is (41.2%, 47.4%).

On the other hand, suppose a random sample survey of a group of animals in their natural habitat had been taken and 4 out of the 17 animals had a particular disease. The sample proportion is  $0.235 = \frac{4}{17}$ . Using the Normal approximation:

## Test and CI for One Proportion Method

p: event proportion

Normal approximation method is used for this analysis.

### Descriptive Statistics

N	Event	Sample p	95% CI for p
17	4	0.235294	(0.033654, 0.436934)

### Test

Null hypothesis  $H_0: p = 0.5$

Alternative hypothesis  $H_1: p \neq 0.5$

Using the exact method:

## Test and CI for One Proportion Method

p: event proportion

Exact method is used for this analysis.

### Descriptive Statistics

N	Event	Sample p	95% CI for p
17	4	0.235294	(0.068108, 0.498993)

### Test

Null hypothesis  $H_0: p = 0.5$

Alternative hypothesis  $H_1: p \neq 0.5$

The exact method is the “gold standard”. It gives (0.07, 0.50), which is substantially different from the Normal approximation (0.03, 0.44).

As noted above, a general guideline is that both the numerator of the sample proportion (the number of successes) and the difference between the denominator and numerator (the number of failures) should be greater than 10 to use the Normal approximation.

▷ **QUESTION:** Consider a sample proportion of  $\frac{2}{17}$ , and find a 95% confidence interval using the Normal approximation. Is it symmetric? What is going on at the lower end of the interval?

▷ **QUESTION:** Now consider a sample proportion of  $\frac{0}{17}$ , and try to find a 95% confidence interval using the Normal approximation. What do these results indicate?

▷ **QUESTION:** Find the exact confidence intervals for  $\frac{2}{17}$  and  $\frac{0}{17}$ .

▷ **QUESTION:** If the data really were diseased animals, what issue might arise about the applicability of the technique?

It is sensible to ask: why use the Normal approximation if we have the exact method easily available to us? The answer is: we shouldn't, if the exact method is available. But it is often not available.

#### 5.3.4 Confidence interval for difference between two population proportions, $\theta_1 - \theta_2$

Let  $X_1 \stackrel{d}{=} \text{Bi}(n_1, \theta_1)$  and  $X_2 \stackrel{d}{=} \text{Bi}(n_2, \theta_2)$ . If  $n_1$  and  $n_2$  are large, then

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \stackrel{d}{\approx} N(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2})$$

and an approximate 95% confidence interval for  $(\theta_1 - \theta_2)$  is given by

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm 1.96 \sqrt{\frac{\frac{x_1}{n_1}(1-\frac{x_1}{n_1})}{n_1} + \frac{\frac{x_2}{n_2}(1-\frac{x_2}{n_2})}{n_2}}$$

For the comparison of two proportions in MINITAB, you should use Stat → Basic Statistics → 2-Proportions .... The set-up is very similar, but there is no exact method.

▷ **EXAMPLE.** Suppose that  $\frac{443}{1000}$  said that they would vote ALP in a random sample in Sydney, and a similar random sample in Melbourne gave  $\frac{389}{1000}$ . We are interested in an estimate and 95% confidence interval for the difference between the two proportions.

## Test and CI for Two Proportions

### Method

$p_1$ : proportion where Sample 1 = Event

$p_2$ : proportion where Sample 2 = Event

Difference:  $p_1 - p_2$

### Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	1000	443	0.443000
Sample 2	1000	389	0.389000

### Estimation for Difference

Difference	95% CI for
	Difference
0.054	(0.010862, 0.097138)

*CI based on normal approximation*

### Test

Null hypothesis  $H_0: p_1 - p_2 = 0$

Alternative hypothesis  $H_1: p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	2.45	0.014
Fisher's exact		0.016

We could report this as follows: the estimated difference between Sydney and Melbourne in the percentages intending to vote ALP was 5.4%, with a 95% confidence interval of (1.1%, 9.7%).

### 5.3.5 A general form for confidence intervals

We have now seen quite a few examples of a general form of confidence interval. In many situations we are interested in estimating a parameter (or a parametric function),  $\lambda$  (say), and we have an estimator,  $U$  (say), whose distribution is at least approximately Normal, due to the Central Limit Theorem. Further, in many cases the estimator  $U$  is unbiased, so that  $E(U) = \lambda$ , and we can find the variance of  $U$ ,  $\sigma_U^2$ . It then follows that a 95% confidence interval for  $\lambda$  is at least approximately given by  $u \pm 1.96\sigma_U$ .

Often  $\sigma_U$  will be a function of one or more unknown parameters. These need to be replaced by estimates, giving us the standard error of the estimate,

$\text{se}(U)$ . This gives us the generic form:

$$u \pm 1.96 \text{se}(U).$$

Examples of this we have seen are the large sample, approximate confidence intervals: for a single proportion and a difference between two independent proportions, and for means when we do not assume Normality for the data. But there are many other examples: log odds ratios, log hazard ratios, regression slopes and others. Since many important estimators of interest are, at bottom, based on averaging, the Central Limit Theorem applies and we have an approximate confidence interval for “large samples” available to us, of this form.

The wide applicability of this is the second reason for the importance of the Normal distribution.

If the distribution of the data is Normal and we are estimating means, the standard error will involve estimated variances and instead of a large sample result based on the Central Limit Theorem, we have confidence intervals of a similar form for any sample size, based on the application of the  $t$  distribution.

Examples of this we have seen are the confidence intervals for means and differences of means from Normally distributed data (without requiring a large sample).

In summary, many confidence intervals (although not all!) are of the form

$$\text{estimate} \pm k \times (\text{standard error}),$$

where  $k$  is obtained from an appropriate distribution.

## 5.4 Confidence interval for a correlation, $\rho$

We have considered the sample correlation,  $r$ , in Chapter 2. We may wish to use a sample correlation coefficient to obtain a confidence interval for the corresponding population correlation coefficient, which is usually labelled  $\rho$ .

In fact, we have already seen an application of this, in the pre- and post-tuning data used in section 5.2.3, where the correlation was found to be  $r = 0.80$ , with a 95% CI of (0.57, 0.91).

There is a Central Limit Theorem result for the sample correlation, which should not be surprising, because its formula has averaging at its core. But because  $r$  and  $\rho$  are between  $-1$  and  $+1$ , the distribution of  $R$  is very skew and the sample size needs to be extremely large for the normal approximation to be adequate. So this approach is not used.

In other words, we do not have a reasonable option for an approximate con-

fidence interval of the form estimate  $\pm k \times$  standard error, for correlations. Rather, a transformation is used to obtain an approximate result in large samples (but  $n$  does not have to be extremely large). The formal result is as follows.

We assume that we have a random sample from a bivariate distribution  $(X, Y)$ , made up of pairs  $(X_i, Y_i), i = 1, 2, \dots, n$ . We calculate the sample correlation,  $R$ .  $R$  is a random variable, with a distribution.

If  $(X, Y)$  has a bivariate Normal distribution and  $n$  is large then

$$Z = \frac{1}{2} \ln \frac{1+R}{1-R} \stackrel{d}{\approx} N\left(\frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right).$$

In fact,  $Z = \tanh^{-1} R$ , or  $R = \tanh Z$ ;  $\tanh$  is a hyperbolic trigonometric function (the hyperbolic tangent) and  $\tanh^{-1}$  is its inverse.

We use this result to first find an approximate 95% CI for  $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ , and then transform back to the  $\rho$  scale to get an approximate 95% CI for  $\rho$ .

One desirable feature of this approach is that we always get confidence limits that are inside the interval  $(0,1)$ .

This is a technique that arises in other contexts, beyond the scope of this course. Ratio measures such as the relative risk, odds ratio and hazard ratio have approximately Normal distributions for very large sample sizes. But inferences can be obtained for modest sample sizes (while still ‘large’) by first finding a function of the estimator that approaches Normality faster than the estimator does itself. Inference is obtained on the transformed scale (e.g. the log scale) and then back-transformed to get the inference of real interest.

▷ **EXAMPLE.** iBobbly intervention for suicide prevention

In the iBobbly study, the researchers measured K10 and the “PHQ-9” at baseline, on all 61 participants. The PHQ-9 is a measure of depression severity, so we could reasonably expect it to be correlated with the K10 variable, an overall measure of distress based on questions about anxiety and depressive symptoms over the past four weeks.

The following plot shows the association between these two variables, the sample correlation, and the 95% confidence interval for  $\rho$  using the method outlined above.

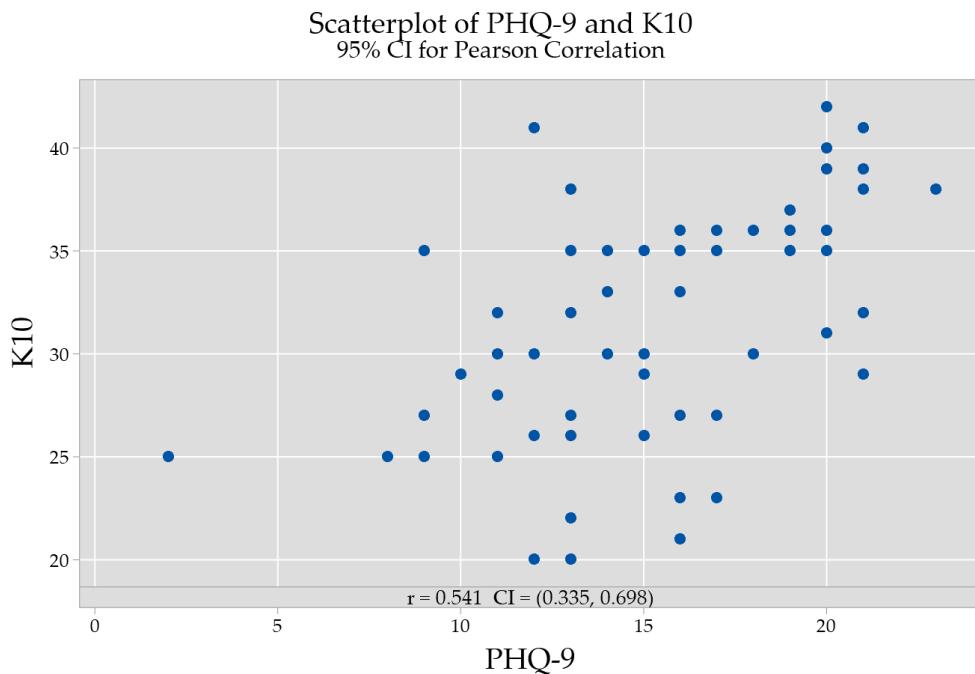


Figure 58: Scatterplot of PHQ-9 versus K10 at baseline, iBobbly study, showing the sample correlation and 95% confidence interval.

For these data  $r = 0.54$  and the 95% confidence interval for  $\rho$  is  $(0.34, 0.70)$ .

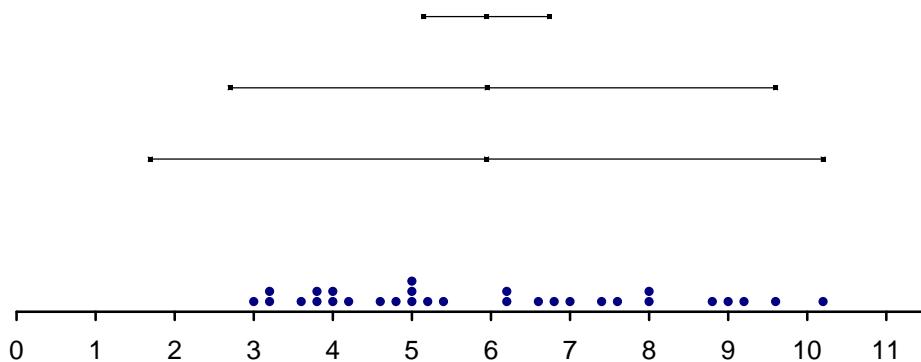
In MINITAB: Stat > Basic Statistics ▶ Correlation. It is common to seek all the correlations between a number of variables, and their associated inferences. To get this, choose Pairwise correlation table in the Results. There are also different options to display with the graphs; see the Options dialogue box.

Note that the 95% confidence interval for a correlation is not, in general, symmetric around  $r$ . This makes sense. Suppose we find that  $r = 0.99$ . For a 2-sided confidence interval, the upper limit must be less than 1, and hence must be within 0.01 of  $r$  in this case. But we could reasonably expect more uncertainty on the lower side, and hence the lower limit of the confidence interval will be further away from 0.99 than the upper limit. For estimates that are close to  $-1$ , the same feature is found, but with the effects on the limits reversed.

This has a direct analogy with the exact confidence interval for a population proportion  $\theta$ ; corresponding phenomena occur for sample proportions that are close to zero or one.

## 5.5 Exercises

- 5.1 A 95% confidence interval for a parameter is such that it contains the unknown parameter with probability 0.95 (see Section 4.1). We call this a “success”. So, the probability that a 95% confidence interval is successful is 0.95. And it is a failure (i.e. does not contain the parameter) with probability 0.05.
- (a) Suppose we have four independent 95% confidence intervals. Show that the probability that all four of these intervals are successful is 0.8145.
  - (b) (a) Suppose we have 20 independent 95% confidence intervals, what is the probability that all 20 are successful?
  - (b)\* How many of these intervals do you ‘expect’ to be successful?
  - (c)\* What is the distribution of the number of successful intervals?
  - (d)\* Using iii., find the probability that the number of successful intervals is equal to 20? 19? 18?
  - (c) If the number of independent 95% confidence intervals is increased to 100, what is the probability that all 100 intervals contain the true value of the parameter?
  - (d)\* Show that the probability that four independent  $100Q\%$  confidence intervals are all successful is  $Q^4$ . Hence show that if we have four independent 98.73% confidence intervals, the probability that all are successful is 0.95.
- 5.2 The cheddar cheese data was introduced in Chapter 2. As a reminder, there were 30 samples of cheese. “Taste” is the outcome variable of interest. The taste scores were obtained by combining the scores from several tasters. Taste scores could range from 0 to 60; a higher score reflects a more positive taste rating. Three of the chemicals whose concentrations were measured were acetic acid, hydrogen sulfide and lactic acid. For acetic acid and hydrogen sulfide (natural) log transformations were taken. The data are stored as `cheese.mwx`.
- (a) First consider the hydrogen sulphide data, for which  $\bar{x} = 5.94$  and  $s = 2.13$ . The figure below shows a dotplot of the data, and three intervals. The midpoint of each interval is the mean.



One of the intervals (lines) on the figure is the 95% confidence interval for the true mean for the log of hydrogen sulphide. Which one? Explain your choice.

For the next questions, consider only the variable `lactic`, which is the concentration of lactic acid in the samples.

- (b) Draw a dotplot and a histogram of the data (using the software). Does it look like it might have come from a Normal distribution?  
*Later in the course we are going to look at more formal ways of assessing whether the underlying distribution is Normal or not.*
- (c) Assuming that the data comes from an underlying Normal distribution, (°) find a 95% confidence interval for  $\mu$ , the average lactic acid level in samples from this cheese.
- (d) Will a 50% confidence interval for  $\mu$  be wider, or narrower, than the 95% confidence interval? Find a 50% confidence interval for  $\mu$ .  
[ You will need to click on Options and change the Confidence level to 50%. ]
- (e)\* What would happen if the confidence level was made even smaller, 20% say? What is the 0% confidence interval?
- (f) Find a 99.9% confidence interval for  $\mu$ .
- (g) A food scientist says: "I understand that it is desirable for the average lactic acid level in that cheese to be 1.8 units." Consider the 95% confidence interval; do you consider the cheese consistent with this standard? What would be your conclusion if you had used the 99.9% confidence interval?
- (h) A colleague produces a draft report and attempts a brief explanation of a 95% confidence interval that you have calculated. Which do you prefer?
  - A. "It's a range that contains 95% of our data."

- B. "The chance that the next sample mean falls in the first 95% confidence interval is 0.95."
- C. "There's a 95% chance that it contains the true parameter value."
- D. "There is a 95% chance that the sample mean falls in the 95% confidence interval."

Explain your choice.

- (i)\* What issues are there to be considered in recruiting individuals to rate the taste of the cheese?

5.3 An experiment is conducted involving 84 children divided into 42 pairs. The members of each pair are carefully matched so that they are similar with respect to sex, age, IQ, and are also matched on their scores on a mathematical knowledge test. They have not yet studied multiplication. One member of each pair is randomly chosen and taught multiplication using method A. The other member of the pair is taught multiplication using method B. Test scores for the 84 children after the learning experience are stored in iq.mwx .

pair	1	2	3	4	5	...	42
Method A	51	46	30	54	24	...	31
Method B	38	44	31	46	21	...	28

- (a) What is the research question that is being addressed here?
- (b) Does the design of the study seem appropriate? Explain why or why not.
- (c) Produce an appropriate visual display (or displays) of the data. What do you conclude from the display(s)?  
 (You can calculate the difference between each pair of children using Calc > Calculator: store the result in C4; type C2–C3 in the Expression box; click OK. Alternatively, you may find it easier to type the command let C4=C2–C3 in the Session Window.)
- (d) Find an estimate and a 95% confidence interval for  $\mu_A - \mu_B$ , the difference between the mean scores obtained by children taught by the two methods.  
 [Stat > Basic Statistics > Paired t with Each sample in a column, select Method A for the First sample and select Method B for the Second sample; click **OK**. Alternatively, you can work directly on the difference stored in column 4 and do the following: Stat > Basic Statistics > 1-Sample t and One or more samples, each in a column, select C4; click **OK**.]
- (e) Based on this interval, what do you conclude about the two methods? Would you recommend one method over the other? Explain why or why not.

- (f)\* In the experiment, Method A is taught by Ms. Potts and Method B is taught by Mr. Pan. Is this a strength or a weakness of the study? Explain.
- 5.4 The National Health and Nutrition Examination Survey is an important longitudinal study designed to assess the health and nutritional status of adults and children in the USA. The first wave of the study was in 1971. Open the MINITAB worksheet NHANES.mwx. Each row represents data from one subject in the survey. The data in this file is an extract from the full dataset. The most important aspect of this is that all subjects in this dataset were *current cigarette smokers* at the first wave of the study in 1971. Some follow up information from interviews in 1982 is also included.
- (a) Examine the data file.  
Does this data file have data from children?
  - (b) Find the variable called cholesterol. This is the serum cholesterol, in mg/100 ml, measured in 1971. The units are also commonly described as mg/decilitre, or mg/dL. It is said to be desirable for your level of serum cholesterol to be less than 200 mg/dL. Using only basic descriptive information, what can you say about the percentage of NHANES smokers in 1971 that have a serum cholesterol greater than 200?
  - (c) The variable “race” codes for ethnicity in a very simple way. There are two groups, ‘White’ and ‘Black/Other’. <sup>17</sup> Produce a plot that compares the distribution of cholesterol for smokers in the two racial groups.
  - (d) What is the difference ('White' – 'Black/Other') in mean serum cholesterol (mg/100 ml) for current smokers?
  - (e) Without assuming equal population variances, find a 95% confidence interval for the true difference in means, assuming the data in each group come from a normal distribution.
  - (f) Now carry out the same analysis restricted to smokers whose birthplace was Florida. Florida has the code '12' for birthplace. You can isolate the Florida subjects by using Data > Subset Worksheet ....

<sup>17</sup>The detailed notes for the 1971 survey are as follows, for ‘race’: “The race of the respondent was marked by observation and it was assumed that the race of all related persons was the same as the respondent unless otherwise learned. The race categories were “White”, “Negro” or “other”. If the appropriate category could not be marked by observation, then race was asked. Persons of race other than White or Negro, such as Japanese, Chinese, American Indian, Korean, Hindu, Eskimo, etc. were reported as “Other”. Mexicans were included with “White” unless definitely known to be American Indian or of other nonwhite race.”

- (g) Your friend asks “How come the confidence interval for Florida is so much wider than the one for the whole of the USA?” Provide a good answer.
- (h) What assumptions are made in the construction of these confidence intervals?
- 5.5 A study was undertaken to assess the effect of intake of paracetamol-containing analgesics (e.g. Panadol) on kidney function and other health parameters. A group of women were identified from city workplaces, with high intake of paracetamol-containing analgesics. The level of NAPAP (N-acetyl-P-aminophenyl) in urine was used as a marker of paracetamol intake. This constitutes the “study” group. A second group were identified from the same workplaces and with normal NAPAP levels, who had low or no paracetamol intake. The women were examined at baseline in 1995 and also in 2005, and had their kidney function evaluated by several laboratory tests.
- The data set renal.mwx contains the data on serum-creatinine levels (an important index of kidney function) for both the study group and the control group.

<code>id</code>	<code>age</code>	<code>group</code>	<code>creat_1995</code>	<code>creat_2005</code>	<code>diff (2005-1995)</code>
1	41	2	0.97	1.00	0.03
2	47	2	0.88	1.12	0.24
3	48	1	1.48	0.75	-0.73
4	42	1	0.78	.	.
5	43	2	0.96	0.85	-0.11
6	49	1	0.79	0.95	0.16
7	43	1	0.80	1.20	0.40
8	44	2	0.74	1.10	0.36
9	49	1	0.84	0.97	0.13
10	39	2	0.85	0.87	0.02
:	:	:	:	:	:

The control group is labelled group 1, the study group is group 2.

- (a) (a) Are the 1995 study group and 1995 control group paired, or independent samples? How would you compare these groups?
- (b) Are the 1995 study group and 2005 study group samples paired, or independent samples? How would you compare these groups?
- (b) Do the two groups have different serum-creatinine profiles at baseline?
- Find a 95% confidence interval for the difference in the mean levels of the study and control group at baseline.

- 5.6 A colleague has analysed the data from Problem 5.3, and shows you the output below.

## Two-Sample T-Test and CI: Method A, Method B

$\mu_1$ : mean of Method A

$\mu_2$ : mean of Method B

Difference:  $\mu_1 - \mu_2$

*Equal variances are not assumed for this analysis.*

### Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Method A	42	38.48	7.67	1.2
Method B	42	33.17	9.02	1.4

### Estimation for Difference

Difference	95% CI for
	Difference
5.31	(1.67, 8.95)

### Test

Null hypothesis  $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2.91	79	0.005

- (a) Compare the point estimate of the mean difference in test scores that you obtained (in Problem 5.3) with the result your colleague found.
  - (b) Compare the 95% confidence interval for the mean difference in test scores that you obtained (in Problem 5.3) with the result your colleague found.
  - (c) Why are the results different? (Examine the output and consider the assumption your colleague has made about the data.)
  - (d) Which analysis is more appropriate? Explain why.
- 5.7 An investigator wants to test a new eye-drop that is supposed to prevent allergic ocular itching. To study the drug she uses a contralateral design, in which one eye gets the active drug (*A*) and the other gets a placebo (*P*). For each subject the eye receiving the active drug is assigned randomly. The subjects use the eye-drops three times a day

for a week and then report their degree of itching in each eye (0=none, 1=mild, 2=moderate, 3=severe) without knowing which eye-drop is used in each eye. 100 subjects are recruited into the study. The results are given in `eye.mwx`.

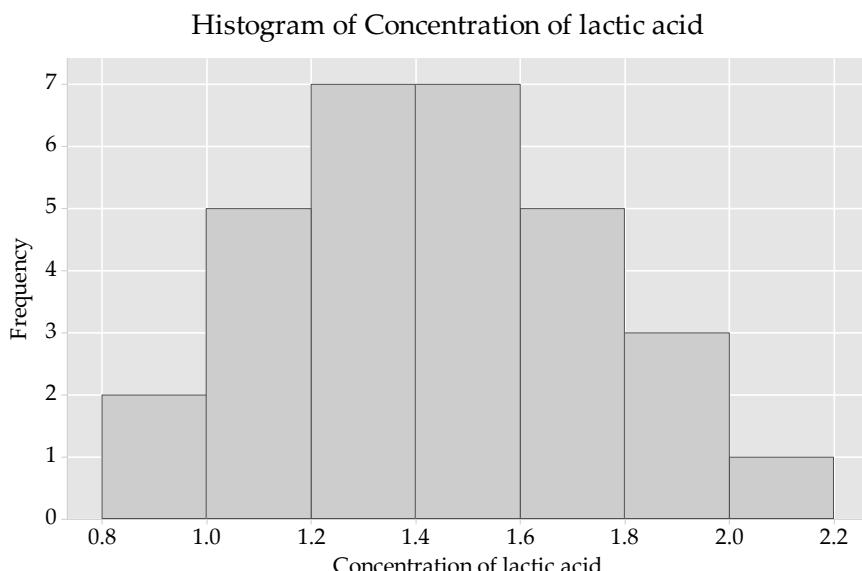
drug applied to left eye	A-eye	P-eye	diff (A-P)	sign
A	2	3	-1	-1
P	1	2	-1	-1
P	0	2	-2	-1
A	3	2	+1	+1
P	0	1	-1	-1
A	1	2	-1	-1
A	1	2	-1	-1
A	1	1	0	0
P	3	3	0	0
:	:	:	:	:

Find an estimate and a 95% confidence interval for

- (a) the probability that the eye receiving drug A is scored better than the eye receiving the placebo;  
 [Obtain a frequency count of the sign of the difference using: Stat > Tables > Tally Individual Variables ... for variable sign, and hit **OK**.]  
*Note that A-eye is better than P-eye if sign is negative.*
- (b) the probability that both eyes are given the same score;
- (c) the probability that the eye receiving drug A is scored no worse than the eye receiving the placebo.

## 5.6 Answers

- 5.1 (a) As the four confidence intervals are independent, we can multiply the individual success probabilities to find the joint success probability:  $0.95^4 = 0.8145$ .
- (b) (a)  $0.95^{20} = 0.3585$
- (b)\* As the confidence coefficient is 0.95, we would expect 95% of the confidence intervals to be successful on average. This is 19 of the 20 confidence intervals.
- (c)\* Let  $X$  be the number of successful 95% confidence intervals in a set of  $n$  confidence intervals.  $X \stackrel{d}{=} \text{Bi}(n, 0.95)$ .
- (d)\*  $\Pr(X = 20) = 0.358, \Pr(X = 19) = 0.377, \Pr(X = 18) = 0.189$
- (c)  $\Pr(X = 100) = 0.0059$ .
- (d)\* Let  $X$  be the number of successful  $(100Q)\%$  confidence intervals in a set of  $n$  confidence intervals.  $X \stackrel{d}{=} \text{Bi}(n, Q)$ . For any one interval, the probability of success is  $Q$ . When events are independent, their combined probability is the product of the individual probabilities. Hence the probability that all four intervals are successes is  $Q \times Q \times Q \times Q = Q^4$ . If  $Q = 0.9873$  then  $Q^4 = 0.95$ .
- 5.2 (a) The top line is the 95% confidence interval. The confidence interval is a statement about the unknown population mean. It is not a statement about the sample, or about an item selected from the population.  
The bottom line on the figure shows  $\bar{x} \pm 2s$ ; it covers all but the last observation. The middle line is the shortest interval, symmetric around the mean, that includes at least 95% of the data. Neither of these lines represents a 95% confidence interval for the true population mean.
- (b) Both a dotplot and a histogram will give you at least some idea of whether the data are consistent with an underlying normal distribution. But it takes a bit of experience to make good judgements about this visually, especially in small samples. The histogram is symmetric and roughly bell-shaped; a sample like this could be drawn from an underlying normal distribution.



The histogram looks reasonably consistent with an underlying normal distribution.

You can use: Graph > Histogram ▶ With fit. A Normal distribution is superimposed on the histogram.

- (c) Stat > Basic Statistics ▶ 1-sample t; select lactic for One or more samples, each in a column, then click OK. The 95% CI is (1.33 to 1.56).
- (d) Narrower: it will give us an interval within which we are less confident that the true mean lies. The 50% CI is (1.40 to 1.48).
- (e)\* A 20% confidence interval for the true mean lactic acid level is (1.42,1.46).

A 0% confidence interval is 1.44; this is the point estimate of the true mean lactic acid level.

- (f) A 99.9% confidence interval for the true mean lactic acid level is (1.24 to 1.64).
- (g) These data are quite inconsistent with this standard, in the sense that we are 95% confident that the average lies between 1.33 and 1.56, and the upper limit of this interval is considerably less than 1.8. If we increase our confidence to 99.9%, the interval gets wider: (1.24 to 1.64); even so, the value of 1.8 is still well outside the interval.

- (h) Choice A is not correct as the confidence interval does not refer to the range of the data.

Choice B is not correct as the confidence coefficient refers to the probability of including the true population mean in the long run, not in the 'next' interval.

Choice D is incorrect because the confidence interval was calculated with the sample mean in the centre; there is a 100% chance

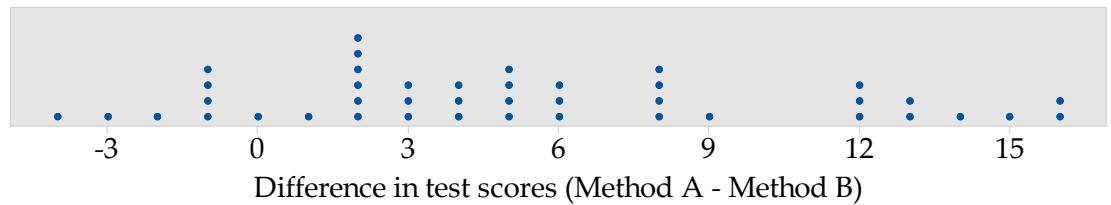
that the sample mean is in the confidence interval.

Answer C is correct.

- (i)\* Being a ‘taste tester’ can require skilled judgements and food companies often use individuals from a ‘taste panel’ who have been trained to use rating scales to identify a variety of different characteristics of the food of interest. Recruiting untrained individuals (people in the street or in a supermarket, for example) could introduce unwanted variation in the ratings.

- 5.3 (a) Is there a difference between effectiveness of the two teaching methods?
- (b) The study design seems reasonable provided the children used are a random sample from the population about whom conclusions wish to be drawn. It would also be important to know how closely matched the children were on the mathematical knowledge test.
- (c) As the data are paired, a dotplot of the differences between the pairs is appropriate. From the dotplot, 35 of the 42 differences are positive and the seven negative differences are all quite small. It appears that method A tends to result in higher scores.

To plot the differences use Graph > Dotplot ► Simple, and select difference.



An analysis of the differences between the two teaching methods using a one sample  $t$  procedure is appropriate.

- (d) The relevant output is:

**Paired T-Test and CI: Method A, Method B**

**Descriptive Statistics**

Sample	N	Mean	StDev	SE Mean
Method A	42	38.48	7.67	1.18
Method B	42	33.17	9.02	1.39

**Estimation for Paired Difference**

Mean	StDev	SE Mean	95% CI for
			$\mu_{\text{difference}}$
5.310	5.403	0.834	(3.626, 6.993)

$\mu_{\text{difference}}: \text{mean of (Method A - Method B)}$

**Test**

Null hypothesis  $H_0: \mu_{\text{difference}} = 0$

Alternative hypothesis  $H_1: \mu_{\text{difference}} \neq 0$

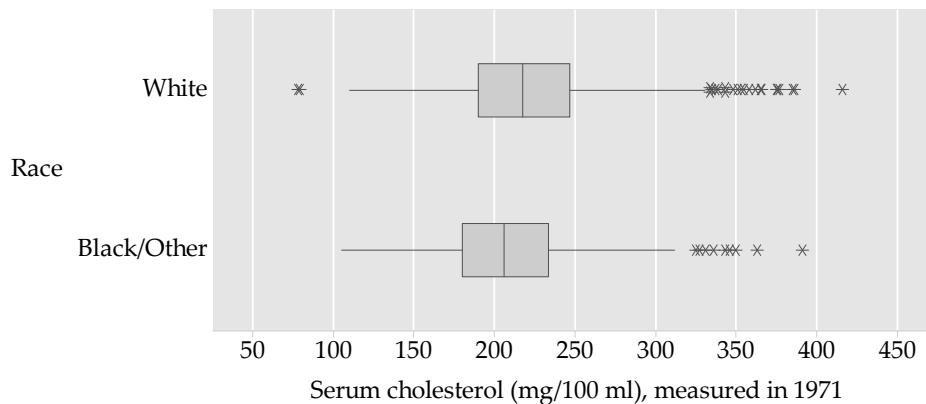
T-Value P-Value

6.37 0.000

The estimated mean dif-

ference in test scores (method A – method B) is 5.31; the 95% confidence interval is (3.63, 6.99).

- (e) Based on the confidence interval the true difference in the means of the two methods may be around 3.6, and could be as high as 7, with method A tending to give higher scores than those for method B. An understanding of the multiplication test is needed to be able to say whether the (estimated) difference between the methods is of any practical importance.
  - (f)\* The study aimed to compare the effectiveness of the different teaching methods. However the teaching methods were taught by different people; hence the ‘treatment’ experienced by the different groups of children involved both different methods and different teachers. Ms. Potts may simply be a better teacher than Mr. Pan. This is a weakness of the study.
- 5.4 (a) There are various ways to explore the data to answer this question. You could use a histogram or a boxplot to examine the distribution of age. You could obtain summary statistics. As the minimum age is 25 years, there are no children in the data set.
- (b) The median serum cholesterol was 216 mg/100 ml, and the lower quartile was 188.75. This means that the percentage of NHANES smokers with a serum cholesterol greater than 200 must be between 50% and 75%.
- (c) Given the relatively large sample sizes, boxplots are appropriate.



- (d) The output from a two-sample procedure for comparing means is below.

## Two-Sample T-Test and CI: Serum cholesterol (mg/100ml), race Method

$\mu_1$ : mean of Serum cholesterol (mg/100ml) when race = White

$\mu_2$ : mean of Serum cholesterol (mg/100ml) when race = Black/Other

Difference:  $\mu_1 - \mu_2$

*Equal variances are not assumed for this analysis.*

### Descriptive Statistics: Serum cholesterol (mg/100ml)

race	N	Mean	StDev	SE Mean
White	1491	220.9	45.2	1.2
Black/Other	235	210.8	46.8	3.1

### Estimation for Difference

Difference	95% CI for Difference	
10.10	(3.66, 16.53)	

### Test

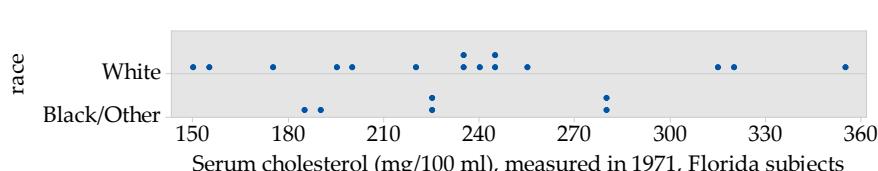
Null hypothesis  $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
3.09	306	0.002

The mean serum cholesterol was 10.1 mg/100 ml higher for smokers classified as 'White', compared with smokers classified as 'Black/Other'.

- (e) From the output above, the 95% confidence interval for the true difference in means, assuming the data in each group come from a normal distribution, is for 'White' - 'Black/Other' (3.7 to 16.5) mg/100 ml, or equivalently for 'Black/Other' - 'White' (-16.5 to -3.7) mg/100ml.
- (f) The results for Florida are shown below:

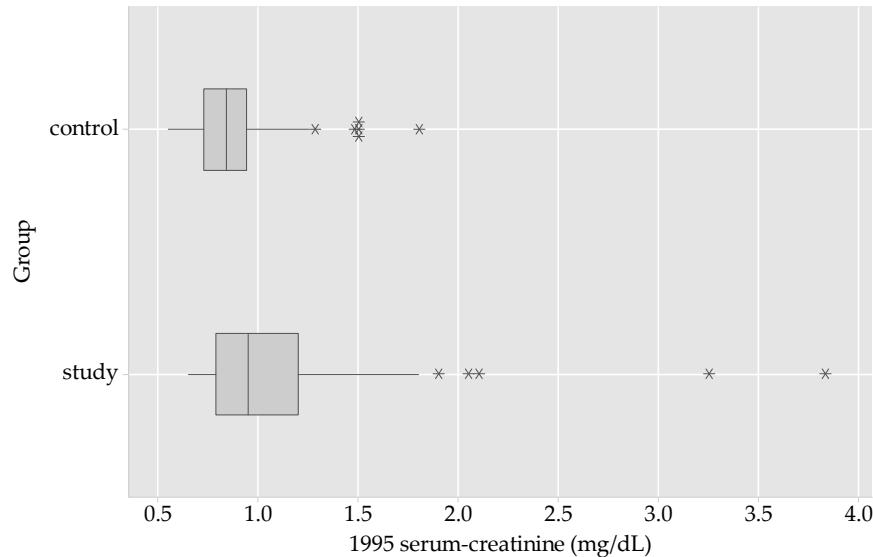


A dotplot is preferable for these small sample sizes.

For the Florida subjects, the difference in means ('White' – 'Black/Other') for smokers is 5.7 mg/100 ml; the 95% confidence interval is (-44.1 to 55.6) mg/100 ml.

- (g) The width of a confidence interval depends on the sample size and the variability in each of the groups being compared. When the whole data set is considered, there are nearly 1500 in the 'White' group and over 200 in the 'Black/Other' group. When the data for Florida are considered, there are 15 in the 'White' group and 6 in the 'Black/Other' group. Hence there will be less uncertainty in the estimates of the means and the difference in means for the entire data set than for Florida.

- (h) To draw an inference from these confidence intervals to a population, we need to assume that the NHANES survey sampled at random from the correspondingly defined populations (adults aged over 25 who are current smokers in 1971 etc.), without biases such as selection and response bias.<sup>18</sup> The analysis assumes the data in each group are a random sample from a normal distribution. No assumption is made about the equality of population variances.
- 5.5 (a) (a) The two groups (study and control) are comprised of different individuals; they are classified according to their paracetamol intake. This is an independent samples design. The groups can be compared using a confidence interval for the difference in mean serum-creatinine levels.
- (b) The 1995 study group and the 2005 study group are the same individuals, measured on two different occasions. This is a paired samples design. The groups can be compared using a confidence interval for the mean difference in serum-creatinine levels between 1995 and 2005.
- (b) Boxplots of the serum-creatinine levels in 1995 for each group are shown below.



The Minitab output:

#### Estimation for Difference

Difference	95% CI for Difference
-0.1834	(-0.2882, -0.0785)

<sup>18</sup>The 1971 documentation described the study as “a multistage, stratified, probability sample of loose clusters of persons in land-based segments”.

The estimated mean difference in serum-creatinine levels (control – study) is  $-0.18$ , with a 95% confidence interval:  $(-0.29, -0.08)$ .

- 5.6 (a) The point estimate of the mean difference is  $5.3$  in the results obtained in question 4.3 and in the output your colleague provided.
- (b) In your analysis (question 4.3), you found a 95% confidence interval for the mean difference (Method A – Method B) of  $(3.63, 6.99)$ . In your colleague's output the 95% CI for the (mean of Method A – the mean of Method B) is  $(1.67, 8.95)$ .
- (c) The analysis provided by your colleague has treated the groups as independent samples.
- The confidence interval provided by your colleague is wider than the correct confidence interval; the confidence interval for independent samples combines the variability in test scores from both of the two groups. This can reflect variation relating to sex, age, and so on, as well as variation due to the methods of teaching. The paired design 'removes' the unwanted sources of variability.
- (d) The analysis that takes the pairing into account is more appropriate; the pairing attempts to take into account a number of factors that might impact on the variability of the test scores.

- 5.7 (a) The estimate of  $\theta$ , the probability that the eye with drug A scores better than the eye with placebo, is  $42/100 = 0.42$ . The 95% confidence interval for  $\theta$  is:

Method based on normal approximation:

$$\frac{42}{100} \pm 1.96 \sqrt{\frac{0.42 \times 0.58}{100}} = 0.42 \pm 0.0960, \text{ i.e. } (0.32, 0.52).$$

- (b) The estimate of  $\theta$ , the probability that both eyes receive the same score, is  $48/100 = 0.48$ . The 95% confidence interval for  $\theta$  is:

Method based on normal approximation:

$$\frac{48}{100} \pm 1.96 \sqrt{\frac{0.48 \times 0.52}{100}} = 0.48 \pm 0.0979, \text{ i.e. } (0.38, 0.58).$$

- (c) The estimate of  $\theta$ , the probability that the eye receiving drug A scored the same or better than the eye receiving the placebo, is  $(42 + 48)/100 = 0.90$ . The 95% confidence interval for  $\theta$  is:

Method based on normal approximation:

$$\frac{90}{100} \pm 1.96 \sqrt{\frac{0.90 \times 0.10}{100}} = 0.90 \pm 0.0588, \text{ i.e. } (0.84, 0.96).$$

## 6 Hypothesis testing

One of the main activities of applied analytics is obtaining insights, drawing inferences, making conclusions about a general situation, based on analysis of data. Classically, we may think of this as drawing an inference about a ‘population’ from a ‘sample’. This traditional mental framework for the inference process is useful, because it is tangible and accessible, and fits well with much actual inferential work. It can be expanded to a more general, and perhaps more elusive idea: that inference works from the particular to the general, from the specific to the universal.

The general context we seek to make an inference about cannot be scrutinised itself, typically. This means that there must be uncertainty, or imprecision, when we carry out inferences. Whatever method of inference is used reflects that uncertainty, and attempts to quantify it according to sound statistical principles of design and analysis. In Chapter 5, we introduced one approach to quantify this uncertainty—calculating confidence intervals around an estimate of a population parameter. In this chapter, we consider hypothesis testing where uncertainty is quantified using a *P*-value. The calculation of a *P*-value assumes an underlying statistical model. This statistical model represents a hypothesis or theory about the true state of things, and so procedures to find *P*-values are known as hypothesis tests; sometimes they are called significance tests or statistical significance tests.

We first consider an intuitive way of thinking about a *P*-value. This is important because:

“The *P*-value is one of the most misunderstood quantities in psychological research.” (Cohen, 1994)

In this, psychological research is not alone.

### 6.1 The concept of a *P*-value

In a report of a formal statistical analysis, including the vast majority of empirical academic research papers, you will commonly see a probability reported, called a *P*-value. *P*-values are pervasive in statistics; they require careful description and interpretation. The *P* in *P*-value refers to probability, meaning the *P*-value is a probability between zero and 1.

*P*-values arise when data are used to test theories about parameters. A simple example is a randomised controlled trial that compares a new treatment for HIV/Aids with a standard treatment. The outcome of interest is the rate of survival, and researchers want to investigate differences in the mean survival time between the two treatments.

Typically, the calculation of the  $P$ -value for theory testing is based on a pessimist's view of what's true in the world. The pessimist's view is that the true efficacy of the treatments we are comparing do *not* actually differ. In this view, the mean survival time will be the same for the two treatments.

Once the data are collected, a  $P$ -value is calculated. We can think of the  $P$ -value as a potential challenge to the pessimist's view of the world. A small  $P$ -value indicates that the data we have are somewhat surprising if the pessimist's view of the world is correct. The  $P$ -value measures the plausibility of the results we have, if the pessimist's view of the world is right.

Most often, researchers are keen to find small  $P$ -values. At heart, they are optimists, believing, for example, that new treatments do better than old, and they want to challenge the pessimistic view. Indeed, that is why they wish to carry out such a study.

The  $P$ -value does not tell us directly about the plausibility of the pessimist's view of the world – we don't know if this view is right or wrong. It gives us an idea of the plausibility of the data we've sampled, given that view. More formally, the  $P$ -value tells us the probability of the result that we've obtained or results that would challenge the pessimistic view more strongly, if the pessimist's view is right.

## 6.2 The null hypothesis

What kind of hypothesis or theory does this statistical model underlying the calculation of the  $P$ -value represent? In the hypothesis testing approach to statistical inference, the idea is (almost always) to test the *absence* of a true effect. For example:

- A zero difference between population means;
- Two equal population proportions;
- A population correlation of zero.

These are 'null hypotheses': hypotheses about population parameters which specify a particular value for the parameter. In our examples above, the parameter values (difference of means, difference of proportions, correlation) were all zero. If the parameter of research interest is a ratio, the null hypothesis will usually be that it equals one.

A null hypothesis, almost always, entails a theory of 'no effect' in some sense, but this does not always mean that the specified parameter value is zero, as we have just seen with the example of a ratio. 'Null' is the Latin word for 'not any', or 'none'. The null hypothesis is labelled " $H_0$ ".

In the randomised trial of two treatments for HIV/Aids, the null hypothesis is that there is no difference between the population death rates for the two

different treatments. This hypothesis is, of course, the opposite of what the researchers hope to find.

It is usually the case that the null hypothesis is the opposite of what researchers would like to find. The logic of the approach is:

- Assume nothing is going on, that is, there is no true effect;
- Prove me wrong (if you can) with some data that are inconsistent with the null hypothesis;
- If you can't, I'll say the data are consistent with no true effect.

This leads to a more formal definition of the  $P$ -value: it is the probability of the observed result or something more extreme, assuming that the null hypothesis is true.

Sometimes the null hypothesis is referred to as a “chance hypothesis” and the  $P$ -value as indicating the probability that the results are “due to chance”. This language is quite imprecise and a little confusing, so it should be avoided.

#### ▷ EXAMPLE. The Higgs Boson

On the 4th July 2012, the science world was abuzz with news of an important advance in physics: the discovery of the Higgs boson. It made headlines around the world.

What is the Higgs boson? It is an elementary sub-atomic particle. But why was it so important? It's theorized in what is known as the “Standard Model” of particle physics, but before 2012 there was a lack of convincing evidence to support its existence. It was a missing part of the Standard Model as it is the particle that explains why matter has mass.

The Higgs boson was detected with experiments run in the Large Hadron Collider — a particle accelerator built in a tunnel under the French-Swiss border. The circumference of the tunnel is 27 kilometres. Protons can be fired at each other inside this vacuum at very high speeds. The observations made in the experiment detecting the Higgs boson were the number of collisions of protons occurring at various voltages, measured in Giga-electron volts (GeV). The science behind the experiments is highly technical, but the evidence for the discovery is based on a hypothesis test. The null hypothesis (pessimistic view) is that there is no Higgs boson, represented by the red line on Figure 59.

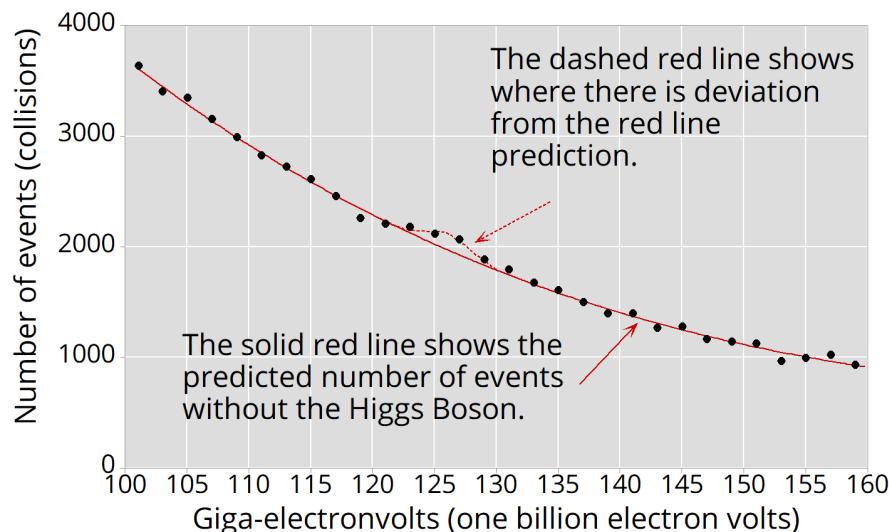


Figure 59: Scatterplot showing the number of collisions versus voltage in the experiments providing evidence of the Higgs boson

However, physicists discovered a new particle with a mass around 125 GeV, where the observed result was not consistent with the red line. This was reported as a ‘5 sigma’ result, a shorthand way of referring to a  $P$ -value equal to 0.0000006. The probability is obtained as the probability that a normal distribution is 5 standard deviations from its mean, or further; hence ‘5 sigma’. The interpretation was that the result observed (or one even more extreme) was highly unlikely, and that this provided evidence of the existence of the Higgs boson.

### 6.3 Formulating statistical hypotheses

The research agenda in which hypothesis tests are commonly used is driven by substantive questions about the way things really are: the truth about the world, if you like. A research question can be, in some senses, general, such as: “Is this new drug a good treatment for people suffering from HIV-AIDS?” When we ask the question in this way, we do not yet have concrete hypotheses to consider. We need to clarify what “good” means: good for what purpose? And good, in relation to what alternatives?

Further thought might lead to considering survival after diagnosis as an outcome, and comparisons with established treatment. Here are some hypotheses about this outcome.

Compared with patients on the established treatment, the average length of survival for patients on the new drug is:

- (a) the same

- (b) different
- (c) longer
- (d) 5 years longer

All of these hypotheses are theories about the population, or the real situation. They are not statements about the data. But they are what we want to consider; in research, we are nearly always interested in the wider context, or the population.

We can formulate such hypotheses in terms of a statistical parameter. If we define:

$$\begin{aligned}\mu_N &= \text{mean survival on new drug} \\ \mu_E &= \text{mean survival on established treatment} \\ \mu_D &= \mu_N - \mu_E\end{aligned}$$

... then we can say that the corresponding statistical hypotheses are:

- (a)  $\mu_D = 0$
- (b)  $\mu_D \neq 0$
- (c)  $\mu_D > 0$
- (d)  $\mu_D = 5$

Note that the first hypothesis is a null hypothesis, whereas the other three describe some effect of the new treatment on the outcome and so reflect the research agenda. These are forms of the research hypothesis.

## 6.4 An example: water quality

We consider a simple example. Like many examples we meet initially, it is artificial in some respects, but it does have important features common with a realistic context.

Drinking water has several physical and microbiological properties of interest, and any government authority that looks after the quality of a community's drinking water usually monitors this with a testing program. One of the physical properties of interest is the pH of the water.<sup>19</sup>

The pH scale measures the acidity or alkalinity of a fluid; it is related to hydrogen ion concentration. A pH of 7 means the solution is neutral; values less than 7 are more acidic, and values greater than 7 are more alkaline.

---

<sup>19</sup>Actually, pH is of lesser importance, typically, than other features, such as microbiological contamination.

While there is no well-established health guideline for the pH of drinking water, if the pH is too low (acidic), corrosion in pipes can occur, while if it is too high (alkaline), skin irritation is a risk. For water supplied to a community for personal use (drinking and washing), pH values in the range 6.5 to 8.5 are generally thought desirable.

Suppose that we nominate a preferred mean value for the pH:  $\mu = 7.5$ . Suppose, further, that it is known that the population standard deviation of pH values is 0.5. That is, we want  $\mu = 7.5$  and we assume  $\sigma = 0.5$ .

The monitoring authority takes a random sample of 25 observations at their sampling taps, placed throughout the city.<sup>20</sup> Define  $X$  to be the pH of a sample from one of these taps, and assume that  $X$  is normally distributed.

Assuming independence of the 25 samples, which may be approximately reasonable if the taps are randomly placed, the distribution of the sample mean,  $\bar{X}$ , is itself normally distributed, in fact if  $\mu = 7.5$ , as we hope:

$$\bar{X} \stackrel{d}{=} N\left(7.5, \frac{0.5^2}{25}\right).$$

This means that the standard deviation of  $\bar{X} = 0.1$ , so that the distribution of  $\bar{X}$  is as shown in Figure 98.

Sampling distribution of the mean for samples of size 25

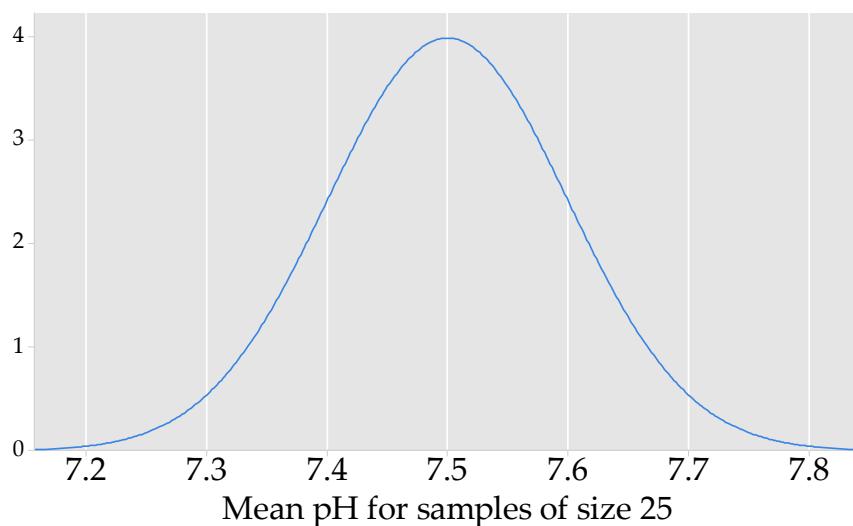


Figure 60: Sampling distribution for the mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5.

We are making a number of assumptions here, but the key *hypothesis* is that  $\mu = 7.5$ . This may or may not be true. We hope it is true, but there could be

---

<sup>20</sup>This process does occur in Melbourne; there are Melbourne Water sampling taps in random locations just outside the boundaries of residential properties.

a problem: the water might be much more alkaline than that, in which case  $\mu > 7.5$ , or it might be much more acidic, in which case  $\mu < 7.5$ .

Now an actual sample of 25 is obtained, and it is found that the observed sample mean is 7.75. That is,  $\bar{x} = 7.75$ . Obviously this is larger than 7.5. But we don't expect to get a sample mean exactly at 7.5, there will be some random variation. What we want to do is to use the *known* sampling distribution of  $\bar{X}$  to address this question: to say whether or not we should regard  $\bar{x} = 7.75$  as indicative of a problem or not.

We've observed  $\bar{x} = 7.75$ . Is this acceptable?

The way we think about this is to ask: how strange is this observed value  $\bar{x} = 7.75$ , if the true mean is really 7.5? How unusual is it, if the desired situation is the reality? We can answer this by working out the probability of obtaining an observed sample mean this far away from 7.5, or further. It could have been just as distant from the desired  $\mu = 7.5$  on the low side, and that would also start to raise concerns. The picture is as shown in Figure 62. As we can see, this probability is quite small. It is not zero, but the fact that it is small is suggestive of evidence *against* the hypothesis that "everything is OK" and  $\mu = 7.5$ .

Sampling distribution of the mean for samples of size 25

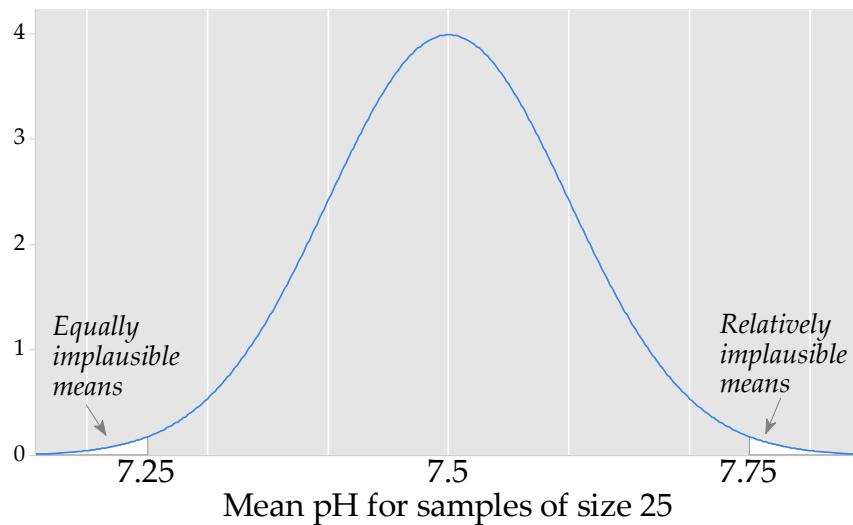


Figure 61: Sampling distribution for the mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5. The white areas reflect the probability of a sample mean at least as extreme as the observed  $\bar{x} = 7.75$ .

Now we consider the calculation of the  $P$ -value, based on the reasoning above. It can be calculated, in this case, using the normal distribution. We find that the probability of a sample mean of at least 7.75, when the sample size is 25, the standard deviation of an individual observation is 0.5, and the

true mean is  $\mu = 7.5$ , is 0.006210. This makes the  $P$ -value twice this value, or  $P = 0.012$ .

Sampling distribution of the mean for samples of size 25

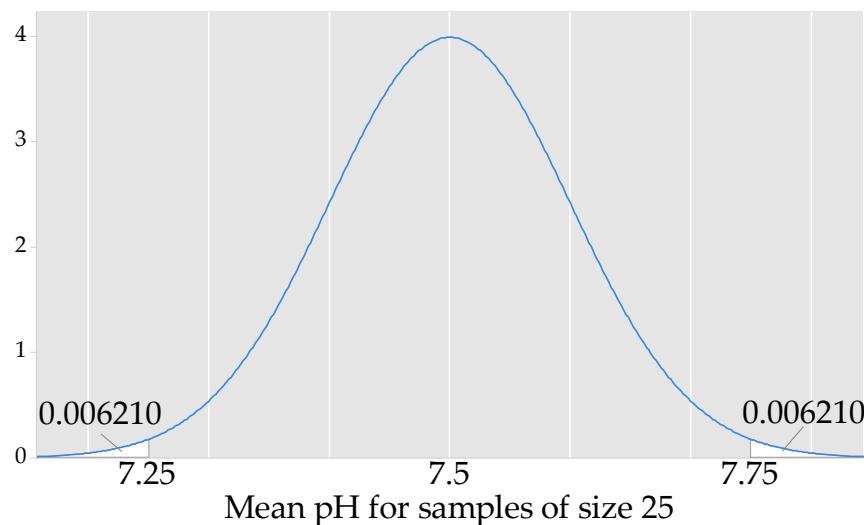


Figure 62: Sampling distribution for the mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5. The white areas reflect the probability of a sample mean at least as extreme as the observed  $\bar{x} = 7.75$ , and hence give the  $P$ -value:  $P = 0.012$ .

## 6.5 Structure of the hypothesis test

Let's now consider the structure of the hypothesis as illustrated by the example of assessing water quality.

The research agenda is monitoring water quality with the aim of detecting water that is considered alkaline or acidic. In this context, the null hypothesis ( $H_0$ ) corresponded to average pH levels corresponding to the standard:  $H_0: \mu = 7.5$ .

Having formulated a research question and a related null hypothesis, we need relevant data to make an inference: to draw a conclusion, with uncertainty, about the null hypothesis.

There is more than one way to think about this process; in this course, we use the conventional and widely applied “frequentist” approach.<sup>21</sup>

The idea of data that are “relevant” to testing a null hypothesis is enormously important, of course. It is at this point that we need to think about research design: the way that we design our study will affect the extent to which we can draw sound and reliable inferences. The formal study of the

---

<sup>21</sup>You might find the “Bayesian” approach mentioned in some contexts: it is a different approach to making inferences, not covered in this course.

design of experiments is a topic in itself, and we touch on it to some extent in this course.

Suppose that we have collected data relevant to our question. We may ask: "Are these data consistent with the null hypothesis, or not?" That is, we examine the study result and consider whether the data in the study could have come from a population model in which the null hypothesis is true.

To make this idea applicable we need what is called a **test statistic**.

- The test statistic is a function of the observations (i.e. a single value summary such as  $\bar{X}$  or, in a context of comparison  $\bar{X}_1 - \bar{X}_2$ ), which is used to make an inference about  $H_0$ .
- The test statistic needs to have a known distribution, at least approximately, when the null hypothesis is true, and this distribution needs to change when  $H_0$  is not true.

In the water quality example above, relevant data came from the collection of water samples, and the test statistic was the mean pH level. In order to set up a model for this test statistic, under the null hypothesis, we needed to make some assumptions:

- The distribution of pH of drinking water is approximately normal;
- The population standard deviation of pH values is 0.5.

Together, the null hypothesis and the assumptions allow us to specify what the distribution of the relevant test statistic — the sample mean — would be, if the null hypothesis is true.

We collected data and calculated the mean for our sample of pH measurements. The question then is: what does the sample mean tell us about the population mean? In particular, if the true population mean is 7.5, (°) how likely is it that we observe a sample mean as big as we did observe?

Statistical theory allows us to actually calculate this probability. We define

$$P\text{-value} = \Pr \left( \begin{array}{l} \text{result at least as extreme as that obtained,} \\ \text{given that the null hypothesis is true} \end{array} \right)$$

The  $P$ -value provides a numerical measure of uncertainty and expresses quantitatively the level of surprise in the observed data, assuming that the null hypothesis is true. The direction of the measure is that the *smaller* the  $P$ -value, the *greater* the level of surprise. If the  $P$ -value is *small*, this is evidence against the null hypothesis. If the  $P$ -value is large, nothing very unusual has happened, on the assumption of the null hypothesis.

### Choosing the test statistic

For any given situation there is likely to be a number of reasonable test statistics. But under specific assumptions (e.g. Normality of a numerical outcome of interest) there is often a ‘best’ test statistic; i.e. one that is most sensitive to  $H_0$  not being true. The best test statistic is determined by the design, and properties of the data. There are standard test statistics for standard situations, and these are well-known (e.g. *t*-test, analysis of variance and so on). But once the context is even slightly more complicated, the test statistic is not very obvious at all, and statistical science has been used to derive the form of the test statistics.

As we will see in upcoming chapters, understanding and considering the validity of the assumptions is an important aspect of carrying out a hypothesis test correctly. There will be times when it is better to use a test statistic whose validity does not require stringent assumptions, such as Normality, than to use the ‘best’ test statistic.

## 6.6 Misconceptions about hypothesis testing and *P*-values

In some disciplines, estimation and calculation of confidence intervals has been the primary approach to drawing statistical inferences. In other disciplines, hypothesis testing has a long tradition as being the standard approach. At the same time, there has been some trenchant criticism of the use of hypothesis testing. There are two broad classes of criticism. The first relates to the use of hypothesis testing in practice and concerns that arise from the misuse and misinterpretation of hypothesis tests and *P*-values. The second concerns relates to the philosophy and logic of the method itself. We consider the first class of criticism here.

### There's more to life than statistical significance

Statistical significance is a term sometimes used to signify finding a small *P*-value in the results of an analysis. Common convention is that *P*-values less than 0.05 are considered to be small, and so findings with  $P < 0.05$  are deemed to be statistically significant. For some people, finding statistically significant results is the holy grail of statistical analysis.

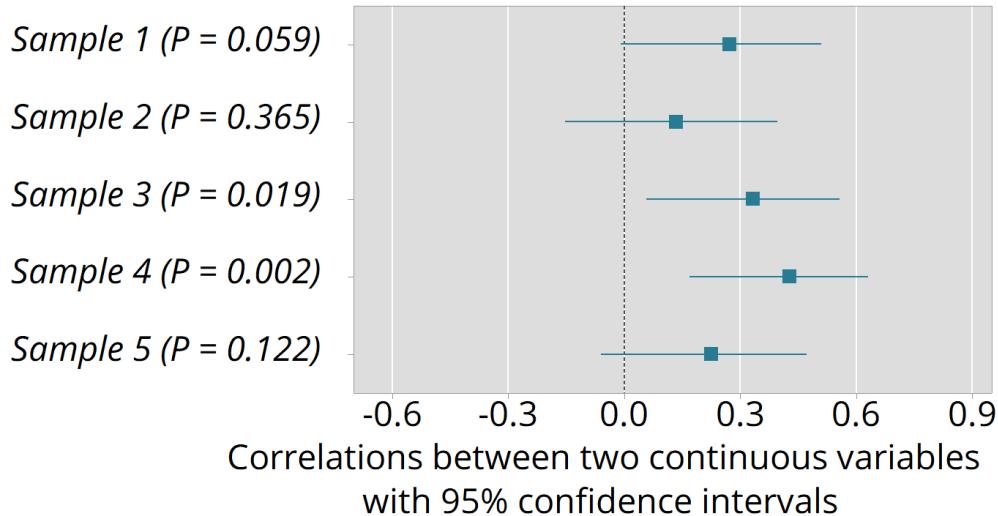
If the *P*-value is small, then we may wish to specify a threshold of “statistical significance”. Note that the structure of hypothesis testing does not really require us to, and we may be content to simply report the *P*-value, however large or small. But sometimes an action is required if we conclude that the null hypothesis is false, and in such cases the preference for a cut-off or threshold has led to some commonly used values, the most famous of which is  $P = 0.05$ , or the 5% level of significance, due to R.A. Fisher. Any such thresholds, even if they have the weight of convention, are ultimately arbitrary, just like the value of the confidence level for a confidence interval.

### Blinkered interpretation and binary thinking

Too much focus on statistical significance can result in blunt and over-simplified interpretation of results. While a threshold of 0.05 is commonly used, it is ultimately arbitrary, making it quite bizarre to form drastically different conclusions according to whether  $P = 0.051$  or  $P = 0.049$ . Indeed, two psychologists famously said:

“...surely, God loves the .06 nearly as much as the .05.” (Rosnow and Rosenthal, 1989, p.1277)

Consider, for example, the results from five different samples, each with a sample size of 50 — you can think of them as five different studies of the same population. Figure 63 shows the correlation between two continuous variables measured for each sample. A 95% confidence interval for the true correlation is shown. The results are simulated here, so we know the true correlation; it is 0.3.



*Each sample has 50 observations,  
and is from the same underlying population.*

Figure 63: Confidence intervals around an observed correlation for five different samples of the same size

The labels on the plot show the  $P$ -values, in each case, for testing the null hypothesis that the true correlation  $\rho = 0$ . The line at zero on the plot corresponds to the null hypothesis that the true correlation is zero. If we interpret the results through the blinkers of statistical significance, the results will seem to be inconsistent. Using the usual definition of statistical significance ( $P < 0.05$ ), two of the results are statistically significant, and three are not.

Note also the 95% confidence intervals; in the two cases where  $P < 0.05$ , the 95% confidence intervals do not include the null hypothesis value of zero. This is not a coincidence; we learn about this connection in section 6.10.

But remember — these results were all sampled from the same population. As expected, in the plot the sample correlations (the squares) vary around the true value of 0.3.

There are associated problems of interpretation when we think of, or limit the report of a statistical test, to the binary “significant/not significant” outcome. When the result of the hypothesis test is found to be not statistically significant, this is sometimes interpreted to mean that the null hypothesis is true. Logically, this is *wrong*.

The null hypothesis entails a very precise specification of the parameter of interest, and it is essentially certain not to be true. In the water quality example, the null hypothesis is  $H_0 : \mu = 7.5$ , that is, the true mean pH is exactly 7.5.

If we obtained a  $P$ -value from relevant data and found it to be, say,  $P = 0.60$ , then this does not imply that  $\mu = 7.5$ . Rather, it says that *if* the null hypothesis is true (i.e.  $\mu = 7.5$ ), *then* the data we obtained would not be unusual: the result, or data more extreme, would occur 60% of the time. However, to reverse this and say that this implies that the null hypothesis is true, is not a correct argument. For one thing, consider testing an alternative null hypothesis, that, say  $\mu = 7.501$  (a number very close to 7.5). The  $P$ -value for testing this different null hypothesis would surely be very similar to 0.6; maybe 0.61. So do we then conclude that  $\mu = 7.501$ ? The parameter  $\mu$  is a fixed but unknown number. It cannot simultaneously be 7.5 and 7.501.

However, it is still quite common for researchers to report and discuss their findings as if it were true that a  $P > 0.05$  means that the null hypothesis is true. It can be seen even in the titles of articles.

Of course, it is possible for the data to be *consistent with* both  $\mu = 7.5$  and  $\mu = 7.501$ , and that is a perspective that is encouraged by a confidence interval, with a range of values for the unknown parameter.

## 6.7 Principled use of hypothesis tests and statistical inference

### Stop the search for the holy grail

The focus on achieving a small  $P$ -values can lead to researchers to search for statistically significant results. But occasionally unlikely results can occur, when the null hypothesis is true. Repeatedly testing hypotheses can increase the chance of this. How much does this matter? If a research ‘finding’ has arisen in the context of many tests, but the full range of tests conducted

is suppressed, the statistical significance of the test is misleading. This can occur in more or less subtle ways, in the reporting of statistical material. This is sometimes called “**data-dredging**” or “**p-hacking**”. Clearly, this is a dubious practice, and in extreme cases might amount to scientific fraud.

We might read something like the following:

“There was no significant difference overall between the active treatment and placebo. However, among women aged 30 to 34 who had more than one child, the difference was statistically significant ( $P = 0.04$ ).”

Any account of a study that is along these lines needs to be treated very sceptically.

This issue is related to the so-called “multiple comparisons problem”, which will be discussed in more detail later.

In many situations, there is more than one test that can be used: for example, a test based on the Normal distribution, and a distribution-free test. Statistical tests have different properties, and there may be features of the data or study that lead us to prefer using one test and not another.

However, it is wrong to carry out a few tests, and then just report the result of the test with the smallest  $P$ -value. Essentially, the choice of statistical test should be “blind” to the  $P$ -values produced.

### Theory driven variable selection

Hypothesis testing can be a useful tool in finding statistical models, especially when we are attempting to fit models with a large number of potential explanatory variables. This is because it gives us an apparently objective framework for making decisions about the model and the influence of variables.

However, it may be important — here as elsewhere — to consider the estimated effect of potential explanatory variables, regardless of statistical significance. And in the fitting of some models, we may want to include some variables known to be relevant, even if they do not prove to be statistically significant for the data set we have. Such variables might be age and sex, for example, when fitting models to epidemiological data.

## 6.8 Principled reporting of hypothesis tests and statistical inference

Any study that only tells you about the statistical significance of the findings is selling itself short. There is important information in the estimate of the effect observed and the precision of the estimate of the effect (the confidence

interval). We also need to know the magnitude of the  $P$ -value, rather than whether it is above or below a cut-off.

Thinking about the estimate and its precision allows you to think about the substantive meaning of the results. Figure 64 shows six hypothetical studies, with differing sample sizes, but with the same observed correlation (the squares at 0.3). The top result, for a sample size of ten, has a wide confidence interval. Again, the vertical line on the plot corresponds to a null hypothesis that the true correlation is zero. For a sample size of ten, the result is not statistically significant. However, as the sample size increases, the confidence intervals narrow. Once the sample size is 50, the null hypothesis falls outside the interval: the result is statistically significant. But here the effect estimated — the correlation — has exactly the same magnitude. The studies haven't estimated a different effect; rather they've estimated the same effect with different precision. The different sample sizes here mean that there is different potential for finding small  $P$ -values.

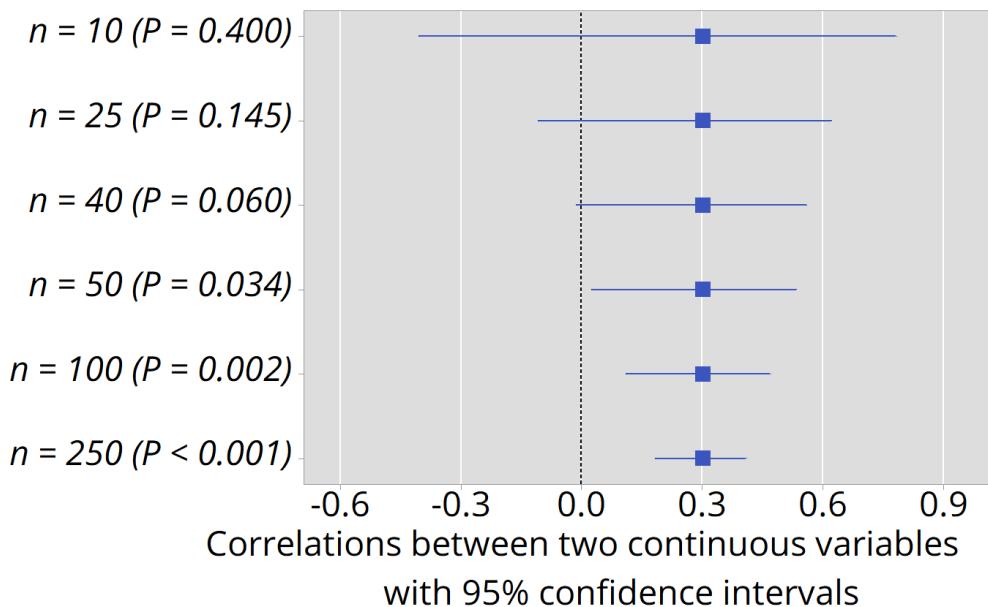


Figure 64: Confidence intervals around an observed correlation of 0.3, for varying sample sizes

For much of the history of significance testing, the results of tests were reported as either “statistically significant”, or “not statistically significant”. Sometimes, the threshold of statistical significance in such reporting is not even stated, but it is nearly always 5%, or  $P < 0.05$ . It is still possible to find such reporting in the literature.

Consider a randomized controlled trial, comparing a new treatment for

HIV-AIDS with a standard treatment, reported along the following lines:

“In our randomized trial, the difference between the mean survival on the new treatment and on the established treatment was statistically significant at the 5% level.”

The problem here is not that the result has been wrongly calculated, but that the report of it is so impoverished. A lot of expense and effort may have gone into designing and carrying out the study, and all that one is left with is a conclusion that is qualitative, and has only two possible outcomes.

Notice that the above report does not even tell us the direction of the result: was the new treatment better, or worse? Even reporting the point estimate of the effect gives more information than this, and better still is to report the 95% confidence interval.

Hence a principled report of the inference carried out would be:

“In our randomized trial, the mean survival on the new treatment was greater than on the established treatment by 3.5 years, with a 95% confidence interval of (1.2 to 5.8) years; the  $P$ -value was 0.003.”

Here the null hypothesis is not explicitly stated, and that is fairly common; we are required to see that the null hypothesis tested was that the difference between the true mean survival on the two treatments is zero.

### Report precise $P$ -values

The  $P$ -value is a continuous quantity, and so should be reported as such, to a sensible number of decimal places. Let’s assume the results of some analysis find a  $P$ -value of 0.005, to three decimal places; this is a reasonable way of reporting the  $P$ -value. However, there are a number of other ways you might see such a result reported, none of which is recommended.

### Avoid the following:

- $P = 0.0050706$ : this is too many decimal places.
- 0.005 in a table labelled “Sig.” at the top: the  $P$ -value is reported not “Sig.”, always use the label  $P$ -value.
- $P < 0.05$ : this only indicates the  $P$ -value in relation to an arbitrary threshold.
- Using the “star” system, and indicating that the result is, in this case, less by 0.01 by using \*\*: again this does not provide precise information.

- Stating that the result is statistically significant: this only has implied numerical meaning.

If you are reporting  $P$ -values in an academic paper or thesis, it's good practice to report the actual value to three decimal places. If the  $P$ -value is very small, common practice is to report it as  $P < 0.001$ , and there is good reason for that: the distributions on which  $P$ -values rely are often approximate, and the approximations are likely to be least adequate in the tails of the distribution, where small  $P$ -values are obtained. But, of course, it's not sufficient to only report a  $P$ -value; relevant estimates and confidence intervals should also be provided.

### Failure to launch? Thinking about large $P$ -values

How does a researcher think about large  $P$ -values? Often, it seems, an initial reaction is disappointment, but again this arises from a very narrow view of the meaning of a  $P$ -value. A large  $P$ -value indicates the result observed is consistent with the null hypothesis. However, this does not mean that we should draw an inference that the null hypothesis is true or correct. Large  $P$ -values might arise if a study is poorly conducted, if there is substantial variability or the sample size is small. We need to know more about the estimate and the confidence interval before any substantive interpretation is made. Here's a cartoon about tempting but misleading ways to think about  $P$ -values:

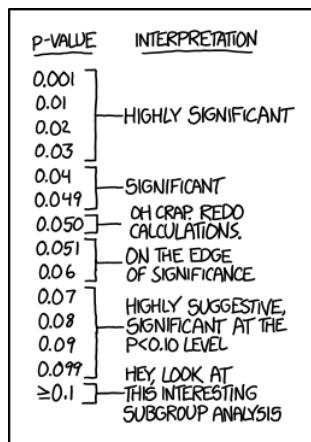


Figure 65: *How not to think about p-values*, xckd cartoon number 1478

### Practical or clinical 'significance'

Always keep in mind that statistical significance is not equivalent to the practical importance of a result. The reporting of  $P$ -values has taken precedence over confidence intervals, and even summary statistics, in some disciplines. If the estimates or the confidence interval are left out, the practical meaning cannot be considered.

### Avoid the slippery slide

The term statistical significance is often used as short-hand. However the term signifies the relative magnitude of the  $P$ -value, nothing more. A problem arises when statistical significance is taken to mean significance of the findings in a more general sense. Findings might be described as:

- statistically significant in summarising the results ...
- significant in discussing the findings, ...
- and, important in wrapping up the conclusions.

Take care to avoid this slippery slide, where statistical significance morphs into importance. The findings may or may not be substantively significant or important. This will depend on the magnitude of the effects, rather than on the statistical significance *per se*.

### Avoid the obsession with $P = 0.05$

The obsession that some researchers have with achieving statistical significance leads to all sorts of linguistic gymnastics. Here are some real examples where the author tries to claim statistical significance in some sense when the  $P$ -value is greater than 0.05.

- a barely detectable statistically significant difference ( $p = 0.073$ )
- a little significant ( $p < 0.1$ )
- a margin at the edge of significance ( $p = 0.0608$ )
- a robust trend toward significance ( $p = 0.0503$ )
- almost clinically significant ( $p < 0.10$ )
- an important trend ( $p = 0.066$ )
- approaches but fails to achieve a customary level of statistical significance ( $p = 0.154$ )
- barely escaped statistical significance ( $p = 0.07$ )
- did not reach the usually accepted level of clinical significance ( $p = 0.07$ )
- flirting with conventional levels of significance ( $p > 0.1$ )
- narrowly eluded statistical significance ( $p = 0.0789$ )
- perceivable statistical significance ( $p = 0.0501$ )
- tantalisingly close to significance ( $p = 0.104$ )

- teetering on the brink of significance ( $p = 0.06$ )

These are some of the more imaginative wordings from a large collection scraped from the web by Matthew Hankins.<sup>22</sup>

### Avoid the file drawer problem

When statistical significance is the focus, poor decisions may be made about the value of disseminating results. For example, a researcher who uses statistical significance as a measure of importance might believe that any of the results of a particular study that fail to achieve statistical significance can be “left out” of a publication. The “file drawer problem” refers to whole studies with results that are not statistically significant being left in the bottom of a drawer and never published.<sup>23</sup> A consequence of this is that published research literature would be biased.

The importance of any particular analysis should be determined *before* the results are in: any outcome considered important enough to measure should be considered important enough to report on.

## 6.9 One- or two-sided statistical tests

The null hypothesis specifies the population model when there is no effect; we test the null hypothesis. In the definition and calculation of the  $P$ -value, as described above, we can consider results at least as extreme as that observed, in either direction, that is, in both tails of the distribution of the test statistic. In doing so, we are carrying out a two-sided or two-tailed statistical test.

In relatively simple inferential contexts, such as our water quality example or in a comparison of two means, it is possible to conduct one-sided or two-sided hypothesis tests. These contexts involve a single parameter. So-called directional hypothesis tests do not arise in more complex inferential contexts.

### Alternative hypothesis

Behind the choice between a one- or two-sided hypothesis test is the population model we consider when the null hypothesis is *not* true, known as the **alternative hypothesis**, denoted by  $H_1$ . In fact, it often corresponds to what we hope to show or demonstrate. In a general sense, the alternative hypothesis says that “there is an effect” (of some sort), while the null hypothesis says that there is no effect.

Unlike the null hypothesis, the alternative is usually not precisely specified, although it can be. In the HIV-AIDS example, the null hypothesis was that

---

<sup>22</sup>See <https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/>

<sup>23</sup>Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3), 638.

$H_0 : \mu_N - \mu_E = 0$ . Note that this defines the value of the parameter of interest — the difference in mean survival between the two treatments — to be an exact figure, namely, zero.

The alternative hypothesis in this case could be taken to be that there *is* a difference in mean survival between the two treatments, and hence that  $\mu_N - \mu_E \neq 0$ .

### One or two-sided alternative hypotheses

The idea of one- or two-sided alternative hypotheses can be hard to grasp, partly because it touches on quite deep concepts about the structure of hypothesis testing.

When we are considering a research question about a new or alternative treatment or method, it is likely that we think that the new treatment, if it has an effect at all, will do so in a particular direction.

If we are investigating a method in education, we hope that it may lead to improved results. Usually we are not interested in methods that could make things worse, and in some areas of research, notably medicine, it would be unethical to do so.

Conversely, if we are investigating a possible hazard, for example, a disease in plants, we usually presume that any effect will be an adverse one: it will make things worse. We do not consider that a disease could improve the health of plants.

These natural ideas can be described as “directional” research ideas or hypotheses: they imply that any effect will be in a particular direction. They do not envisage an effect in the opposite direction.

However, while we may have a strong belief that any effect will be in a particular direction, that does not mean that we can be certain — before the data are collected — that any observed effect will be in the hypothesized direction. We might be surprised by the result: the result might turn out to be contrary to our research idea.

These notions about the direction of effects translate into the nature of alternative hypotheses and tests.

In simple situations we can consider either one-sided or two-sided  $H_1$ s. For example, for the null hypothesis  $H_0 : \mu_1 = \mu_2$ ,  $H_1 : \mu_1 > \mu_2$  is a one-sided alternative, whereas  $H_1 : \mu_1 \neq \mu_2$  is a two-sided alternative.

For our water quality example, the null hypothesis  $H_0 : \mu = 7.5$ . The hypothesis  $H_1 : \mu \neq 7.5$  is a two-sided alternative. If we were only concerned about alkalinity,  $H_1 : \mu > 7.5$  is the relevant one-sided alternative.

Should we choose a one or a two-sided alternative hypothesis and test?

There is a strong and conservative convention to prefer two-sided tests. This is because a one-sided alternative hypothesis means that the only possible

theories you are prepared to consider are:

- no effect, or the null hypothesis (e.g.  $\mu = 7.5$ );
- an effect in a particular direction (e.g.  $\mu > 7.5$ ).

In particular, with a one-sided alternative hypothesis you are asserting that you are *sure* that there cannot be any effect in the opposite direction. Even if you think or hope that the effect will be in one direction, it is another thing altogether to have such conviction about this that you refuse to contemplate the possibility of an effect in the opposite direction.

If you are carrying out a one-sided test, no amount of data pointing to an effect in the opposite direction to the one you hypothesized will convince you of such an effect. Hence the choice of a one-sided test can be seen as a form of scientific arrogance, and it is usually recommended that a two-sided test be carried out.

This choice impacts on the calculation of the  $P$ -value. It does so because of the word “extreme” in the definition of the  $P$ -value:

$$P\text{-value} = \Pr \left( \begin{array}{l} \text{result at least as extreme as that obtained,} \\ \text{given that the null hypothesis is true} \end{array} \right)$$

### Calculation of a one-sided $P$ -value

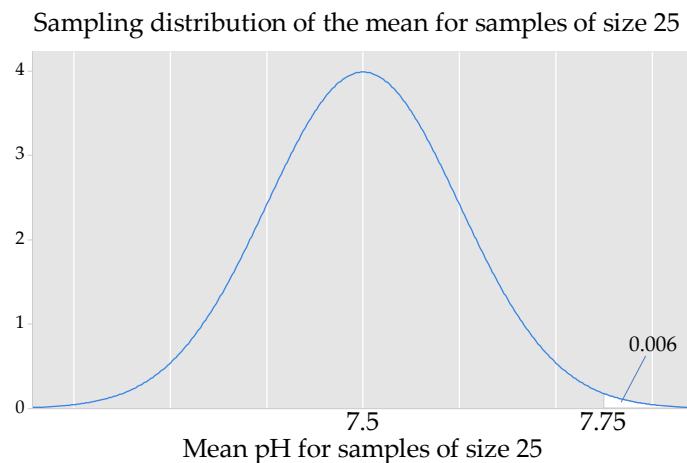
To calculate the  $P$ -value for a one-sided alternative we look at the one-sided alternative to determine how to interpret “extreme”, in the definition of the  $P$ -value. In the water quality example above, if we observe  $\bar{x} = 7.75$ , say, more extreme values would be  $\bar{x} > 7.75$ .

With the null hypothesis and one-sided alternative as above, then, we calculate the one-sided  $P$ -value as

$$P = \Pr(\bar{X} \geq 7.75, \text{ given } H_0 \text{ is true}).$$

▷ **QUESTION:** This will be a number smaller than 0.5. Why?

Figure 66 shows the one-sided  $P$ -value, based on the reasoning above. We find that the probability of a sample mean of at least 7.75, when the sample size is 25, the standard deviation of an individual observation is 0.5, and the true mean is  $\mu = 7.5$ , is 0.006.



**Figure 66:** Sampling distribution for the sample mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5. The white area reflects the probability of a sample mean at least as extreme, in one direction, as the observed  $\bar{x} = 7.75$ , and hence give the P-value:  $P = 0.006$ .

What happens if we observe a mean pH for our sample that is in the opposite direction to that which we expect? For example, if we observe  $\bar{x} = 7.25$  as shown in Figure 67 — a value just as far from the null hypothesis as 7.75 — it will not incline us to the conclusion that  $\mu$  is less than 7.5, since we did not entertain that as a possibility, and we may not change our hypotheses according to the data. So if we observe  $\bar{x} = 7.25$ , say, the P-value will be:

$$P = \Pr(\bar{X} \geq 7.25, \text{ given } H_0 \text{ is true}).$$

▷ **QUESTION:** The one-sided P-value will be a number greater than 0.5. Why?

You can see the exact P-value in Figure 67.

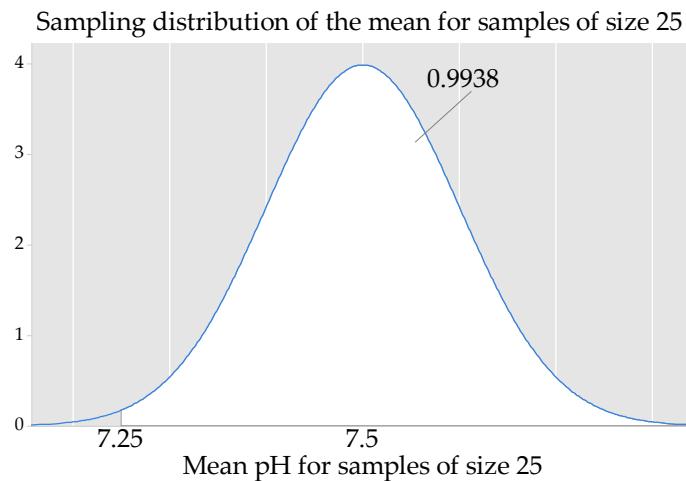


Figure 67: Sampling distribution for the mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5. The white area reflects the probability of a sample mean at least as extreme, in one direction, as the observed  $\bar{x} = 7.25$ , and hence give the  $P$ -value:  $P = 0.9938$ .

### Calculation of a two-sided $P$ -value

In our water quality example, we have seen that the two-sided  $P$ -value was calculated as double the one-sided  $P$ -value; in many of the cases we cover in this course the distribution of the test statistic is symmetric when the null hypothesis is true and hence doubling the one-sided  $P$ -value is appropriate.

For a two-sided alternative, when the distribution of the test statistic is not symmetric, the  $P$ -value can be worked out in more than one way, and we do not cover this here.

## 6.10 Confidence intervals and hypothesis tests

In an inference setting, the calculation of a confidence intervals and the  $P$ -value are commonly based on the same underlying statistical theory. When that is the case, there is a link between them. It is useful to know how this relationship works.

A confidence interval contains a range of values for the parameter of interest that are plausible, given the data. This should resonate with testing; when we carry out a hypothesis test, we are asking whether the data are consistent with a particular parameter value: is it plausible?

For inference on a single parameter using the same method for both the interval estimation and the hypothesis test, we can state the connection as follows: a 95% confidence interval consists of all parameter values which, if tested, would give  $P > 0.05$ .

More simply and informally, the 95% confidence interval consists of all pa-

parameter values with which the data are consistent. We operationalise “consistent” as meaning  $P > 0.05$ .

▷ **EXAMPLE. Fuel efficiency of cars with 1.6 litre engines**

Consider the data on tuning the engines of used cars, which we considered in chapter 5. We considered the difference in fuel economy, comparing pre-tuning with post-tuning; a positive difference reflects a benefit of tuning. The estimate of the mean difference was 0.28, with a 95% confidence interval for  $\mu_D$ , the true mean difference, of (0.03, 0.53) litres/km. This was shown in Figure 54, part of which is repeated here in Figure 68.

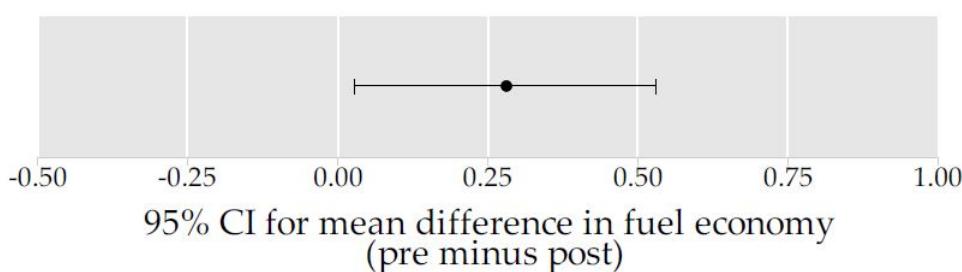


Figure 68: Estimate and 95% confidence interval for the mean difference of fuel economy, pre- minus post-tuning, cars with 1.6 litre engines.

Here a natural null hypothesis is  $H_0: \mu_D = 0$ . We will learn in Chapter 9 how to carry out a test for this; we find that the  $P$ -value is 0.03.

Note that the 95% CI here does not include zero (although the lower limit is not much greater than zero). In this sense, the confidence interval suggests that the data are *not* consistent with  $\mu_D = 0$ . (Of course, we realise this is not a definite conclusion.) Correspondingly, the  $P$ -value for testing  $H_0: \mu_D = 0$  is found to be less than 0.05, in fact  $P = 0.03$ .

- If the null hypothesis we are considering is outside the 95% confidence interval, the  $P$ -value for that null hypothesis will be less than 0.05.
- Conversely, if the null hypothesis we are considering is inside the 95% confidence interval, the  $P$ -value for the null hypothesis will be greater than 0.05.
- If we test the two ends of the 95% confidence interval as null hypothesis values for the unknown parameter, we obtain  $P = 0.05$ .

This can be stated in another way: a 95% confidence interval consists of all parameter values which would have a  $P$ -value greater than 0.05 if we tested those parameter values as the null hypothesis. There is a link between the “95%” and the value “0.05”, here:  $0.05 = 1 - 0.95$ . For a different

confidence coefficient, such as 90%, the link would need to be expressed correspondingly: a 90% confidence interval consists of all parameter values which would have a  $P$ -value greater than 0.10 if we tested those parameter values as the null hypothesis.

Consider, as another example, the five confidence intervals and hypothesis tests carried out for a correlation in Figure 63. Look carefully at the 95% confidence intervals, and the location of zero in relation to them. In the three cases where  $P > 0.05$ , the corresponding 95% confidence intervals include zero, the null hypothesis value. In the two cases where zero is outside the 95% confidence interval, the corresponding  $P$ -values are less than 0.05. Look at the first case: the 95% confidence interval just includes zero, and  $P = 0.059$ , a value just greater than 0.05.

This correspondence is often used in discussing research and is important to understand.

## 6.11 Exercises

- 6.1 A multiple choice test (on Applied Aesthetics) consists of twenty questions with five alternative answers only one of which is correct. Ian is taking the test, but has absolutely no idea about the subject, and so he guesses the answer to each question. This means he has a probability of 1/5 of getting each question correct.
- What is the distribution of Ian's score on the test? (.)
  - We suspect that Sue may have some idea. What is an appropriate null hypothesis to test this? What is an appropriate alternative hypothesis? What is the parameter of interest here?
  - Assuming that Sue is guessing (.) find the probability that Sue gets 10 or more.
  - Sue gets 10 correct on the test. What do you conclude?
- 6.2 In consumer affairs law in Australia, there is a rule that says: "Rule 1: The declared quantity on a package should accurately reflect the quantity being supplied." A company producing honey in 200 gram jars has some concerns about its new filling machine. The label on each jar states that the amount of honey in the jar is 200 grams. They take a random sample of 30 honey jars, filled by the new machine.
- The company is concerned about the amount of honey in the population of jars they produce. Which one of the following is an appropriate parameter of interest in this case?
    - the labelled value: 200 grams;
    - the average weight of the sample jars;
    - the mean amount of honey per jar;
    - the amount of honey in one jar;
    - the deviation of the weight from 200 g;
    - the average amount in the thirty honey jars sampled.
  - For the question the company has, define an appropriate null hypothesis for this parameter.
  - The data are collected, and the sample mean from the 30 jars is 199.6 grams. The company's statistician carries out a 2-sided test of the null hypothesis, and finds that the  $P$ -value is  $P = 0.8$ . Assuming that the test was correctly carried out, describe, in words, the meaning of this  $P$ -value.
  - The statistician now finds a 95% confidence interval for the true mean weight from jars filled by the new filling machine. What are two things you can say about this confidence interval, based on the information in part (c)?

6.3 FEV<sub>1</sub> denotes the forced expiratory volume after one second. For ten year-olds (male or female), FEV<sub>1</sub> is Normally distributed with a mean of 2.6 and a standard deviation of 0.4. You are interested in the proposition that ten year-old children living in a smoking household (i.e. a household where one or more adults smoke) have a lower FEV<sub>1</sub> than ten year-olds in the general population.

- (a) What is the parameter of interest?
- (b) What is the null hypothesis of interest?
- (c) What is the two-sided alternative hypothesis?
- (d) You test a random sample of 20 ten year-old children, who live in a smoking household. The mean FEV<sub>1</sub> for the sample of 20 children is found to be 2.43.

What is the probability of obtaining a sample mean of 2.43 or less, if the distribution of FEV<sub>1</sub> among these ten year-olds is the same as for ten year-olds in the general population? (°)

What is the probability that the sample mean is more than 0.17 away from the population value?

What is the *P*-value for the test of the null hypothesis?

Recall that if  $X \stackrel{d}{=} N(2.6, 0.4^2)$ , then  $\bar{X} \stackrel{d}{=} N(2.6, \frac{0.4^2}{20})$ .  
Assume that the standard deviation of FEV<sub>1</sub> scores for ten year-old children from smoking households is 0.4.

- (e) What do you conclude about the mean FEV<sub>1</sub> for ten year-olds from smoking households?

## 6.12 Answers

- 6.1 (a) Ian's score on test can be thought of as the number of successes in 20 trials (20 questions) where the probability of success is  $1/5 = 0.20$ . Let  $X$  be Ian's test score, then  $X \stackrel{d}{=} \text{Bi}(20, 0.2)$ .
- (b) If we suspect that Sue has some idea, our research hypothesis is that Sue has some knowledge of Applied Aesthetics. An appropriate null hypothesis is that Sue does not know about Applied Aesthetics and that she would be guessing on the test. We can formulate this hypothesis in terms of the proportion of successes Sue will have on test items with five multiple choice answers:  $H_0 : \theta = 0.20$ . If we believe that the only possible consequence of some knowledge of Applied Aesthetics is that your proportion of successes would be *larger* than when guessing, then a one-sided alternative is implied, namely,  $H_1 : \theta > 0.2$ . Otherwise, we may take the two-sided alternative hypothesis:  $H_1 : \theta \neq 0.20$ . The parameter of interest is the true proportion of items that Sue gets correct.
- (c) If Sue is guessing, then Sue's test score  $X \stackrel{d}{=} \text{Bi}(20, 0.2)$ .  
 $\Pr(X \geq 10) = 1 - \Pr(X \leq 9) = 1 - 0.9974 = 0.0026$ .
- (d) If Sue was guessing, it is quite unlikely that she would get 10 items or more correct on a 20 item test. Her result is very surprising if the null hypothesis — that she is guessing — is true. This probability is the *P*-value for the one-sided alternative. If we use the two-sided alternative hypothesis, the *P*-value is twice the probability calculated in (c), namely,  $P = 2 \times 0.0026 = 0.0052$ .
- 6.2 (a) We can think of the amount of honey in a jar as a random variable. Ideally all jars would contain exactly 200 grams of honey, but in reality the amount will vary from jar to jar. If we think of all jars produced by the company, we can consider the distribution of the amount of honey in the population of jars. The parameters of this distribution are the mean and standard deviation of the amount of honey per jar. If the label states that the jar should contain 200 grams, we would hope that the mean amount of honey per jar would be 200 grams. The parameter of interest is the (population or true) mean amount of honey per jar ( $C$ ).
- (b) For the company, an appropriate null hypothesis (no deviation from the stated amount on the jar) is:  $H_0 : \mu = 200$ .  $\mu$  is the mean amount of honey per jar in the population of jars produced by the company.
- (c) The *P*-value is relatively large, suggesting that the (sample) mean of the sample of jars is not at all surprising if  $H_0 : \mu = 200$  is true.

A (sample) mean of 199.6 for a sample of 30 jars is quite consistent with the hypothesis that the true mean is 200 grams.

- (d) [1] As the  $P$ -value associated with  $H_0 : \mu = 200$  is 0.8, the data are consistent with the null hypothesis: hence the 95% confidence interval based on these data will include the value 200.
- [2] The confidence interval will be centred at 199.6 and, as argued above, will include the value 200. By symmetry, we know the confidence interval will also include the value 199.2.
- 6.3 (a) We are considering the distribution  $\text{FEV}_1$  for ten-year-old children living in smoking households. As we are interested in a general proposition about these children, the distribution of interest is the population distribution. As  $\text{FEV}_1$  is a continuous measure, the parameter of interest for this distribution is  $\mu$ , the population mean.
- (b) The null hypothesis is that the mean  $\text{FEV}_1$  for ten-year-olds from smoking households is no different from the general population of ten year-olds, i.e.  $H_0 : \mu = 2.6$ .
- (c) The alternative hypothesis is that the mean  $\text{FEV}_1$  for ten-year-olds from smoking households is different from the general population of ten-year-olds, i.e.  $H_1 : \mu \neq 2.6$ . This is a two-sided alternative hypothesis.
- (d) Let  $X$  be the  $\text{FEV}_1$  for ten-year-old children living in smoking households. We assume that the distribution of  $\text{FEV}_1$  for children from smoking households is the same as children in general:  $X \stackrel{d}{=} N(2.6, 0.4^2)$ . Then,  $\bar{X} \stackrel{d}{=} N(2.6, \frac{0.4^2}{20})$ .  
 $\Pr(\bar{X} \leq 2.43) = 0.0286$ . Hence the two-sided  $P$ -value is  $P = 2 \times 0.0286 = 0.057$ .
- (e) The  $P$ -value is relatively small; we need to know more about the clinical implications of the lower mean  $\text{FEV}_1$  for ten-year-olds in smoking households. We should also consider the 95% confidence interval for the true mean: (2.25, 2.61).

## 7 Models for data

There are a large number of discrete and continuous distributions; we have seen only a few. They are vitally important in statistical science, forming part of the framework in which formal inferences occur.

### 7.1 Distributions for modelling data

We have already seen distributions used to model data. Previously, however, we motivated the relevant distribution from first principles, for example, by thinking about the physical device of a die, and how it is rolled, or considering the way random sampling works.

Here, the perspective shifts, and we start to see how inferences use models for data, without there being a clear or simple probabilistic mechanism.

▷ **EXAMPLE. Dockets at a cafe**

A customer gets a coffee at a cafe most afternoons, somewhere between 2 pm and 6 pm, and usually later in the day. The cafe uses a docket system to keep track of everyone's order. The machine printing the dockets starts at 1, prints consecutive numbers until 99, and then starts at 1 again. It is reset each day; when the cafe opens in the morning at 6:30 am the first docket is '1'.

The customer collects the docket he receives each day and notes the number. Over a period of years, he has a sample of 1094 dockets.

What distribution will the customer's number follow? A distribution we might consider is the integer distribution for the numbers  $1, 2, \dots, 99$ , with equal probability for each outcome. Is that right? Are the data consistent with this? Before we think about that model further, we look at the data.

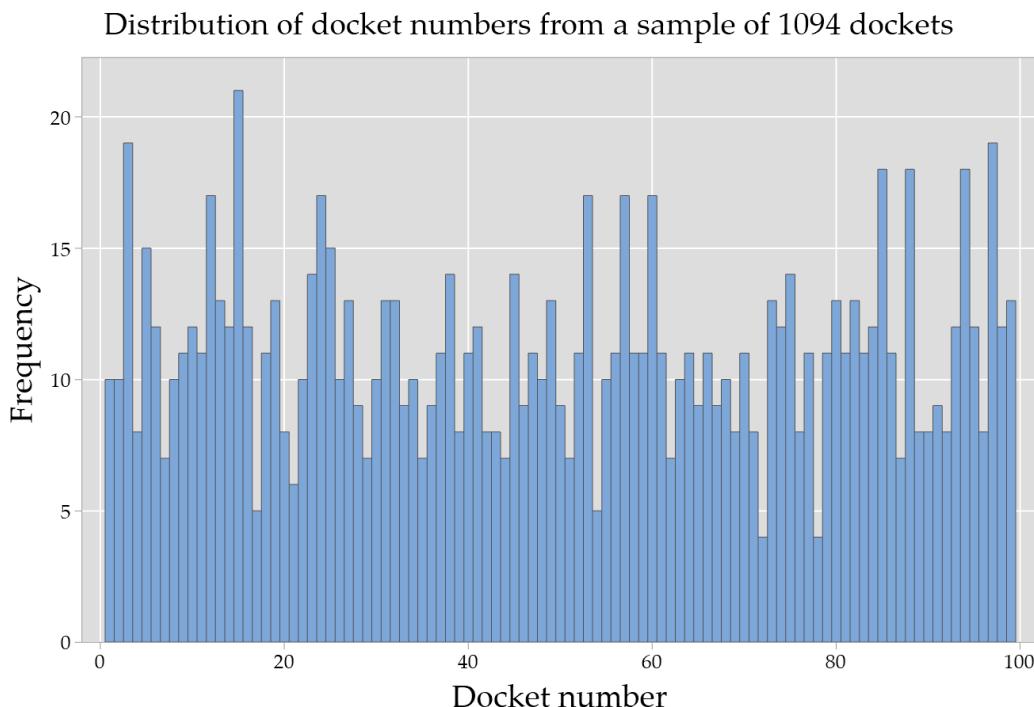


Figure 69: Distribution of the docket number (1 to 99) for 1094 dockets collected from a cafe.

A histogram is shown in Figure 69. It has an unusual construction; generally speaking, it would be preferable to have some collapsing into bins of wider intervals. This could be done here, although for assessing the proposed model, we would need to be careful: it would be desirable to have bins arranged so that they include the same number of integers in each case, and this could only really be done by 11 bins of ‘width’ 9 (the first one including 1,2,…,9, etc., or nine bins of ‘width’ 11 (the first one containing (1,2,…,11, etc.).

The unusual construction is deliberate here because it shows the complete distribution and, with care, the grid lines allow us to read off the distribution without error. For example, we see from Figure 69 that among the sample of 1094, there were (exactly) 10 dockets printed with the number ‘1’, also 10 with ‘2’, 19 with ‘3’ and so on.

Now we consider what might be a plausible model. If the integer distribution is assumed (like rolling a die), the expected number of times each docket number appears would be  $1094 \times \frac{1}{99} = 11.05$ . We see that there are indeed many counts of 11, and quite a few counts of 10 and 12 too.

But ... isn’t this completely different from rolling a die? The machine can give exactly 99 outcomes, like the 6 we can get from a fair die, but why should they be equiprobable? If the customer made sure that he was the first person to be served each day, he would get docket number ‘1’ every day. If

he came in the first 20 minutes of the day, his docket numbers would tend to be lower numbers and he might never have a docket number over 30, say. What might justify thinking that the integer distribution is appropriate? The actual number he obtains each afternoon depends on the sequence of events each day, on the combined decisions of hundreds of customers and their arrivals on the day, and the timing of his own arrival. How is this random? Could we consider the outcomes to be equiprobable?

This is a simple example of real data in which we may postulate that a particular statistical model applies, based on a combination of considerations.

The first is that the intrinsic variation in the cycles of the machine, the pace of different days, and the timing of the customer, could indeed combine to produce data that is essentially coming from the integer distribution: there is enough chaotic disturbance of timings by the afternoon that it really is like rolling a 99-sided die.

The second is empirical: we look at the distribution and ask ourselves: is it consistent with the integer distribution? This is a version of the perspective in Chris Wild's diagram that we looked at in Chapter 4, Figure 39, in which we *see* the empirical distribution of the data and we *imagine* an underlying statistical model, and ask ourselves how they match up.

Look again at Figure 69. Do you think the data could be considered a random sample of size 1094 from the integer distribution? There doesn't seem to be a concentration of numbers in any area. We can clearly see from the histogram that all 99 docket numbers are represented. The smallest count is 4, for two docket numbers (72 and 78) and the largest count is 21 (docket number 15). The visual impression suggests consistency with the integer distribution.

Given what we have learned in Chapter 6, we may reasonably ask: "Can we test the hypothesis that these data come from the integer distribution?" This question is a different type from the hypothesis tests we have considered so far, which have been about a single parameter. However, it is a hypothesis about a distribution underlying the data, and hence is, indeed, a question that can be addressed using a hypothesis testing framework. We do not cover the theory for this here, but — for what it is worth — the  $P$ -value for testing that these data come from the integer distribution on the numbers  $1, 2, \dots, 99$ , is  $P = 0.362$ . Hence the data are consistent with the integer distribution.

#### ▷ EXAMPLE. Train trip times

Consider a study of a particular train trip in a timetable, from an outer suburban train station to a central city station. The purpose of the study is to check whether the actual times are adhering to the train timetable. The train's departure time and arrival time are recorded for 250 days. The times are measured to one second accuracy, so that times such as 42 minutes and

7 seconds (42:07) are recorded.

The average time recorded, to the nearest second, was 2598 seconds (= 43:18 minutes, or 43.30 minutes as a decimal). The minimum time was 2466 seconds (= 41:06 minutes, or 41.10 minutes), and the maximum time recorded was 3747 seconds (= 62:27 minutes, or 62.45 minutes).

If we regarded this random variable as discrete with possible values at every second ( $\dots, 2800, 2801, 2802, \dots$ ), and sought to model it using a discrete distribution, we would need to come up with probabilities for each second separately. In the absence of a theoretical basis for doing so, we might consider using the data gathered, to estimate these probabilities.

Think of the consequences of doing this. There are 1282 discrete times (to the nearest second) in the range of the data, from 2466 to 3747 seconds. With  $n = 250$  observations, most of these discrete times will not appear in the data, that is, they will have a frequency of zero. Of the rest, most will have a count of one, some will have two, and a handful of the discrete times might have occurred three or more times in the data set.

It would therefore be quite cumbersome to model the times as discrete, and not effective. It is convenient to think of the times as coming from an underlying theoretical distribution which is continuous. We can then look at the continuous distribution and its properties, to understand more about the pattern of the train trip times.

A related point is that it is unlikely that we would ever need to consider the lengths of the train trips at the detailed level of the individual, discrete times. It is much more likely that we would be interested in, say, the percentage of trips between 42 and 43 minutes, or the fraction of trips that are five minutes late or worse, and so on, rather than the chance that a train trip is 43 minutes 30 seconds.

Figure 70 shows a histogram of the 250 train trips. Alongside it is the probability density function (pdf) of a continuous random variable  $X$ . This pdf is given by the following:

$$f(x) = \frac{\exp - [\ln(x - 41) - 0.4]^2 / 2}{\sqrt{2\pi}(x - 41)}, \quad x > 41.$$

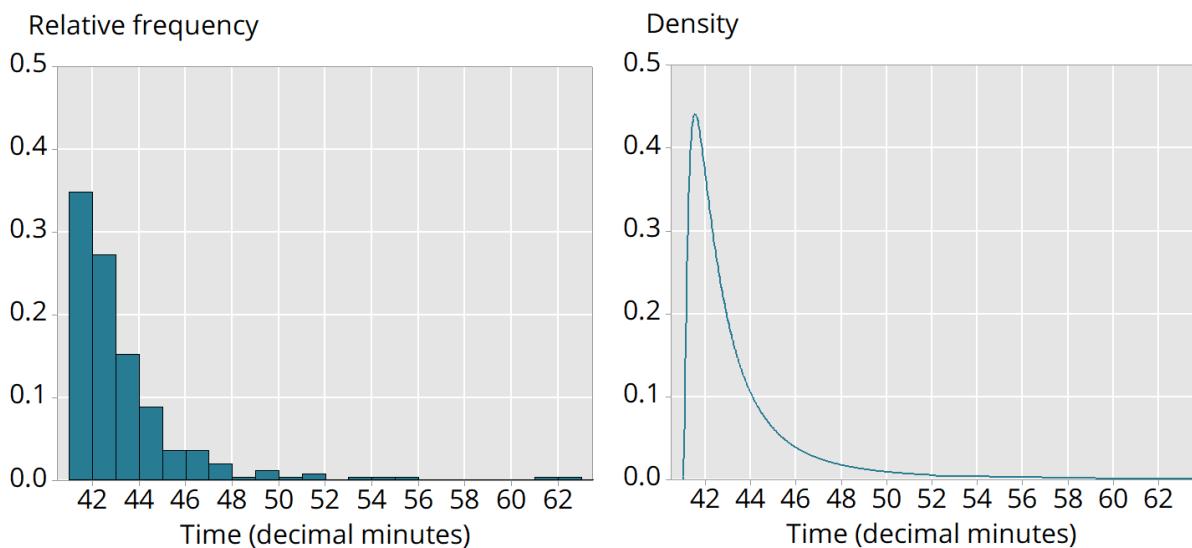


Figure 70: Comparison of data on train trip durations ( $n = 250$ ), showing relative frequencies, and a corresponding probability density function for a postulated model.

The relative frequencies in the histogram correspond to the areas under the curve in the probability density function. We can assess whether the model is a good fit to the data by looking at the probabilities from the pdf, and asking how close they are to the relative frequencies in the histogram. There were 87 trips between 41:00 minutes and 41:59 minutes; this is a relative frequency of  $87/250 = 0.348$ . How close is this to the probability implied by the pdf? Calculating the area under the curve for this pdf is beyond the scope of this course; it is found to be 0.345, which is very close to the relative frequency. The following table shows the relative frequencies, and the probabilities from the pdf, for the first five one-minute intervals.

Time interval	Number	Relative frequency	Probability from model
41:00 - 41:59 minutes	87	0.348	0.345
42:00 - 42:59 minutes	68	0.272	0.271
43:00 - 43:59 minutes	38	0.152	0.142
44:00 - 44:59 minutes	22	0.088	0.081
45:00 - 45:59 minutes	9	0.036	0.049

The train trip time illustrates the way that continuous distributions are often used, as a useful approximation to a discrete random variable, when the discrete random variable is on a very fine scale. This occurs in a wide variety of contexts, such as measurements of human heights (centimetres), IQ (integers), test scores or exam marks, and so on.

In circumstances where we do not have a clear basis for choosing a particular pdf to model data of this sort, the relative frequencies from the histogram serve as a guide: we obtain estimates of probabilities for given intervals, directly. We then look for a continuous distribution that can closely reflect

these estimates.

## 7.2 Models for residual variation

The second important way in which distributions are used in applied analytics is in statistical models with a systematic and random component. Actually, this is just an extension of the modelling of data (above), which can be framed as a basic case of this.

You will encounter many such models in statistical science, and this idea will be developed further in subsequent chapters.

In the general context of such models, there is a component that aims to capture some of the variation in a response variable by a systematic component, expressed as a mathematical function of one or more explanatory variables. However, the systematic component does not explain all of the variation in the data. There is always some “left-over” or residual variation that is thought of as ‘purely’ random, in the sense of being unexplained by available explanatory variables.

The nature of unexplained variation can take different forms. A very large number of statistical models assume that the unexplained variation can be modelled by a Normal random variable with mean zero and unknown constant variance. We will see many applications of this; the various specific models (regression, analysis of variance, analysis of covariance) come under the rubric of the “linear model”.

Another key example of this is logistic regression, in which the binomial distribution features. Essentially, data of a binomial nature are modelled in such a way that the probability of success is itself related to explanatory variables.

There are many other statistical models that fit into this overall structure, with the application of yet more underlying distributions to capture the unexplained variation.

## 7.3 Model complexity

There is a famous quote, generally attributed to George Box, one of the eminent mathematical statisticians of the 20th century. He expressed it a variety of ways in different publications, but it is usually reported as follows.

“All models are wrong, but some are useful.”

Box’s point was that all models are approximations, involving assumptions — implied or otherwise — that are never exactly true. It is obvious that the complexity of the natural world, and measurements in a research context,

cannot be reduced to a few simple mathematical formulae. But a model may be *adequate* or even *good* for a purpose such as decision making, inference about causation, estimation or prediction. And great advances in science have been made by the use of relatively simple models.

This makes statistical science difficult, and also interesting: there is a challenge and a skill in learning what to worry about. Which assumptions matter? What aspects of the model are robust, meaning that they will perform well, or satisfactorily, even if the assumptions are not met? When there is evidence of departure from assumptions and it is a matter of degree, what extent of violation of the assumptions matters?

These are demanding questions. We touch on some of them in this course, but they do not get extensive coverage.

### 7.3.1 Simple models

One of the simplest models we encounter in statistics is (simple) linear regression. It says that the average value of a response variable  $Y$ , is a linear function of an explanatory variable,  $x$ . We will look more closely at this model in chapter ??; it is mentioned here as an illustration of George Box's quote.

In almost any situation in which a simple linear regression is used, no one seriously believes that there is truly a straight line relationship as implied. Rather, the linear relationship may be a sufficiently adequate model to gain some insight into the association between the variables, or even, if  $x$  is chosen by design and its levels randomised, to draw causal inferences. It might prove to be adequate for prediction, with some associated imprecision that is able to be estimated.

However, this does not mean that simple linear regression is guaranteed to be adequate or useful, as we shall see. The viability of the model and its assumptions always need to be considered, according to first principles, and by empirical inquiry. We look at these approaches further in later chapters.

### 7.3.2 Complex models

In many fields of research, models of considerable complexity are used. There can be a real tension between the claimed merits of a complex model, on the one hand, and the accessible elegance and simplicity of a basic model, on the other.

Some general areas where complex statistical models are used include the following.

- Macro-economic models that attempt to capture the relationships be-

tween important variables in a nation's economy, such as demographic changes, unemployment, economic growth, debt, tax and so on.

- Climate change models that consider the inter-connectedness of green-house gases and other pollution and changes in temperature, rainfall, sea level, the melting of solar ice caps, and extreme events.
- Weather models that use inputs from weather stations around the world to make forecasts of the weather at a particular location.

Complex models like these all have features common to those we have considered: response variables (possibly several), postulated relationships with explanatory variables, and random variation of an assumed form.

## 8 The linear model — a broad view

In the previous chapter we considered the general idea of a statistical model. Here we introduce a very extensive class of models known as the ‘linear model’. This class not only includes many specific forms of model, it also leads naturally into extensions, making yet other classes of model. A very high fraction of statistical models used in practice are linear models, or extensions.

The generic form of this model has these features:

- A single response variable  $Y$ . This is the random variable that we hope to be able to predict, or explain.
- A function of one or more explanatory variables, which expresses the average value of  $Y$ , for any given values of the explanatory variables. This function is linear in the way that unknown parameters appear in the function.
- Random errors, which, in the standard linear model, are assumed to be independent and Normally distributed with mean zero and standard deviation  $\sigma$ ;  $\sigma$  may be described as the ‘residual standard deviation’ and is a parameter that must be estimated. The mean of zero is not really an assumption since the appropriate centering of the data can be incorporated in the linear function of the parameters; in general, there will be a constant term.

The model can then be expressed as

$$Y = \text{linear function} + \text{random error}.$$

When we are being mathematically precise, we need subscripts and indices to capture the details of the model, but this shows the basic idea.

### 8.1 Simple linear models

To understand the linear model, we start with a couple of simple examples. The first is the simplest model — so simple, in fact, that in practice we do not really consider it. But its simplicity aids understanding, and we can build up from the simplest case.

#### A single mean

Consider a random sample  $Y_i, i = 1, 2, \dots, n$  from the Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . We looked at this context when discussing a confidence interval for  $\mu$  in chapter 5.

To fit this in the general linear model framework, we write the structure of the model this way:

$$Y_i = \mu + E_i, \quad i = 1, 2, \dots, n$$

where we assume that the random error terms,  $E_i$ , are independent and have a Normal distribution with mean zero and standard deviation  $\sigma$ .

Note that we don't get to observe the random errors. Why not? Why can't we subtract off  $\mu$  from both sides, to obtain  $E_i$ ? We can, but  $\mu$  is a parameter, thought of as a fixed but unknown number, so we will not actually know the value of  $E_i$ . We return to this point when thinking about assessment of the assumptions of the model.

Sometimes we think of the model this way. The observations are centered at  $\mu$ ; that is where they 'start'. An additional, unpredictable amount is then added: this is the random error. In this perspective, the random error encapsulates all the left-over, unexplained variation, which includes all the variation we don't identify.

Is this new? Is it a departure from what we assumed when we worked out a confidence interval for  $\mu$  in chapter 5? No: it is just a different way of looking at the same assumptions. In fact the model above implies that  $Y_i \stackrel{d}{=} N(\mu, \sigma^2)$ , as before.

### Simple linear regression

This can also be framed as a general linear model example. The observations consist of pairs  $(x_i, Y_i), i = 1, 2, \dots, n$ .

We conceptualise  $x$  as a fixed explanatory variable and hence use lower case  $x$ ; it is non-random.  $Y$  is a random variable: the response. We write the simple linear regression model as

$$Y_i = \alpha + \beta x_i + E_i \quad i = 1, 2, \dots, n,$$

where the  $E_i$  terms are random errors, assumed to be independent, Normally distributed with mean zero and standard deviation  $\sigma$ . In this case the mean value of  $Y$  depends on  $x$ . We can think of  $Y_i$  as 'starting' at  $\alpha + \beta x_i$ ; the random errors are then added, taking the observation up or down from the line, depending on the size and sign of the random error.

We will consider this model and extensions of it in Chapter ??.

You should have seen by now the commonality between the different cases. For one thing, the statement of the random errors as independent, Normally distributed with mean zero and standard deviation  $\sigma$  is something that is repeated every time: it is the same assumption in each case. The second common element is the response, a random variable, on the left hand side of the equations. The difference between the cases is (only) the actual form

of the ‘systematic’ part of the model.

## 8.2 Response variable

Consider, again, the general form of the linear model:

$$Y = \text{linear function} + \text{random error}.$$

$Y$  is the (numerical) random variable that we hope to be able to predict, or explain. We think of this as the key response or outcome variable.

Commonly, the choice of  $Y$  is determined by the context: it is likely to play a key role in the research question we are asking. In a study of the superannuation balances of women workers, the balance might be the response variable, and personal, work and demographic variables might be potential explanatory variables.

In other contexts,  $Y$  is not so obvious, and may be chosen as the variable of interest because it is hard to measure, and we therefore want to develop a model that predicts it using variables that are more accessible and simpler to record. A study of elephant characteristics modelled the weight of the elephant in terms of the chest girth and sex of the animal. Here the weight was taken as the response variable  $Y$ . Chest girth and weight are both just physical characteristics of an elephant. We could be asking about predicting chest girth from the weight. But the chest girth is relatively easily measured (although you might want some training and assurances before attempting it), whereas the weight, which is important to know, is difficult to measure without an elaborate process. So a linear model can help, if chest girth (and other variables) work well as predictors.

Then there are cases where a number of variables could be considered as the outcome variable  $Y$ , and the choice is really governed by the particular research focus. In a study of environmental dispositions of young people, researchers might measure their knowledge of environmental issues, their involvement in environmental activism, and their support for environmental action. Is the response variable obvious? Not necessarily. One might model ‘involvement’ as the outcome with ‘knowledge’ and ‘support’ (and perhaps other factors) as explanatory variables. But it could also be reasonable to model ‘support’ as the  $Y$  variable.

This discussion is a reminder that statistical modelling must always relate coherently to the question being addressed.

## 8.3 Linear function

The linear function of one or more explanatory variables does the work of predicting or explaining the response. It offers insights into the way the re-

response variable changes, on average, for different configurations of the explanatory variables. There are two aspects to this: the variables themselves, and the form of the linear function.

When there are many candidate explanatory variables, choosing which ones to use is an issue that must be dealt with. This topic is the subject of much research. Imagine a situation with 20 potential explanatory variables. We could easily have a big data set with this property. How many possible models are there? Even if we restrict ourselves only to models in which a variable is present in the simplest way (for example, as a linear term) there are a very large number of possible models. The first variable may be included or not (2 choices); the second variable may be included or not (also 2 choices) and so on up to the 20th variable. These possibilities multiply, so there are  $2 \times 2 \times \dots \times 2 = 2^{20}$  possible models.  $2^{20}$  is more than a million, and that is a huge number of possible models to consider. There are some disciplines, such as genetics, where the number of candidate explanatory variables is itself very large – hundreds, or thousands – making the problem even more acute.

Then there is the form of the linear function. A key point to understand about the nature of the linear model is that the ‘linearity’ is all about the unknown parameters of the model, rather than explanatory variables. This is best explained by some simple examples; here we avoid the use of subscripts for the data, to focus on the form of the model. The following two models are linear models.

$$Y = \alpha + \beta \log(x) + E$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2 + E$$

At the start of these and other previous equations you have seen different Greek symbols,  $\alpha$  and  $\beta_0$  in these two models. In each case, they represent a ‘constant’ or ‘intercept’ term: think of it, roughly, as representing the mean level of  $Y$  when the explanatory variables are at their baseline levels, commonly, zero. Generally speaking, this term is not of direct interest, but we need to include it, otherwise we are assuming a model in which the mean value of the response is equal to zero at baseline levels of the explanatories. That will usually not be the case.

This ‘constant’ term is represented by different Greek letters sometimes. Does this matter? No, and the variety is a good thing to get used to, because authors write models in several ways and they do not necessarily use the same symbols.

The two previous models are linear models; on the other hand, the following model does not fit in the general linear model framework, because the relationship of the response variable with the parameters is not linear.

$$Y = \alpha x^\beta + E.$$

Note, however, if we take logarithms of the systematic part of this model we obtain

$$\log(\alpha) + \beta \log(x).$$

We may be prepared to consider the following, different model, in this case:

$$U = \log(Y) = \log(\alpha) + \beta \log(x) + E.$$

In this modified form, we are back to an example of a general linear model: a simple linear regression, with response variable  $U = \log(Y)$  and explanatory variable  $\log(x)$ .

Note that the random error term  $E$  is generic, and not the same as in the non-linear model above.

## 8.4 Random error

The linear function of the explanatory variable cannot fully explain the response variable; if it did, we would not really be considering statistical models, but completely specified functional relationships. Among a group of four-legged tables, the number of legs is four times the number of tables. The amount of income tax payable for a given annual taxable income, according to the tax rules in Australia, is a more complicated function, but there is no uncertainty in the result. Functional relationships such as these are important for various purposes, but they are not what statistical modelling is about. Recall that variation is at the core of applied analytics, and the contexts we are considering have the fundamental feature that the linear function of the explanatory variable only partially explains the response variable.

The way that this is dealt with is to allow an extra term in the linear model, known as the random error. It captures the variation which goes by a variety of names, all attempting to define the same idea: the ‘left-over’, ‘residual’, ‘background’, ‘unexplained’ variation.

There are three important assumptions made about the random errors in the typical case, and it is by no means guaranteed that, in general, they are correct.

- Independence: The random error terms are assumed to be independent, for different observations. This is an assumption that needs to be addressed by considering the structure of the data. There are many possible causes of correlated random errors. A simple structure involving non-independent random errors is a time series; usually there will be ‘serial correlation’ in the random errors for observations close together in time.

Another common source of correlated random errors occurs if there are clusters of similar units in the data. In a study of suburban behaviour, a cluster might be a household; in an educational survey, a cluster could be a school or a class; in an animal study a cluster could be a pen or litter of animals. Such clusters lead to units within a cluster being more similar to each other than units in different clusters; this violates independence. There are ways to accommodate this, but not within the standard linear model we consider here.

- Constant standard deviation: The random errors are assumed to have the same standard deviation  $\sigma$ . Again: there is no law of nature that says that the residual standard deviation must be constant, and there is often quite strong evidence that it is not so. This assumption can be examined empirically; we discuss this further in Chapter 11.
- Normality: Finally, the random errors are assumed to follow a Normal distribution. What are the grounds for this assumption? If we believe that the unexplained variation comes from a large number of minor sources, then it is likely that the errors are Normally distributed, due to the ‘Central Limit Theorem’: the sum of a large number of independent random variables tends to a Normal distribution. The nature of the residual variation does not have to be of this form, and when it is not, the random errors may show non-Normality. For example, there may be a single contribution to the random errors that is occasionally large and positive, leading to skewness, and the occasional very large random error.

It is useful to have a perspective on the relative importance of these assumptions. There is some sort of hierarchy: the first two are of higher priority than the last.

If the observations are not statistically independent, assuming that they are can lead to badly incorrect inferences: generally, you will be claiming more precision than is justified. To correctly deal with lack of independence, the right model needs to be fitted; these are beyond the scope of this subject, but you need to be aware of the issue and the potential for problems.

Ignoring this issue is a common error and a serious one.

▷ **EXAMPLE. Multiple blood pressure measurements**

Suppose we conduct a randomised controlled trial to reduce blood pressure in patients with hypertension. There are 20 subjects in each of the treatment group and control group. The outcome is diastolic blood pressure (DBP). At the relevant time point, the DBP is measured five times on each subject. The data analyst proposes to compare the two groups with  $n_1 = 100$  observations in the treated group and also  $n_2 = 100$  from the control group, treated as 200 independent observations. Does this seem correct to you?

It should not: the five repeat measurements on an individual are not statistically independent. The conclusions will not be justified if the analysis assumes they are: typically, the analysis will give a (falsely) smaller  $P$ -value and a (falsely) narrower confidence interval than the correct analysis.

Violation of the constant variance assumption can also lead to misleading inferences. This issue can often be dealt with, for example, by a transformation of the data.

Finally, the assumption of a Normal distribution for the random errors is arguably the least important assumption, because of the Central Limit Theorem. For large sample sizes, the inferences we draw about parameters of the model entail an underlying averaging process, so the sampling distribution of the estimators of the unknown parameters will tend to have a Normal distribution, even if the random errors themselves do not. This is why the assumption of Normality of the random errors is less important than the other two.

There is a simple misconception that naïve analysts sometimes have, and some textbooks reinforce. They become concerned that ‘my data are not Normally distributed’. If this conclusion has been drawn by exploring the distribution of the data overall, it may be a misplaced concern. Systematic structure in the data can lead to patterns overall that are distinctly non-Normal in shape. But if this structure can be captured in the linear function, using explanatory variables, it may be that the residual variation has an acceptable distribution, meeting the required assumptions. The assumptions listed previously are all about the residual variation, not the response variable itself.

We consider these assumptions further in Chapter 11.

## 8.5 Types of explanatory variables

When a linear model is used, there are two types of explanatory variables: factors and covariates.

- A **factor** is a categorical variable, one that takes two or more levels. The simplest type of factor has two levels, and this is commonly required, for variables such as treatment group (active treatment or placebo) or any other explanatory variable with two levels. These are sometimes called ‘binary’ or ‘dichotomous’. A factor can have more than two levels and commonly does, but generally this is a relatively small number; it is unusual and unwieldy to use a factor with many levels.
- A covariate is a variable on a numerical scale, such as pressure, a monetary amount, or (see the elephants case study) chest girth. A covariate does not have to be used on its original scale; a transformation of

it may be used, such as the logarithm or power of the variable, if that proves useful or desirable. That is, if the explanatory variable is  $x$ , we may use  $\log(x)$  or  $x^k$  for some power  $k$ .

In one version of this kind of usage,  $x$  and integer powers of  $x$  may be used as explanatory variables. Sometimes there is evidence of curvature in the relationship between the response variable and an explanatory variable  $x$ ; we then might consider using both  $x$  and  $x^2$  in a linear model of this form:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E.$$

It might occur to you to wonder: is this a linear model? Doesn't the inclusion of the  $x^2$  term make it non-linear? No, it does not: recall that the linear form is about the way the unknown parameters are present in the model; not the explanatory variables. Viewed as a function of the parameters  $\beta_0$ ,  $\beta_1$  and  $\beta_2$ , the form of the function is linear.

The taxonomy of variables includes one other type: ordinal. How can these be used in a linear model as predictors? Not straightforwardly!

This is a context that illustrates the challenging nature of an ordinal variable. Recall that an ordinal variable is at least categorical. So, we can always treat it legitimately as a factor, although we may feel that we are losing some of the potential sensitivity of its effect then, since we are then not taking into account the order. On the other hand, if we code it numerically in equal intervals, such as  $1, 2, \dots, k$  (something we have counselled against, elsewhere), and use it in that form as an explanatory covariate, we are assuming that estimating the change in the response variable per one 'unit' of the ordinal variable is reasonable. And it may not be reasonable.

## 9 Inference — numerical outcome, paired samples

In Chapter 1 we considered data from a paired design and in Chapter 5.2.3, inference on the mean difference from such data. The inference assumed an underlying Normal distribution.

We return to this context here, moving to a focus on hypothesis testing. Different tests are presented, which depend on the assumptions about the underlying data generating mechanism, or “population”.

### 9.1 Assuming Normality for the data: the $t$ test

▷ **EXAMPLE. Fuel economy, engine tuning study (oz\_cars\_1.6\_litres.mwx)**

This example has been considered in Chapters 1 and 5. It involves data on fuel consumption from 22 cars with 1.6 litre engines, pre- and post-tuning. As we have seen, it is statistically efficient to consider the differences between the pre- and post-tuning data when drawing inferences. These differences are shown in the following table.

Car	1	2	3	4	5	6	7	8	9	10	11
Pre tuning	11.51	10.75	11.00	10.77	11.32	11.00	10.68	9.47	10.90	9.50	9.98
Post tuning	11.04	10.27	11.27	9.42	11.20	10.79	10.46	10.02	10.61	9.56	10.47
Difference	0.47	0.48	-0.27	1.35	0.12	0.21	0.22	-0.55	0.29	-0.06	-0.49
Car	12	13	14	15	16	17	18	19	20	21	22
Pre tuning	9.72	8.04	10.00	10.47	11.01	9.39	9.45	9.04	10.78	10.47	9.93
Post tuning	9.55	7.71	9.41	10.40	9.32	8.10	9.62	8.94	10.64	10.03	10.19
Difference	0.17	0.33	0.59	0.07	1.69	1.29	-0.17	0.10	0.14	0.44	-0.26

We saw the dotplot of the differences in Chapter 5 and repeat it here for convenience.

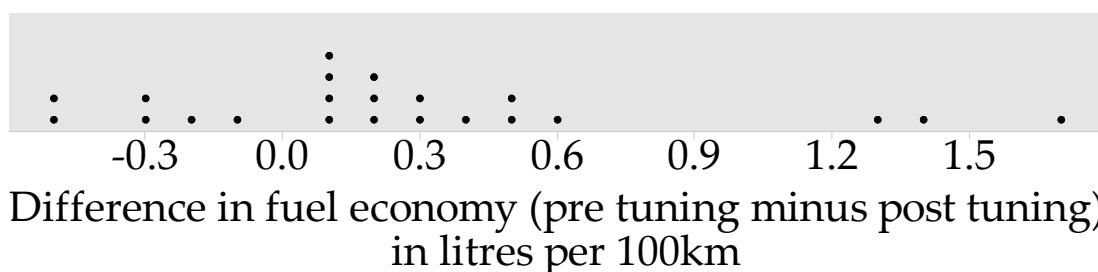


Figure 71: Dotplot of the differences in fuel economy (litres per 100 km), pre-tuning minus post-tuning, for 22 cars with 1.6 litre engine capacity

In Chapter 5 we found that the sample mean difference was  $\bar{d} = 0.28$  litres per 100 km, the sample standard deviation of the differences was  $s_D = 0.564$ ,

and, assuming that the data are a random sample from a Normal distribution, a 95% confidence interval for  $\mu_D$  is (0.03, 0.53).

We now turn our attention to testing the null hypothesis that the true, population mean difference,  $\mu_D$ , is equal to zero. That is, we test  $H_0 : \mu_D = 0$ . Everything said about hypothesis testing in Chapter 6 is true in this situation. It is preferable to focus on estimation here, and to consider the precision of the estimate, reflected in the confidence interval. There may be some interest in testing the null hypothesis, from a policy point of view, for example.

First, some theory that was used in Chapter 5, but which we now make more explicit.

In this context, the assumption of Normality for the underlying data is important; this is not a Central Limit Theorem application. For a random sample of differences  $D_i, i = 1, 2, \dots, n$  from a Normal distribution,  $\bar{D} \stackrel{d}{=} N(\mu_D, \frac{\sigma_D^2}{n})$ . We can standardise this result to obtain:

$$\frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{n}} \stackrel{d}{=} N(0, 1).$$

If we knew the value of  $\sigma_D$ , we could use this result to carry out the test. But we have now abandoned unrealistic scenarios where population parameter values are known. The result that can be applied is as follows:

$$\frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \stackrel{d}{=} t_{n-1}.$$

We are using here exactly the same set of assumptions, and hence the same statistical approach, as that used in Section 5.2.3

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

The test makes use of the result that, *when  $H_0 : \mu_D = 0$  is true,*

$$T = \frac{\bar{D}}{S_D / \sqrt{n}} \stackrel{d}{=} t_{(n-1)}.$$

This is the test statistic in this case: the quantity whose distribution is known, if  $H_0$  is true.

If the null hypothesis is not true, then  $\bar{D}$  is really positioned around a true mean other than zero, and that will show in an extreme value for the test statistic  $t = \frac{\bar{D}}{S_D / \sqrt{n}}$ .

The form of this test statistic, which is quite a general one, corresponds to the common form of confidence interval,  $\text{estimate} \pm k \times (\text{standard error})$ , noted in Section 5.3.5. This parallel test statistic form, for testing  $\mu_D = 0$ , is

$$\frac{\text{estimate}}{\text{standard error}}$$

arising in the same way as the general form of the confidence interval: either the estimator is approximately Normally distributed (large sample, Central Limit Theorem), or the  $t$  distribution is being used because Normality is assumed for the data, and the inference is about a mean or difference of means.

The procedure is as follows. Find the test statistic  $t = \frac{\bar{d}}{s_D/\sqrt{n}}$ . Then use the  $t_{(n-1)}$  distribution to determine the  $P$ -value, which is the chance of getting a value of the test statistic at least as extreme as that actually observed.

In MINITAB: You can carry out the test in one of two ways, which correspond to the two ways we worked out a 95% confidence interval for such data:

- Without creating a column of the differences, you can use Stat > Basic Statistics ▶ Paired t.... If you ask for a Graph (say, a dotplot) you get a display that is useful for description and inference. It shows the distribution of the differences, the position of the null hypothesis, and a 95% confidence interval for  $\mu_D$ .
- Alternatively, if you have a column of the differences, you can use Stat > Basic Statistics ▶ 1-sample t.... Recall that we used this command for a confidence interval for a population mean, assuming Normality, and these two procedures fit together in exactly the way described in Section 6.10.

For the cars example, we find that the test statistic is  $2.33 = \frac{0.28}{0.564/\sqrt{22}}$ , which we compare to the  $t_{21}$  distribution. This gives a  $P$ -value of  $P = 0.030$ .

In practice, we use the  $t$ -test when it is reasonable to assume that the distribution of the differences is Normal, at least approximately.

This setting can be viewed as a very simple form of the linear model described in Chapter 8. The differences,  $D_i$ , are modelled thus:

$$D_i = \mu_D + E_i, \quad i = 1, 2, \dots, n$$

where we assume that the random error terms,  $E_i$ , are independent and have a Normal distribution with mean zero and standard deviation  $\sigma_D$ .

## 9.2 Distribution-free tests

Sometimes we may be concerned about whether the assumption of Normality is reasonable, in small samples. It is possible to carry out a formal test of the assumption, and we will look at this test later.

One approach to use that relaxes the assumption is the “bootstrap”. Another is a “randomisation test”. These are both very useful techniques; they are not covered here.

Another general analytic strategy is to use so-called “distribution-free” or “non-parametric” tests (actually, neither of these names is entirely appropriate). These are a variety of statistical tests, some with corresponding confidence intervals, that do not assume a *particular* underlying distribution for the population from which the sample came, or, more generally, the data-generating mechanism.

For the context of paired data we consider two such tests in some detail, to introduce the ideas.

## 9.3 The sign test

One way to make inferences on paired data is to use the sign test procedure. The basic approach can be used to find a confidence interval for the median difference, or to carry out a hypothesis test for the null hypothesis, that the distributions of the outcome variable are the same for both groups.

### 9.3.1 Hypothesis test

The sign test is based on the number of positive (or negative) differences among the non-zero differences. Zero differences are ignored in the sign test. The only assumption required is that the differences are independent.

$H_0$ : The distribution of the outcome is the same for both settings (example: there is *no* difference in the distribution of fuel economy obtained pre- and post-tuning).

$H_1$ : The distribution of the outcome differs between the two settings (example: there *is* a difference in the distribution of fuel economy obtained pre- and post-tuning).

*If the null hypothesis is true*, all features of the distributions of fuel economy will be the same for both pre- and post-tuned cars: the means, medians, standard deviations ... This, in turn, means that the distribution of the differences in fuel economy in a sample of cars will be centred around zero. Some differences will be positive, some negative, only because of random variation in measurements on the same car at different times.

We arbitrarily focus on the number of positive differences,  $X$ . (We could use the number of negative differences; the inference would be the same.)

The sign test makes use of the result that, when  $H_0$  is true,  $X \stackrel{d}{=} \text{Bi}(n, 0.5)$ , and  $n$  is the number of non-zero differences; here  $X$  is the test statistic.

The idea behind the test is that — if the null hypothesis is true — the chance that a difference in the data is positive or negative is the same, and hence equal to 0.5. There is no tendency for the differences to be in a particular direction.

The formal process involved in the test is as follows:

1. Determine the sign of the difference between the two members of each pair.
2. Determine  $n$ , the number of non-zero differences, and  $x$ , the observed number of positive differences. Zero differences are left out of the calculation altogether, on the grounds that they are actually positive or negative, but we don't know which: they have been rounded to zero.
3. Find the probability of a result as or more extreme than  $x$  for  $X \stackrel{d}{=} \text{Bi}(n, 0.5)$ : if  $x > \frac{n}{2}$  find  $\Pr(X \geq x)$ ; if  $x < \frac{n}{2}$  find  $\Pr(X \leq x)$ . For large  $n$  use a Normal approximation.
4. The 2-sided  $P$ -value is twice the probability found in (3).

The sign test is most appropriate when it is not possible to order the differences, for example when comparing two drugs using data that consists only of the preferences of a group of subjects. But it can also be useful when we have numerical data (as in the example) as a quick and easily applied test. Informally, we are looking at Figure 71 and asking: “Are the numbers of positive and negative differences about the same?”

#### ▷ EXAMPLE. Fuel economy (continued)

$H_0$ : There is *no* difference in the distribution of fuel economy obtained in the pre-tuning and post-tuning settings.

The sign test will be most effective in detecting a particular type of alternative hypothesis:

$H_1$ : There is a shift in *location* in the distribution of fuel economy obtained from the pre-tuning setting to the post-tuning one.

This form of the alternative hypothesis has the idea that the shape of the distribution is the same in the two settings, but the whole distribution is shifted along by an amount.

The null hypothesis and this form of the alternative hypotheses have the implications:

$$H_0 \Rightarrow \text{population median of differences} = 0$$

$$H_1 \Rightarrow \text{population median of differences} \neq 0 \text{ (i.e. a two-sided alternative).}$$

For that matter, the null hypothesis and this form of the alternative hypotheses also imply statements about the mean:

$$H_0 \Rightarrow \text{population mean of differences} = 0$$

$$H_1 \Rightarrow \text{population mean of differences} \neq 0 \text{ (i.e. a two-sided alternative).}$$

There were 22 cars used, and none of the differences was zero, so  $n = 22$ .

$x$  = the observed number of positive differences = 16.

$\Pr(X \geq 16) = 0.026$ , where  $X \stackrel{d}{=} \text{Bi}(22, 0.5)$ , hence  $P = 0.052$ .

In MINITAB: Stat > Nonparametrics ► 1-Sample sign ...

### 9.3.2 Confidence interval

The confidence interval for the median uses the idea in Chapter 5, Section 6.10: the confidence interval consists of all population median differences that would give a  $P$ -value  $> 0.05$ , if tested as the null hypothesis value.

In the example, we consider the differences as “pre” minus “post”, since mostly the fuel economy pre tuning was higher (worse) than post tuning, and it is generally easier to deal with positive differences.

In this case, a 94.75% confidence interval for the population median differences is (0.07, 0.44) litres per 100 km, and a 98.31% confidence interval is (-0.06, 0.47), so the data are consistent with a true median difference that could be very small, and could be up to about half a litre per 100 km.

Because the confidence interval is based on the discrete, Binomial distribution, a confidence interval with a 95% confidence interval cannot be obtained exactly. MINITAB provides a 95% confidence interval based on non-linear interpolation, which is an approximation.

In MINITAB: Stat > Nonparametrics ► 1-Sample sign ...

This confidence interval is the one produced in the graphical summary of a variable in MINITAB: Stat > Basic Statistics ► Graphical Summary ... It does not require symmetry of the distribution of the differences for its validity; the only assumption is independence of the differences.

The sign test is very simple. It ignores the size of the differences altogether. It is reasonable to wonder: can't we use the size of the differences somehow, but without making any strong assumption about the shape of the distribution from which we are sampling? This leads to the next test we consider.

## 9.4 Wilcoxon matched-pairs signed-rank test

This approach uses the relative magnitudes (ranks) of the positive (or negative) differences among the non-zero differences; zero differences are ignored. The only assumptions required for the test are that the differences are independent observations from the same distribution. It can be used to carry out a hypothesis test and also to obtain a confidence interval.

### 9.4.1 Hypothesis test

The null hypothesis  $H_0$ , and the particular alternative hypothesis  $H_1$  are as for the sign test.

The test is based on the reasoning that, when  $H_0$  is true, the positive and negative differences should have similar magnitude and hence similar ranks.

The logic of this test is as follows.

1. Determine the signed differences ( $d_i$ ) for each pair.
2. Rank the magnitudes of the non-zero differences (i.e. the  $|d_i|$ 's) from smallest to largest (i.e. the smallest non-zero  $|d_i|$  gets a rank of 1, the second smallest gets a rank of 2 and so on). For tied  $|d_i|$ 's, assign the average of the tied ranks to each of the tied differences.
3. Determine  $W$ , the sum of the ranks associated with the positive differences.  $W$  is the test statistic. The ranks are just the numbers  $1, 2, \dots, n$ , so the total of the ranks is  $\frac{1}{2}n(n+1)$ . If the true distribution of differences has a median of zero, which is the case when the null hypothesis is true, we would expect that the sum of the ranks associated with positive differences would be, on average, half of the sum of all the ranks, or  $\frac{1}{4}n(n+1)$ . It is possible to work out the distribution of  $W$ , assuming that  $H_0$  is true. This is a matter of the possible combinations: in general, this is a complicated matter, although the reason it is feasible at all is that it only depends on  $n$ , the number of ranks in the data (here, equal to the number of differences).
4. However, there is a Normal approximation based on the Central Limit Theorem:

$$W \stackrel{d}{\approx} N\left(\frac{1}{4}n(n+1), \frac{1}{24}n(n+1)(2n+1)\right).$$

MINITAB uses this approximation for all values of  $n$ .

Note that the Normal approximation referred to here is not of the data themselves, but for the distribution of  $W$ , the test statistic.

The Wilcoxon test is most appropriate when it is not reasonable to assume a particular form of distribution for the differences (e.g. Normal), but where the differences are quantitative so that they can be ranked. Note that if the observations are simply ranks in the first place, a unique order cannot be determined for their differences, so the Wilcoxon test is not applicable.

#### ▷ EXAMPLE. Fuel economy (continued)

$H_0$ : There is no difference in fuel economy between the pre-tuning and post-tuning settings.

$H_1$ : There is a shift in location.

Again we take the differences as pre-tuning minus post-tuning, remembering that it doesn't matter which way around we do it, provided each difference is worked out consistently.

Car	1	2	3	4	5	6	7	8	9	10	11
Pre tuning	11.51	10.75	11.00	10.77	11.32	11.00	10.68	9.47	10.90	9.50	9.98
Post tuning	11.04	10.27	11.27	9.42	11.20	10.79	10.46	10.02	10.61	9.56	10.47
Difference	0.47	0.48	-0.27	1.35	0.12	0.21	0.22	-0.55	0.29	-0.06	-0.49
rank  diff	15	16	11	21	4	8	9	18	12	1	17
Car	12	13	14	15	16	17	18	19	20	21	22
Pre tuning	9.72	8.04	10.00	10.47	11.01	9.39	9.45	9.04	10.78	10.47	9.93
Post tuning	9.55	7.71	9.41	10.40	9.32	8.10	9.62	8.94	10.64	10.03	10.19
Difference	0.17	0.33	0.59	0.07	1.69	1.29	-0.17	0.10	0.14	0.44	-0.26
rank  diff	6.5	13	19	2	22	20	6.5	3	5	14	10

In MINITAB: Stat > Nonparametrics ► 1-sample Wilcoxon ...

### Wilcoxon Signed Rank Test: Difference in fuel economy

#### Method

$\eta$ : median of Difference in fuel economy

#### Descriptive Statistics

Sample	N	Median
Difference in fuel economy	22	0.21

#### Test

Null hypothesis  $H_0: \eta = 0$

Alternative hypothesis  $H_1: \eta \neq 0$

Sample	N for Test	Wilcoxon Statistic	P-Value
Difference in fuel economy	22	189.50	0.042

$W$ , the sum of the ranks associated with positive differences = 189.5. Based on the Normal approximation, if the null hypothesis is true, the distribution of  $W$  is approximately Normal with mean 126.5 and variance 948.75 ( $= 30.80^2$ ). That is, given the null hypothesis,  $W \stackrel{d}{=} N(126.5, 30.80^2)$ . (The sum of the ranks associated with negative differences = 63.5.) Using the Normal approximation, we find that  $P = 0.042$  for a 2-sided test.

#### 9.4.2 Confidence interval

It is possible to use this approach to obtain a 95% confidence interval for the population median difference. However, to do this, one additional assumption is required, namely, that the distribution of the differences is approximately symmetric. This assumption is satisfied when a shift in location (only) is involved.

For the reasons that apply to the sign test confidence interval, we can't nec-

essarily find a confidence interval using the Wilcoxon method with a confidence coefficient exactly equal to 0.95, or 95%, and this is also beyond the scope of SRW. We find that a 94.9% CI for the true median difference (pre-tuning minus post-tuning) based on the Wilcoxon test is (0.02 to 0.46) litres per 100 km.

## 9.5 Comparing the inferences

We have used three different approaches to the example. When they are compared, we need to pay attention to what is being assumed, since this varies between the approaches. The mean difference was 0.28 litres per kilometre, comparing pre- to post-tuning.

- Assuming the data are Normally distributed, a 95% confidence interval for  $\mu_D$  is (0.03, 0.52), and the test of  $H_0 : \mu_D = 0$  gave  $P = 0.03$ .
- The sign test gave  $P = 0.05$  for testing whether the true median equals zero, and a 95% confidence interval for the true median of (0.07, 0.44).
- Without assuming any particular form for the underlying distribution, the Wilcoxon test gave  $P = 0.04$  for testing  $H_0 : \mu_D = 0$ ; this is also a test of the null hypothesis that the true median equals zero. Assuming symmetry, a 95% confidence interval for  $\mu_D$ , and also for the true median is (0.02, 0.46).

Which approach should we use, here, and how do we decide that, in general? First, keep in mind that this is a “small sample” issue; for larger sample sizes the Central Limit Theorem supports inferences based on Normality for the estimators. Second, note how similar the inferences are; the confidence intervals differ somewhat, and the  $P$ -values differ only slightly.

There is a trade-off between assumptions (if they are justified) and statistical efficiency, generally. If we can make legitimately make the assumptions, it will be better to do so. In the example, the assumption of Normality for the data is appropriate. The  $P$ -values for the three tests, moving from more to fewer assumptions, are 0.03 ( $t$  test), 0.04 (Wilcoxon) and 0.05 (sign test). This is the kind of pattern you might expect.

## 9.6 Exercises

- 9.1 This continues on from exercise 5.7. The investigator wants to test a new eye-drop that is supposed to prevent allergic ocular itching.
- A colleague on the project notices that 58 left eyes have been assigned to  $A$  and 42 left eyes have been assigned to  $P$ . He questions the random method used to assign treatments to eyes: "Shouldn't it have been 50-50 if you used a random method?" Carry out a test to investigate if the assignment observed in the study sample is consistent with a random method of assignment. [Check the colleagues's figures using Stat > Tables > Tally Individual Variables ... for variable = Left eye drug, to determine how many left-eyes were assigned to  $A$ .]
  - Test the null hypothesis that there is no difference between  $A$  and  $P$  in the reduction of itching, using a sign test.  
[Use Stat > Nonparametrics ▶ 1-Sample Sign ... for variable select difference; check Test Median, click **OK**.]
- 9.2 An investigator wished to determine whether epinephrine (adrenaline) has the effect of elevating plasma cholesterol levels in humans. Twelve adult males were selected and given both a placebo and the drug. Blood samples were taken following injection of the placebo and again after injection of epinephrine. Analysis of the blood samples resulted in the following data, which are stored in `epin.mwx`

subject	Cholesterol Levels (mg/100mL)	
	placebo	epiniphrine
1	178	184
2	240	243
3	210	210
4	184	189
5	190	200
6	181	191
7	156	150
8	220	226
9	210	220
10	165	163
11	188	192
12	214	216

- Are these samples paired or independent? Explain.
- Produce an appropriate visual display (or displays) of the data. What do you conclude from your display(s)?

- (c) State the null and two-sided alternative hypotheses which reflect the research question of interest. Why is a one-sided alternative hypothesis not appropriate here? Explain.
- (d) Apply each of the following tests. In each case determine, at least approximately, the  $P$ -value of the test and state your conclusions.
- (i) the sign test;  
[ Suppose C3 contains the difference in cholesterol levels between Epinephrine and placebo. Use Stat > Nonparametrics ► 1-Sample Sign; select C3; click Test median; use the arrow to select the appropriate Alternative hypothesis; click OK.]
  - (ii) the Wilcoxon matched-pairs signed-ranks test;  
[ Stat > Nonparametrics > 1-Sample Wilcoxon; select C3; click Test median; use the arrow to select the appropriate Alternative hypothesis; click OK.]
  - (iii) the t-test.  
[ Stat > Basic Statistics > Paired t; click in the box First sample, select C2; click in the box Second sample, select C1; click OK.  
Alternatively, you can get the same result using Stat > Basic Statistics > 1-Sample t; use One or more samples, each in a column, select C3; check Perform hypothesis tests, enter the value 0 for the Hypothesized mean; click Options and use the arrow to select the appropriate Alternative hypothesis; click OK twice.]
  - (iv) Compare the results of the three tests that you conducted. Which test do you think is most appropriate and why?
  - (v) Use the t-distribution to find a 95% confidence interval for the mean difference in cholesterol levels between the placebo and epinephrine.  
Which do you consider to be more useful, the t-test from (iv) or the confidence interval?
  - (vi)\* What additional information, if any, would you like to have in order to decide whether or not the study was suitable for the stated purpose?

## 9.7 Answers

- 9.1 (a) Define  $X$  to be the number of left eyes assigned to drug A. We formulate a null hypothesis in terms of the proportion of left eyes assigned to drug A: if the assignment is at random, we expect half the left eyes to receive drug A.

Under the null hypothesis,  $X \stackrel{d}{=} \text{Bi}(100, 0.5)$ .

We have observed 58 left eyes with drug A in our sample, so  
 $P\text{-value} = 2 \times \Pr(X \geq 58) = 2 \times (1 - 0.933) = 0.134$ .

The probability of assigning 58 left eyes to drug A or more extreme is 0.134. This result is not particularly surprising if the eyes were assigned to the treatments at random.

- (b) The relevant output from the sign test appears below:

### Sign Test for Median: Difference Method

$\eta$ : median of Difference

#### Descriptive Statistics

Sample	N	Median
Difference	100	0

#### Test

Null hypothesis  $H_0: \eta = 0$

Alternative hypothesis  $H_1: \eta \neq 0$

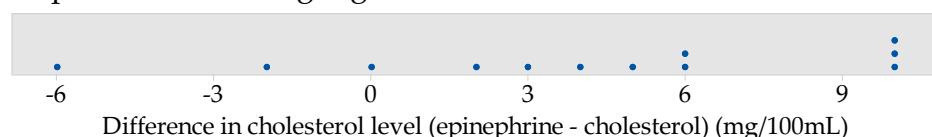
Sample	Number < 0	Number = 0	Number > 0	P-Value
Difference	42	48	10	0.000

There were 42 negative differences, indicating there were 42 cases for which the degree of itching with drug A was rated lower than for placebo. There were 10 cases where the reverse was true.  $P < 0.001$  for the sign test.

- 9.2 (a) Cholesterol levels in the same individuals have been measured under two different conditions: these are matched samples.

- (b) For matched samples we need to consider the distribution of the differences. This must be calculated first.

A useful plot is a dotplot of the differences; 9 of the 12 differences are positive indicating higher levels with adrenaline.



(c)  $H_0$ : There is no difference (in mean plasma cholesterol levels) between the placebo and epinephrine.

$H_1$ : Relative to the placebo, epinephrine produces a different mean plasma cholesterol level.

$H_1$  is two-sided.

A one-sided  $H_1$  would imply that the only possible effect was in one direction; statistical testing should be open to detecting differences in either direction even if we expect higher cholesterol levels with epinephrine than with placebo.

(d) The following table shows the differences,  $d = \text{Epinephrine} - \text{Placebo}$  and the ranks of the absolute values of the differences.

Subject	$d = (E - P)$	Rank of $ d $
1	6	7
2	3	3
3	0	—
4	5	5
5	10	10
6	10	10
7	-6	7
8	6	7
9	10	10
10	-2	1.5
11	4	4
12	2	1.5

(i) The sign test:

### Sign Test for Median: Difference in cholesterol level Method

$\eta$ : median of Difference in cholesterol level

#### Descriptive Statistics

Sample	N	Median
Difference in cholesterol level	12	4.5

#### Test

Null hypothesis  $H_0: \eta = 0$

Alternative hypothesis  $H_1: \eta \neq 0$

Sample	Number < 0	Number = 0	Number > 0	P-Value
Difference in cholesterol level	2	1	9	0.065

There are 9 positive differences;  $P\text{-value} = 0.07$ . The evidence is weak that there are difference in the cholesterol level location between epinephrine and placebo.

(ii) The Wilcoxon Matched Pairs Signed Rank Test:

## Wilcoxon Signed Rank Test: Difference in cholesterol level

### Method

$\eta$ : median of Difference in cholesterol level

#### Descriptive Statistics

Sample	N	Median
Difference in cholesterol level	12	4

Test	N for Wilcoxon		
Sample	Test	Statistic	P-Value
Difference in cholesterol level	11	57.50	0.033

There is some evidence that there are differences in the mean cholesterol levels between epinephrine and placebo:  $P$ -value = 0.03.

- (iii) The  $t$  test:

## Paired T-Test and CI: Placebo, Epin

#### Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Placebo	12	194.67	24.42	7.05
Epin	12	198.67	26.37	7.61

#### Estimation for Paired Difference

Mean	StDev	SE Mean	95% CI for
			$\mu_{\text{difference}}$
-4.00	4.99	1.44	(-7.17, -0.83)

$\mu_{\text{difference}}$ : mean of (Placebo - Epin)

#### Test

Null hypothesis	$H_0: \mu_{\text{difference}} = 0$
Alternative hypothesis	$H_1: \mu_{\text{difference}} \neq 0$
T-Value	P-Value
-2.78	0.018

There is slightly stronger evidence (than for the previous test) for mean differences in cholesterol levels between epinephrine and placebo:  $P$ -value = 0.02. The estimate of the mean difference is 4.0 mg/100mL.

- (iv) If the assumption of normality is met, the  $t$  test is most appropriate as it is most sensitive to detecting differences. Although it is difficult to judge the normality of a small sample, it is not unreasonable in this case.
- (v) Confidence interval for  $\mu_E - \mu_P$ : from the output above, the 95% confidence interval for the mean difference in cholesterol levels in mg/100mL is (0.83, 7.17). The confidence interval is more useful than the  $P$ -value alone. The estimated mean difference was 4.0 mg/100mL; the confidence interval indicates that the sample result is consistent

with true mean differences between 0.83 and 7.17. The true difference could be as small as about 0.8 mg/100mL or as large as about 7 mg/100mL. With expert knowledge about blood cholesterol levels, we can consider whether or not mean differences of this order are of any clinical significance.

A suitable table for summarising the analysis is:

	Mean		Difference in means (Epinephrine – Placebo)		
	Epinephrine	Placebo	Estimate	95% CI	Test statistic
Cholesterol levels (mg/100mL)	198.7	194.7	4.0	0.83, 7.17	$t_{11} = 2.78$

- (vi)\* To what population is it hoped to be able to generalise the results (e.g. people with a particular disease) and can the sample be considered to be a random sample from this population? Was a double blind strategy used? Have possible carry-over effects been allowed for? What other factors (e.g. age, smoking, gender!) might affect the outcome?



# 10 Inference — numerical outcome and one categorical explanatory variable

In this chapter we deal with inferences from some simple forms of the linear model; these are commonly seen in practice, and are the starting point for more elaborate versions of the model. We make “standard” assumptions here. These assumptions vary in their importance. In the next chapter, we will look at how the assumptions can be evaluated and tested, and what strategies are available for dealing with violations of the assumptions.

## 10.1 Independent samples

When we wish to make an inference about the difference between two treatments or interventions, we often do not have a readily available structure for pairing, or matching.

The more likely scenario is that there might be a large population, definable in some way, who could be used for such a study. In such cases, among a pool of potential subjects, allocation can be made at random to each of the two interventions.

We discuss experimental design more in Chapter 16. For the moment, it is important to realize that with a set-up like this, measurements on subjects in the two groups are independent of each other, because they have no structural link with each other.

So in Section 10.2 we consider the comparison of the location of two populations, from which independent random samples have been taken. This means that all the other important factors which may influence the outcome, contribute to the standard deviations within each group.

All other things being equal, it is less efficient to use an independent samples design than a paired design, when a paired design is feasible.

In Section 10.3 we consider the important setting of more than two samples.

## 10.2 *t*-test for independent samples

### 10.2.1 The case of equal variances

We consider inferences for the difference of two means from independent random samples of size  $n_1$  and  $n_2$ , each coming from Normal distributions with the same variance,  $\sigma^2$ . (We deal with the unequal variances case in Chapter 11.)

We use  $Y_{ij}$  to denote the  $j$ th observation from sample  $i$ .

The linear model is written here in two ways. First:

$$Y_{ij} = \mu_i + E_{ij}, \quad i = 1, 2; j = 1, 2, \dots, n_i.$$

where the random error terms  $E_{ij}$  are assumed to be independent and identically distributed, and  $E_{ij} \stackrel{d}{=} N(0, \sigma^2)$ . For this form of the model, the parameter of interest is  $\mu_1 - \mu_2$ .

It is equivalent to write the equation as

$$Y_{ij} = \mu + \alpha_i + E_{ij}, \quad i = 1, 2; j = 1, 2, \dots, n_i.$$

This second form sometimes fits with the way the results are presented in software. It is the form which naturally extends to models with more explanatory variables. For this form of the model, the parameter of interest is  $\alpha_1 - \alpha_2$ . Here we can think of the data as being centred at  $\mu$ , with an additional contribution to the mean coming from  $\alpha_i$ .

The distribution result we apply is the one we previously used to obtain a 95% confidence interval for  $\mu_1 - \mu_2$  in Chapter 5.

$$\bar{Y}_1 - \bar{Y}_2 \stackrel{d}{=} N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

and hence

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \stackrel{d}{=} t_{n_1+n_2-2},$$

where  $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$  is the pooled estimator of  $\sigma^2$ .

Here we are considering hypothesis testing. The null and alternative hypotheses are  $H_0 : \mu_1 - \mu_2 = 0$ ; and  $H_1 : \mu_1 - \mu_2 \neq 0$

The result above implies that when  $H_0$  is true

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \stackrel{d}{=} t_{n_1+n_2-2}.$$

Hence  $T$  is the test statistic and the observed value of  $t$  is compared to the  $t_{n_1+n_2-2}$  distribution to obtain the  $P$ -value; if the null hypothesis is not true,  $T$  will tend to further away from zero than its distribution under the null hypothesis.

In MINITAB: Stat > Basic Statistics ▶ 2-sample t ..., and check the box labelled Assume equal variances.

▷ **EXAMPLE. iBobbly** (iBobbly.mwx) Consider the two-sample example we looked at in Chapter 5, the iBobbly intervention aimed at reducing suicide risk. For convenience, the summary data are repeated here:

Group	$n$	$\bar{y}$	$s$	$s^2$
Waitlist (W)	30	27.83	8.04	64.63
iBobbly (B)	31	23.55	7.76	60.26
Pooled			7.90	62.41

We can test the null hypothesis that the true means,  $\mu_W$  and  $\mu_B$ , are equal; that is,  $H_0 : \mu_W = \mu_B$ . Of course, this is the same as testing  $H_0 : \mu_W - \mu_B = 0$ . For the 2-sided test, we find that  $t = 2.12$ , and  $P = 0.038$ .

Note, in passing, that the sign of the  $t$  statistic depends on the direction of the subtraction in the numerator of the statistic. If the order of the groups was reversed, we would obtain  $t = -2.12$ , and the same  $P$ -value,  $P = 0.038$ .

It is always important to be clear about the direction of the difference. The essence of the inference is the same, provided we interpret it correctly, according to the actual direction used.

- ▷ **QUESTION:** We found previously that a 95% confidence interval for  $\mu_W - \mu_B$  was  $(0.24, 0.83)$ . Suppose that the above test had not been carried out. From this 95% confidence interval *only*, what could you say about the two-sided  $P$ -value?
- ▷ **QUESTION:** Consider the actual  $P$ -value:  $P = 0.038$ . Without any further calculation, what is one of the limits of a 96.2% confidence interval for  $\mu_W - \mu_B$ ? (Check, using MINITAB; you can change the confidence coefficient in the Options).

Recall Figure 56 (page 146) used in Chapter 5 to illustrate the confidence interval for the difference between the means. It is reproduced here in Figure 72, but with a minor addition: the line at the null hypothesis value, zero, is highlighted shown.

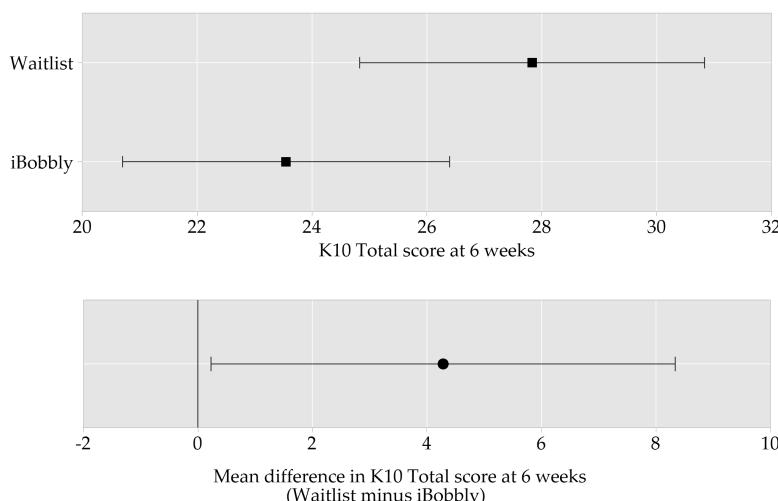


Figure 72: Graph showing the confidence interval (at bottom) for the difference between the two means in the iBobbly data; note the null hypothesis value of zero.

Sometimes the result illustrated in the lower graph in Figure 72 is used to make a simplistic inference: “the confidence interval excludes zero, so the difference is statistically significant”. This is a very limited use of the confidence interval:

- It considers the result of a statistical significance test in terms of an arbitrary threshold, leading to a binary conclusion. This is not good use of the quantitative information available.
- It turns the focus away from the confidence interval, and therefore undermines its utility, which is to provide a plausible range for possible values of the parameter of interest.

The next example demonstrates an important point about the relationship between the separate confidence intervals for the two groups, and the confidence interval for the mean difference.

▷ **EXAMPLE. Bedtime pass** (bedtime.RCT.mwx)

The data are based on an actual study of an intervention designed to help young children settle to bed who have difficulty with this and are “sleep resistant”. The randomised controlled trial investigated the effectiveness of an intervention called a “bedtime pass”, given to the child, which allows the child to get up or be visited once after they have been put to bed. The control treatment was usual care. There were ten children in the treatment group and nine in the control group.

Both treatments were used for some weeks, before the evaluation of the trial commenced.

The data are shown in Figure 73. They are in: bedtime.RCT.mwx.

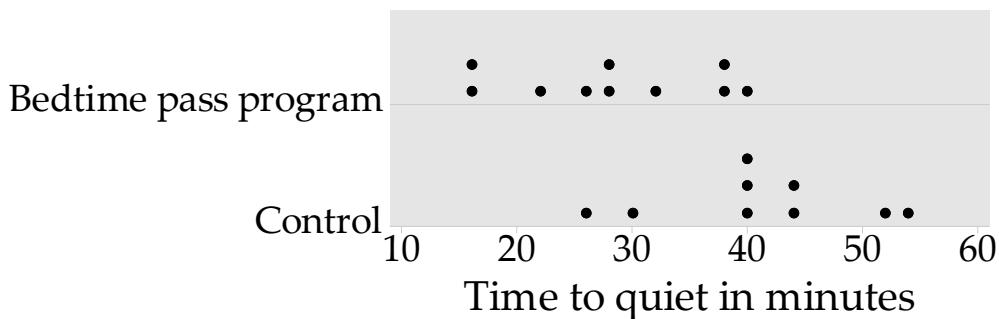


Figure 73: Dotplots of the results of the bedtime pass randomised controlled trial.

The outcome was “time to quiet”: the length of time, in minutes, between when the child was put to bed, and the beginning of the period when the child was quiet for the remainder of the night.

The average time in the control group was 40 minutes and in the treatment group, 28 minutes, a difference of 12 minutes.

Confidence intervals for the two groups separately, and for the difference between the two means, are shown in Figure 74.

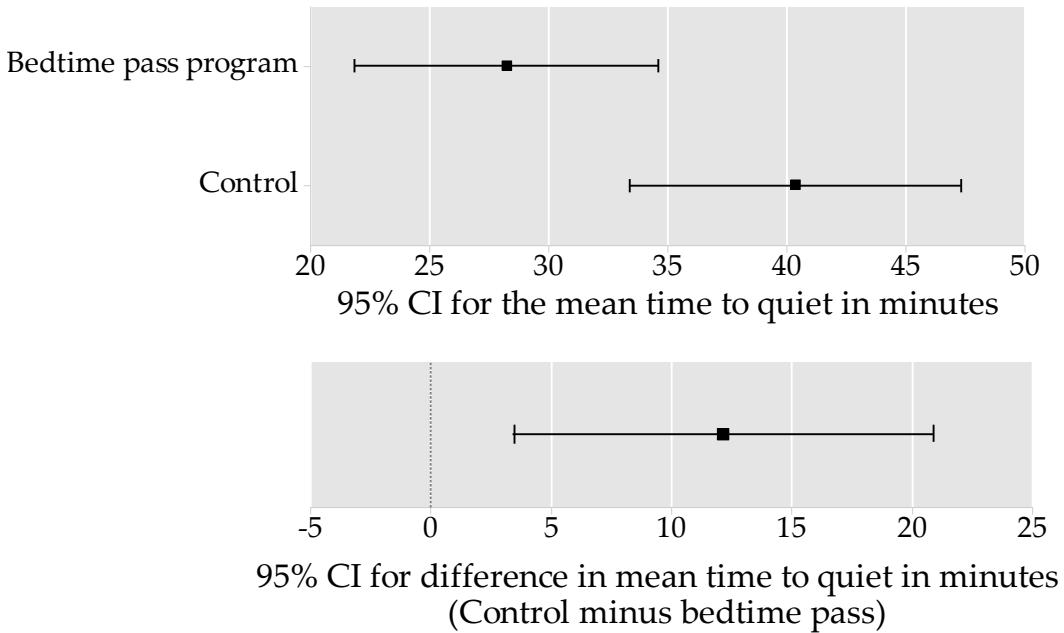


Figure 74: 95% confidence intervals for the bedtime pass randomised controlled trial; separate CIs at the top for the two groups; CI for the difference between the two means at bottom.

These intervals illustrate an interesting phenomenon, which is often found to be counter-intuitive. The separate confidence intervals, for  $\mu_C$  and  $\mu_T$ , say ( $C = \text{control}$  and  $T = \text{treatment}$ ), overlap a little. On the other hand, the 95% confidence interval for  $\mu_C - \mu_T$  is (3.4 to 20.9) minutes, and does not include zero. Given the correspondence between confidence intervals and  $P$ -values, we know from this that the  $P$ -value for the null hypothesis  $H_0 : \mu_C - \mu_T = 0$  must be less than 0.05.

In fact, the  $P$ -value for the null hypothesis is found to be  $P = 0.009$ .

So if the two individual 95% confidence intervals for the group means overlap, we cannot necessarily conclude that the test that the true means are equal will have a  $P$ -value of greater than 0.05. The latter does not inevitably follow from the former, as this example shows. The overlap can be up to about one-third even when there is a significant difference.

On the other hand, if the separate CIs do *not* overlap, then it is safe to conclude that the CI for the difference between the means will not include zero, and therefore that the  $P$ -value must be less than 0.05. Generally, in such circumstances, the  $P$ -value will be quite a bit less than 0.05.

## 10.3 Inference for $k$ means, independent samples

### 10.3.1 Concept of the approach

We now consider the comparison of the means of more than two populations, based on random samples from each of them. This is a qualitatively different matter from the comparison of just two populations. Inferences about the difference between the means of two populations can be represented by a single parameter (e.g.  $\mu_1 - \mu_2$ ). As soon as there at least three populations, this is no longer the case; for  $k$  populations, there are  $\frac{1}{2}k(k - 1)$  pairwise comparisons.

Why are we concerned about location rather than spread, means rather than standard deviations? We can indeed ask inferential questions about the spread of one, two or several populations. Sometimes such questions are of direct interest in themselves.

However, many important research questions are concerned with the mean (or, more generally, the location) of the population distribution, because we want to know whether there is more, or less, of some quantity, in different populations. Does the educational method give a higher score? Which policy intervention leads to lower levels of corruption? Which state has the highest level of charitable giving, per capita? These representative questions, when nailed down to a statistical structure, will involve inferences about the location of the relevant population distributions, and the most commonly used measure of location is the mean.

We now consider the comparison of means of  $k$  ( $> 2$ ) populations from independent samples.

One way to do this could be by comparing pairs of populations using  $t$ -tests. Something like this — as we shall see — is often needed in any case.

But it is sensible to make an overall inference, and hence to formulate tests of the following null hypothesis, versus the alternative hypothesis listed.

$H_0$ : There are no differences between the means of the populations;

$H_1$ : There are some differences in means between the populations.

It is not possible to have a one-sided alternative, because this situation cannot be represented using a single parameter.

▷ **EXAMPLE. White pine** (pine.mwx)

The following data were obtained from an experiment to determine the effect of storage conditions on the moisture content of white pine timber. Eleven samples of timber were used in three different conditions. We assume that the allocation of the samples to the conditions was random; the reason why there are different numbers in the groups is not known.

Conditions	Moisture Content (%)					Mean
1	7.3	8.3	7.6	8.6	8.3	8.0
2	5.4	7.4	7.1			6.6
3	8.5	9.5	10.0			9.3

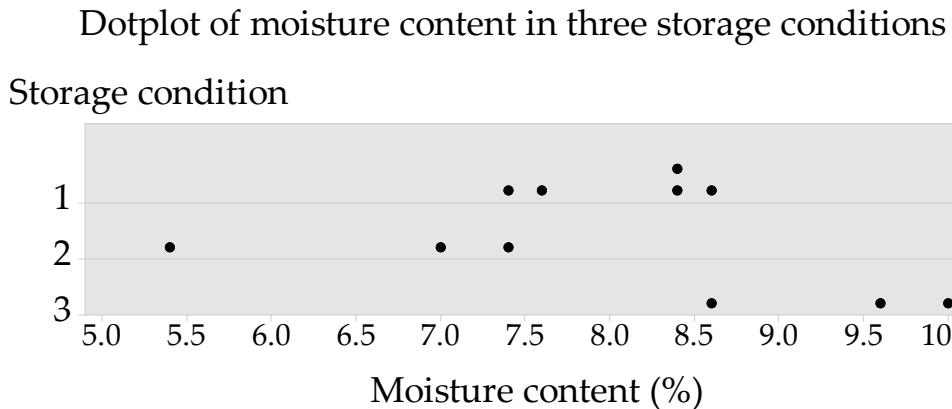


Figure 75: Dotplot of moisture content of white pine in three storage conditions

As you look at Figure 75, ask yourself: are the locations of the populations “behind” these samples the same, or not? What influences your thinking about this? What are the features of the data that shape how you address the question?

For the approach known as ‘one-way analysis of variance’, we assume the following.

- The data come from  $k$  distributions with (possibly) different means  $\mu_i, i = 1, 2, \dots, k$ .
- The variances of each of the  $k$  distributions are assumed to be the same,  $\sigma^2$ .
- The distributions are assumed to be Normal.
- All observations are assumed to be independent.

In other words, we assume that the samples consist of independent observations from  $k$  Normal populations with possibly different means,  $\mu_1, \dots, \mu_k$ , but with the same variance,  $\sigma^2$  ( $\circlearrowright$ ).

One-way analysis of variance is a particular form of the linear model. While there is a jump in complexity from the two-sample case, the form of the linear model is exactly the same. For data  $Y_{ij}, i = 1, 2, \dots, k; j = 1, 2, \dots, n_i$ , the following two forms are equivalent.

$$\begin{aligned} Y_{ij} &= \mu_i + E_{ij}, \quad i = 1, 2, \dots, k; j = 1, 2, \dots, n_i \\ Y_{ij} &= \mu + \alpha_i + E_{ij}, \quad i = 1, 2, \dots, k; j = 1, 2, \dots, n_i \end{aligned}$$

The random error terms  $E_{ij}$  are assumed to be independent and identically distributed, and  $E_{ij} \stackrel{d}{=} N(0, \sigma^2)$ .

We test

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 = \cdots = \mu_k, \text{ against the alternative hypothesis} \\ H_1 : H_0 &\text{ is not true.} \end{aligned}$$

A recurrent idea or analogy that we appeal to here is the “signal to noise” ratio. The test we construct can be viewed in this way. In fact, many test statistics can. In one-way ANOVA, the “signal” about the null hypothesis is the extent to which the sample means are spread out. If they are spread out a lot, there is evidence against the null hypothesis. The “noise” is the background variation that is there no matter what, regardless of whether the null hypothesis is true.

This situation is conceptually very similar to the two-sample  $t$ -test; the only difference is that we cannot now express the size of the variation between the samples as the difference between two means, because there are more than two means. The measure that naturally extends the difference between two means to the variation among several means is the *variance*, which is one reason for the name: **analysis of variance (ANOVA)**.

Analysis of variance is a very general idea within the broad class of linear models. A one-way ANOVA involves one of the simplest forms of linear models, introduced in Chapter 8. Other techniques, known by names such as regression, analysis of covariance, two-way ANOVA, multi-way ANOVA, are all examples of the same general form. For this reason, we consider one-way ANOVA in some detail, because a lot of the features are fundamental to more complicated cases.

The first purpose of the one-way ANOVA is to test  $H_0$ ; along the way, we get a number of other useful results. To test any null hypothesis, we need a test statistic: something we can calculate from the data, whose distribution we know if the null hypothesis is true, and which tends to give a noticeable signal if the null hypothesis is not true. Armed with the test statistic, we can work out the  $P$ -value, which, as always, is the chance of a result at least as extreme as that observed, given that the null hypothesis is true.

The theory below is heading towards that goal: getting the test statistic to test  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ .

### 10.3.2 The $\chi^2$ and $F$ distributions

There are two distributions that we have not previously met, which are used in analysis of variance. They are the  $\chi^2$  distribution and the  $F$  distribution, which are now defined.

## $\chi^2$ distribution

If  $Z_1, Z_2, \dots, Z_p$  are independent  $N(0, 1)$  random variables, then  $U = \sum Z_i^2$  has a  $\chi^2$  (“chie-squared”) distribution with  $p$  degrees of freedom; we write  $U \stackrel{d}{=} \chi_p^2$ . For example,  $Z_1^2 \stackrel{d}{=} \chi_1^2$ . The  $\chi^2$  distribution takes non-negative values and is positively skewed. The  $\chi_3^2$ ,  $\chi_5^2$  and  $\chi_{10}^2$  distributions are shown in Figure 76.

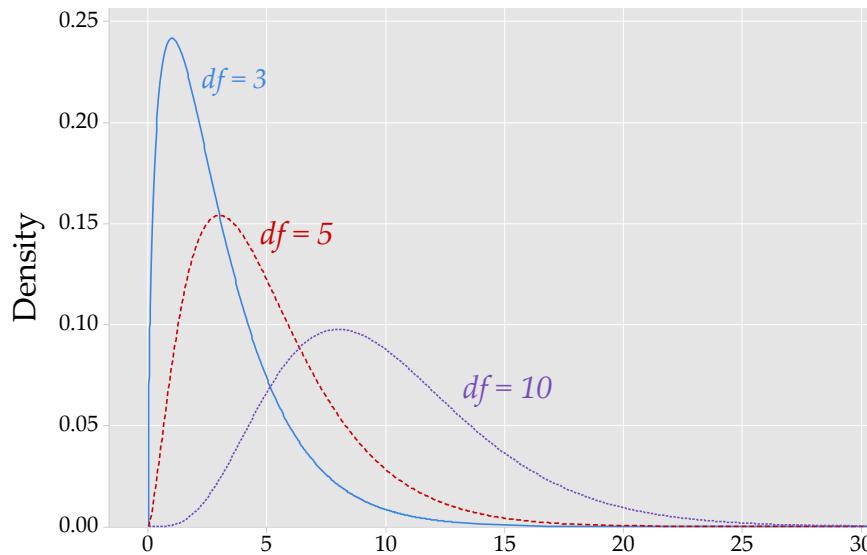


Figure 76: Probability density functions of  $\chi^2$  random variables with various degrees of freedom (df):  $\chi_3^2$ ,  $\chi_5^2$  and  $\chi_{10}^2$ .

## $F$ distribution

If  $U \stackrel{d}{=} \chi_p^2$  and  $V \stackrel{d}{=} \chi_q^2$ , and  $U$  and  $V$  are independent, then  $\frac{U/p}{V/q}$  has an  $F$  distribution with  $p$  and  $q$  degrees of freedom.

Three  $F$  distributions are shown in Figure 77.

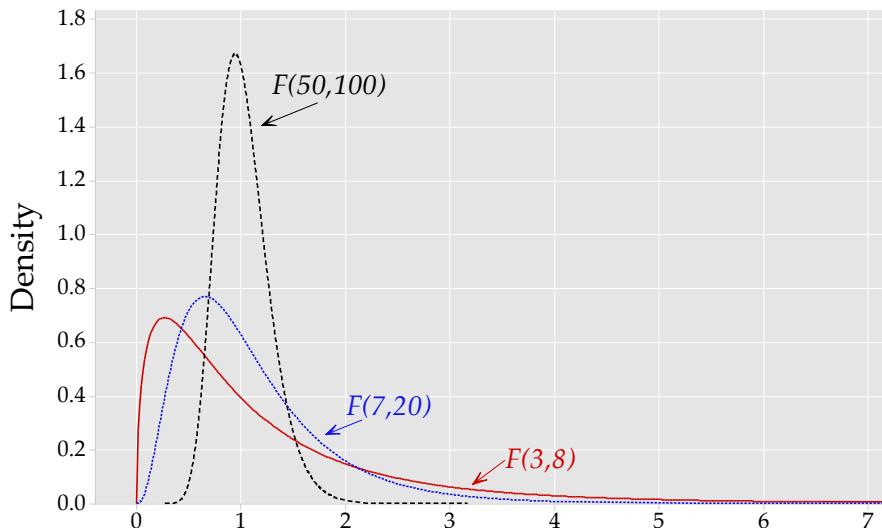


Figure 77: Probability density functions for  $F_{3,8}$ ,  $F_{7,20}$  and  $F_{50,100}$ .

### 10.3.3 Sums of squares and mean squares

A crucial technical quantity that appears throughout this kind of theory is a **sum of squares**. Actually, it is almost always a sum of squared deviations. We have already met such a quantity in the numerator of the formula for a sample variance:  $s^2 = \frac{\sum(y_i - \bar{y})^2}{n - 1}$ .

In order to look at the formulae in detail we need some notation.

term	description
$y_{ij}$	the $j$ th observation from the $i$ th sample, where $i = 1, \dots, k$ and $j = 1, \dots, n_i$
$\bar{y}_i$	the mean of the $i$ th sample
$\bar{y}_{..}$	the mean of all the data, considered as an overall sample

These quantities are written as observations in the table above, and they have their corresponding random variables  $Y_{ij}$ ,  $\bar{Y}_i$  and  $\bar{Y}_{..}$ .

We can think of the total sum of squares as a measure of the total amount of variation in the data. In the notation used, this is

$$\text{total } SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2,$$

It turns out to be possible to split up this total  $SS$  into two parts, one attributable to the variation *between* the samples, and one attributable to the variation *within* the samples.

Of course, even if  $H_0$  is true, we don't expect all the sample means to be

identical; there will be some sampling variation. But the test is based on the idea that for a given total amount of variation, when the variation between the samples is very large relative to the variation within the samples, we have evidence against  $H_0$ .

The split up of the total sum of squares is as follows:

$$\begin{aligned}\text{total } SS &= \text{between groups } SS + \text{within groups } SS \\ \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 &= \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \\ \text{total } SS &= \text{model } SS + \text{residual } SS\end{aligned}$$

1. This equation partitions the total variation in the data into two bits: the variation *between* groups and the variation *within* groups. The two quantities on the right-hand side can be shown to be statistically independent. In terms of the signal to noise ratio analogy, the between groups SS has to do with the signal, because it is reflecting how spread out the sample means are. The within groups SS is about the noise, because it reflects the variation within the groups, around the group means. This is unaffected by how spread out the group means are.
2. The formula

$$\text{total } SS = \text{model } SS + \text{residual } SS$$

is included because it expresses the extension of the analysis of variance concept to more general applications: we split up the total sum of squares of the data into two parts: some variation due to things we can identify (the model  $SS$ ) and other, left-over, residual variation that we cannot attribute to any systematic source, and which we therefore think of as random, residual variation (the residual  $SS$ ).

This is a really important idea. For this reason, we use the terms “within groups  $SS$ ” interchangeably with “residual  $SS$ ”. “Within groups” is a narrow term that applies to one-way ANOVA; it is a special case of the general quantity, “residual  $SS$ ”. And the same goes for the corresponding mean square term.

3. Recall that the *degrees of freedom* of a set of deviations from a mean is one fewer than the number of deviations.

The three  $SS$  in the equation have degrees of freedom associated with them:

- There are  $N$  observations altogether, so the total  $SS$  has  $N - 1$  degrees of freedom.
- The between groups  $SS$  is all about the variation between the  $k$  sample means, so it has  $k - 1$  degrees of freedom.

- The degrees of freedom have to “add up”, so the within groups  $SS$  has  $N - k$  degrees of freedom. Alternatively, one can see that the within groups  $SS$  is a sum of  $k$  within group  $SS$ , which will have (individually) degrees of freedom  $n_i - 1$ . The sum of these is  $N - k$ .
4. In general, the “mean square” is the sum of squares divided by its degrees of freedom. So between groups  $MS = \frac{\text{between groups } SS}{k - 1}$ , and within groups  $MS = \frac{\text{within groups } SS}{N - k}$ .

#### 10.3.4 Developing the formal test of $H_0$

Where is all this going? We need to regroup and see the big picture. We are going to construct a test statistic that captures the sensible conceptual idea: it will measure the extent of the variation between groups, relative to the variation within groups. Extreme values of the test statistic will constitute evidence against  $H_0$ .

1. Regardless of whether the null hypothesis is true, the residual mean square is an estimate of the unknown  $\sigma^2$ , and

$$\frac{\text{residual } SS}{\sigma^2} \stackrel{d}{=} \chi_{N-k}^2, \text{ where } N = n_1 + n_2 + \dots + n_k.$$

2. If the null hypothesis is true then each sample mean estimates the same, single population mean, which has the consequence that the between groups  $MS$  is also an estimate of  $\sigma^2$  and

$$\frac{\text{between groups } SS}{\sigma^2} \stackrel{d}{=} \chi_{k-1}^2.$$

If  $H_0$  is *not* true then the between groups  $MS$  tends to be bigger.

3. These results lead to the  $F$ -test; if  $H_0$  is true, then the  $F$ -statistic

$$F = \frac{\text{between groups } MS}{\text{within groups } MS} \stackrel{d}{=} F_{k-1, N-k},$$

What happens if the null hypothesis is *not* true? In that case, the  $F$ -statistic will tend to be larger. Larger than what? Larger than the pattern we expect based on the  $F$  distribution.

This is where the signal to noise ratio analogy is most relevant, because the test statistic  $F$  really is a ratio. Further, the denominator, the within group  $MS$  (also known as the residual  $MS$ ), is background variation, which can't be removed, just like background noise. And the numerator, the between groups  $MS$ , carries the ‘signal’ corresponding to whether the null hypothesis is true, or not.

The observed  $F$ -ratio is therefore suitable as a test statistic for the overall null hypothesis: it has a known distribution if the null hypothesis

is true, and it is sensitive to departures from the null hypothesis. This is what we want in a test statistic. Specifically, if the null hypothesis is not true, the  $F$ -ratio tends to be large.

So we have got to the point of being able to test  $H_0$ : if  $f$  is the observed  $F$  ratio, the  $P$ -value is given by

$$P = \Pr(F \geq f, \text{given } H_0 \text{ is true}) = \Pr(W \geq f),$$

where  $W \stackrel{d}{=} F_{k-1, N-k}$ . We use software to “look up” the relevant  $F$  distribution and calculate the  $P$ -value.

This process is another example of the  $P$ -value reasoning. In a given case, we obtain the observed  $F$  ratio,  $f$ . Then we ask: does this observed value look like it could be an observation from the  $F$  distribution with degrees of freedom  $k - 1$  and  $N - k$ ? If the answer is ‘yes’, the  $P$ -value will be large, and the result will therefore not be a surprising one, if the null hypothesis is true.

On the other hand, if the observed  $F$  ratio is very large, the  $P$ -value will be small and the result will be a surprising one, if the null hypothesis is true, hence amounting to evidence against the null hypothesis, and in favour of the alternative hypothesis, namely, that there are differences between the means  $\mu_1, \mu_2, \dots, \mu_k$ .

The results are set out in an ‘analysis of variance’ (ANOVA) table as follows:

ANOVA

Source	df	SS	MS	F	P
Between groups	$k - 1$	BetweenSS	BetweenMS = $\frac{\text{BetweenSS}}{k-1}$	BetweenMS ResidualMS	P-value
Residual (within)	$N - k$	ResidualSS	ResidualMS = $\frac{\text{ResidualSS}}{N-k}$		
Total	$N - 1$	TotalSS			

$df$  = degrees of freedom,  $SS$  = sum of squares,  $MS$  = mean square =  $\frac{SS}{df}$ , and  $F$  = ratio of mean squares

The  $P$ -value is found from the  $F$ -distribution with  $(k-1)$  and  $(N-k)$  degrees of freedom. It is equal to the probability of getting a  $F$ -value equal to or larger than that actually observed.

It is important to see that we can always obtain an estimate of  $\sigma^2$  — regardless of whether or not  $H_0$  is true — from the residual mean square; this is the equivalent of the pooled estimator of  $\sigma^2$  in the two-sample case.

In MINITAB, there are several menu commands that will carry out ANOVAs, and one-way ANOVAs in particular:

- Stat > ANOVA ▶ One-way , if the responses are in one column and a second column gives the indices of the groups; MINITAB calls the group variable a ‘factor’, which is a general term for an explanatory variable that is categorical.

- Stat > ANOVA ▶ and select (Response data are in a separate column for each factor level), if the responses from the groups are in different columns, one column for each group. Note that this is a most unwise way to store data, because it is not able to be extended to accommodate more factors.
- Since one-way ANOVA is just a particular form of the linear model, it is a good idea to get used to using the corresponding command in MINITAB, precisely because it can be applied very broadly. This command is Stat > ANOVA ▶ General Linear Model ▶ Fit General Linear Model. For most purposes, we will use this.

### 10.3.5 A connection with the two sample $t$ test

You may have wondered about the connection between the two-sample  $t$ -test and one-way ANOVA. We introduced one-way ANOVA to deal with inference about several population means.

The assumptions of the ANOVA about the data structure are the same as those of the “equal variances” version of the two-sample  $t$ -test. So what would happen if we used a one-way ANOVA in the two-sample situation? The answer — reassuringly — is that the  $P$ -values obtained using either of the approaches are identical. The two sample  $t$ -test has a statistic that gets compared to the  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom. The  $F$  ratio, if one-way ANOVA is used, gets compared to the  $F$  distribution with 1 and  $n_1 + n_2 - 2$  degrees of freedom. These two distributions are related: if  $T \stackrel{d}{=} t_m$  then  $T^2 \stackrel{d}{=} F_{1,m}$ .

### 10.3.6 ANOVA of the white pine data

▷ **EXAMPLE. White pine (continued)**

ANOVA

Source	df	SS	MS	F	P
Conditions	2	10.94	5.47	9.35	0.008
Residual	8	4.68	0.59		
Total	10	15.62			

This result is represented in the following Figure.

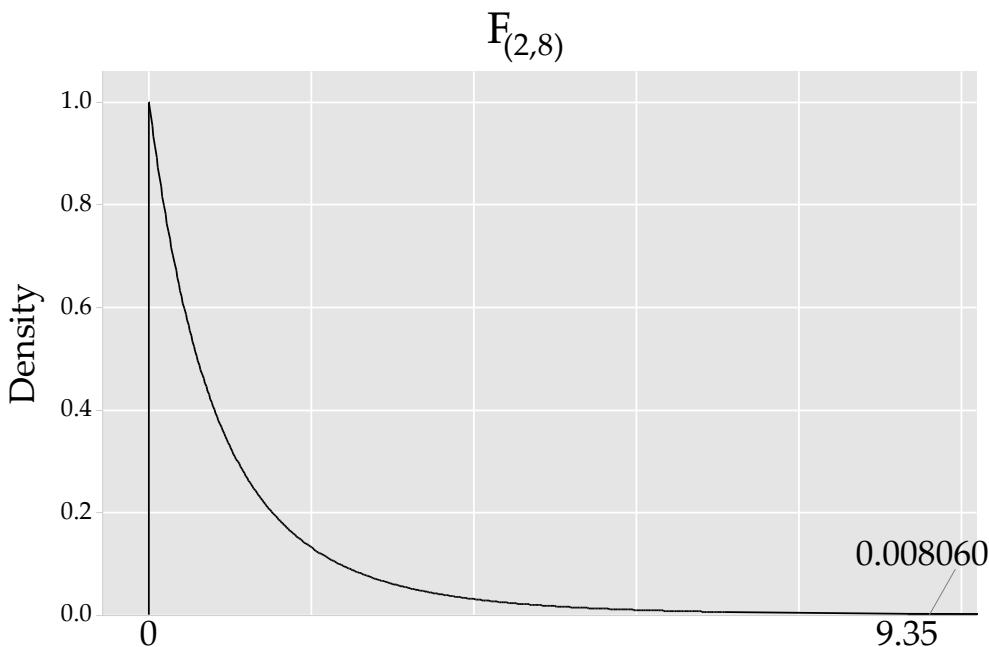


Figure 78: Illustration of the  $P$ -value for the white pine data.

There is strong evidence of a true difference in the mean moisture content of white pine timber kept under the three conditions.

How should the overall test be reported, in a narrative form? It is customary, and advisable, to give more than just the  $P$ -value, although this is what is ultimately interpreted. The purpose of doing so is to give a partial, limited analytic audit trail. If the only thing reported is the  $P$ -value, the reader (or referee!) is obliged to take your analysis entirely on trust. If you report a little more, some of your approach and analysis is open to scrutiny, and this is generally seen as a good thing. So the test result here can be reported like this:

“A one-way ANOVA was carried out to test for differences in the means, between the three conditions. This gave  $F = 9.35$  ( $\text{df} = 2,8$ ) and  $P = 0.008$ .”

### 10.3.7 Exploring changes to the white pine data

Consider what happens if we (arbitrarily) make the means of the second and third groups close to the first, by adding 1.2 to each observation in the second group and subtracting 1.3 from each observation in the third group. Now the dotplot looks like this:

Storage condition

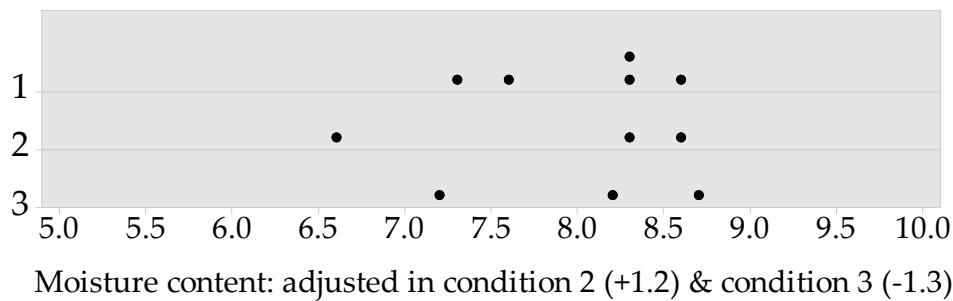


Figure 79: Dotplots of moisture content, with an adjustment.

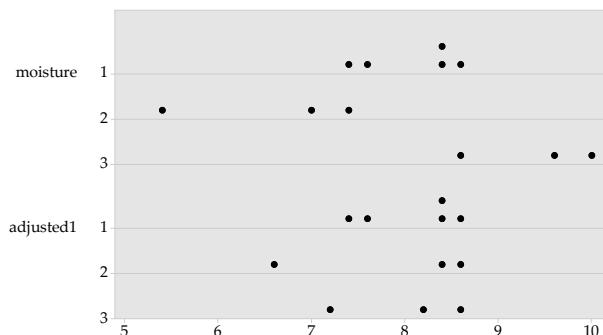
Now the ANOVA for the adjusted data is:

ANOVA

Source	df	SS	MS	F	P
Conditions	2	0.08	0.04	0.07	0.9
Residual	8	4.68	0.59		
Total	10	4.76			

Both the total and between groups  $SS$  are much smaller with the adjusted data but the residual  $SS$  and  $MS$  are the same. For these adjusted data the evidence of a difference is much weaker than in the original data, because the variation between the samples is small relative to the variation within samples.

In Figure 80, the original and adjusted data are lined up to allow the comparisons of interest, and alongside are the means and standard deviations for each group. Think about the consequences for the ANOVA.



Response	Cond.	mean	sd
moisture	1	8.0	0.55
	2	6.6	1.08
	3	9.3	0.76
adjusted1	1	8.0	0.55
	2	7.8	1.08
	3	8.0	0.76

Figure 80: Original and adjusted1 data

Now consider what would happen if we made the data more spread out, but with same group means and overall mean as the original data. Like this:

Conditions	Moisture Content (%)					Mean
1	5.14	9.14	6.34	10.34	9.14	8.0
2	1.70	9.70	8.50			6.6
3	6.00	10.00	12.00			9.3

Now the dotplot looks like this:

Storage condition

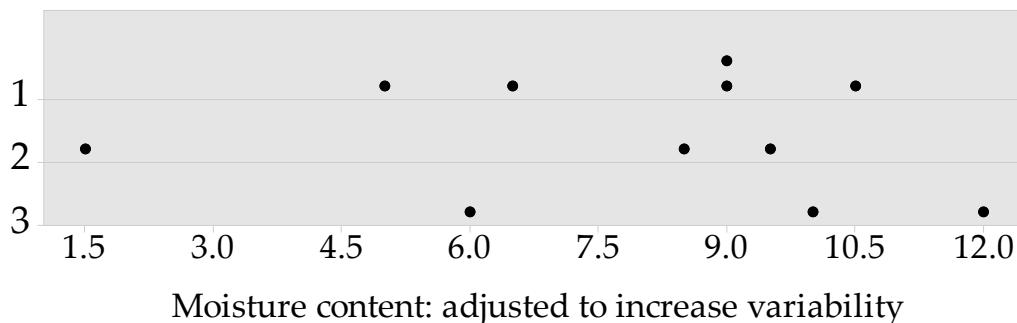


Figure 81: Dotplots of moisture content, with a different adjustment.

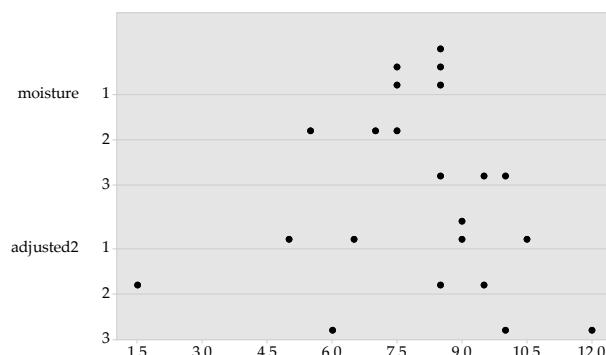
The ANOVA for these adjusted data is:

ANOVA

Source	df	SS	MS	F	P
Conditions	2	10.94	5.47	0.58	0.58
Residual	8	74.90	9.36		
Total	10	85.84			

The between groups  $SS$  is the same as the original data, but the other two  $SS$  are larger. For these adjusted data the evidence of a difference is also much weaker than in the original data, because the variation between the samples is small relative to the variation within samples.

Again, we get good insight into how ANOVA works by lining up the original and adjusted data, and looking at the means and standard deviations.



Response	Cond.	mean	sd
moisture	1	8.0	0.55
	2	6.6	1.08
	3	9.3	0.76
adjusted2	1	8.0	2.18
	2	6.6	4.31
	3	9.3	3.06

Figure 82: Original and adjusted2 data

## 10.4 Confidence intervals and multiple comparisons

### 10.4.1 Confidence intervals

The overall  $F$  test is limited in its usefulness. For one thing, it is a hypothesis test only, and does not focus at all on parameter estimation. Secondly, from it we can only draw an inference about the overall null hypothesis, that all the population means are equal, and not anything more detailed.

It is usually desirable to seek specific inferences about the population means, and very often, inferences of interest are about the differences between pairs of means.

So we wish to find confidence intervals for the difference between particular pairs of means, or to determine  $P$ -values for tests that the means of two groups are equal. This is for the obvious reason that if we conclude that there are some differences between various groups, we would like to know where the differences are. The overall test merely addresses the question of any difference at all.

A 95% confidence interval for  $\mu_1 - \mu_2$  (say) is given by

$$\bar{y}_1 - \bar{y}_2 \pm t_{(N-k)}(0.975) \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)},$$

where  $s^2$  (a *pooled* estimate of  $\sigma^2$ ) is the residual mean square from the ANOVA table. Confidence intervals of this form are sometimes referred to as “Fisher” intervals, or “Fisher’s LSD” method, which refers to “Fisher’s Least Significant Difference”.

Two things to note about this confidence interval:

- It is of the usual general form, estimate  $\pm k \times$  standard error;
- There is a pooled estimate of  $\sigma^2$  being used because we assumed that all the observations have same (population) variance. So each of them can tell us something about  $\sigma$ . In this sense the pairwise inference is exploiting an assumed feature of the data.

▷ **EXAMPLE. White pine** (continued)

A 95% confidence interval for  $\mu_1 - \mu_3$  is

$$8.02 - 9.33 \pm 2.306 \sqrt{0.5852 \left( \frac{1}{5} + \frac{1}{3} \right)} = -1.31 \pm 1.29, \text{ i.e. } (-2.60, -0.03).$$

There are two ways to get all  $\binom{k}{2}$  such confidence intervals in MINITAB.

One one is to use Stat > ANOVA ▶ Oneway ... Comparisons, and check the box labelled Fisher’s, individual error rate, and use the value 5; this stands for a level of significance of 5%, or equivalently, 95% confidence intervals for

the mean differences. You will need to check the box labelled Tests to get the confidence intervals for the mean differences printed out in the session window. You will also get two plots: one showing the mean differences and the associated confidence intervals, and the other showing the estimated means in each group with associated confidence intervals.

The other way is a two step process. First use Stat > ANOVA ► General Linear Model ► Fit General Linear Model. Once the model is fitted, use Stat > ANOVA ► General Linear Model ► Comparisons. Check the box labelled Fisher, click on the name of the factor of interest and then click C = Compare levels for this item. Click on the Results button and check Tests and confidence intervals. This menu applies to other linear models.

#### 10.4.2 Multiple comparisons

When there are  $k$  groups being compared, pairwise inferences between the  $k$  means involve  $\frac{1}{2}k(k - 1)$  comparisons. This introduces an issue known as the “multiple comparisons problem”. Consider the following.

- If we have 100 95% confidence intervals, the probability that they *all* include the relevant parameter values is a lot less than 0.95.
- If we have 100  $H_0$ s, all of which are true, and we test each independently at the 5% level, the chance of at least one significant result is  $1 - (1 - 0.05)^{100} = 0.994$ .

It is often argued that this is a problem: the more tests we do, the greater the likelihood that at least one test will turn out to be significant at the 5% level, even if all of the null hypotheses are true. And there is a similar consequence for simultaneous confidence intervals. Conceptually the way of dealing with this “problem” is to make adjustments as follows:

- Make the individual confidence intervals wider, so that the coverage probability of all the intervals considered as a whole is still 95%;
- Make the individual “error rate” for each of the several tests more stringent, so that overall, the “family error rate” is maintained at 0.05. Or, correspondingly, inflate the  $P$ -values of each of the individual tests.

A number of methods have been developed to do this, the simplest of which is Tukey’s method. It applies to the set of inferences for the pairwise comparisons.

The method consists of using a constant,  $Q$ , which is larger than the corresponding constant from the  $t$  distribution used in the unadjusted inferences. The size of the constant depends on the number of groups,  $k$ , the degrees of freedom of the variance estimate, and the confidence coefficient (e.g. 95%).

Tukey's pairwise confidence intervals calculate

$$\bar{y}_i - \bar{y}_j \pm Q \sqrt{s^2 \left( \frac{1}{n_i} + \frac{1}{n_j} \right)},$$

where  $s^2$  (a *pooled* estimate of  $\sigma^2$ ) is the residual mean square from the ANOVA table. These intervals are always wider than the intervals obtained using the unadjusted, Fisher's LSD, approach.

Tukey's pairwise  $P$ -values are obtained by comparing the usual test statistic with the  $Q$  distribution. These  $P$ -values are always larger than the  $P$ -values using Fisher's LSD approach.

In MINITAB, there are two ways to get Tukey adjustments:

- Use Stat > ANOVA ▶ Oneway ... Comparisons, and check the box labelled Tukey's, family error rate, and enter the value 5, which in this case stands for an overall level of significance of 5%, or, equivalently, *simultaneous* coverage for the confidence intervals, of 95%.
- Use Stat > ANOVA ▶ General Linear Model ... Comparisons ..., choose the relevant categorical variable in "Choose terms for comparisons", and check the Box labelled "Tukey". Under the Results button, check Tests and confidence intervals This gives both 95% confidence intervals and  $P$ -values adjusted for multiple comparisons. Remember to do this after you have fitted the model.

▷ EXAMPLE. White pine (continued)

## Comparisons for moisture

### Tukey Pairwise Comparisons: condition

### Tukey Simultaneous Tests for Differences of Means

Difference of condition Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
2 - 1	-1.387	0.559	(-2.983, 0.209)	-2.48	0.087
3 - 1	1.313	0.559	(-0.283, 2.909)	2.35	0.105
3 - 2	2.700	0.625	(0.916, 4.484)	4.32	0.006

Individual confidence level = 97.87%

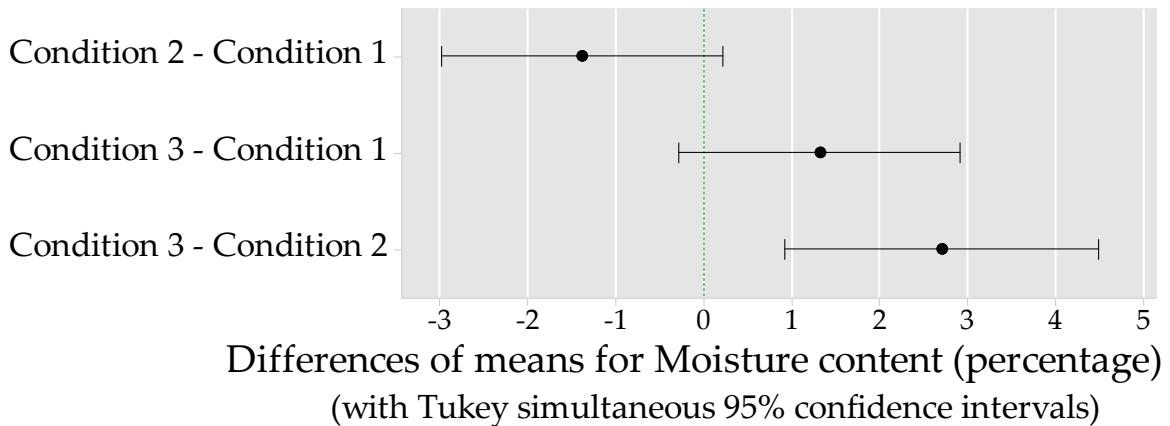


Figure 83: *Pairwise confidence intervals for comparing means in the white pine data, adjusted for multiple comparisons using Tukey's method.*

Figure 83 conveys the pairwise (Tukey-adjusted) 95% confidence intervals and the point estimates in a clean and simple way.

So Tukey's procedure says that  $\mu_1$  and  $\mu_3$  are *not* statistically significantly different at the 5% level; the penalty for multiple comparisons has pushed the  $P$ -value over the threshold of 0.05.

Other multiple comparisons procedures exist; for example, one group may be a control group, and the desired comparisons may be between  $k - 1$  treatment groups and the control group. Then Dunnett's procedure can be used; in MINITAB, you need to specify the control group level, that is, the code for the level of the treatment variable which you are considering to be the control group.

▷ **EXAMPLE. White pine (continued)**

Suppose that the control group happened to be group 3. Then we obtain:

### Comparisons for moisture

#### Dunnett Multiple Comparisons with a Control: condition

#### Dunnett Simultaneous Tests for Level Mean - Control Mean

Difference of condition Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
1 - 3	-1.313	0.559	(-2.799, 0.173)	-2.35	0.080
2 - 3	-2.700	0.625	(-4.361, -1.039)	-4.32	0.005

Individual confidence level = 97.12%

The estimated difference between the means for groups 1 and 3 was  $-1.31$ . For a confidence interval for  $\mu_1 - \mu_3$ , then, we have:

- $(-2.60, -0.03)$ : Fisher's interval, unadjusted for multiple comparisons;
- $(-2.91, 0.28)$ : Tukey's interval, adjusted for all pairwise comparisons;
- $(-2.80, 0.17)$ , appropriate if the third group is a control group and there

is only interest in inference comparing the treatment groups with the control group.

There are a number of other procedures.

There are a few ways of representing these kinds of results.

1. The first, and most helpful representation, is the set of confidence intervals.
2. Sometimes the results are abbreviated to crude ("significant" or "not significant") outcomes, using notation.

This is sometimes represented by superscripted letters, in which two means that share a common letter superscript are deemed to be *not* statistically significantly different.

▷ **EXAMPLE. White pine (continued)**

The white pine data can be represented in this way.

Group	3	1	2
$\bar{y}_i$	9.3 <sup>a</sup>	8.0 <sup>ab</sup>	6.6 <sup>b</sup>

We therefore conclude that, after adjustment for multiple comparisons, at the 5% level, there are no statistically significant differences between the means of populations 3 and 1, nor between populations 1 and 2. These groupings are sometimes referred to as "homogenous subsets". When there are several groups, there can many different homogenous subsets. Or, when the overall  $F$  ratio is small and the overall  $P$ -value is large, there may be just one such set, since then there are no statistically significant differences among the means.

### 10.4.3 The multiple comparisons controversy

Making adjustments for multiple comparisons is common practice. You are likely to meet it in research, including journals. It has been discussed in fields as diverse as agriculture, psychology, epidemiology and genetics. However, it is not universally done and there is a legitimate debate about it.

One situation involving a genuine multiple comparisons problem is when many inferences are carried out, but only the most extreme result is highlighted, or, even worse, only the result with the smallest  $P$ -value is reported. If a survey has 20 numerical scales, suppose that the researcher looks at the 190 correlations among them, picks out the one that is largest, and reports it ( $r = 0.65, P = 0.007$ ) without telling us that she also looked at 189 other correlations. This is a misleading and even fraudulent analytic strategy. On the other hand, if the process was openly reported, then the readers are in a position to judge what the researcher makes of this correlation, in the context of it being one of many examined. Figure 84 illustrates this compellingly.

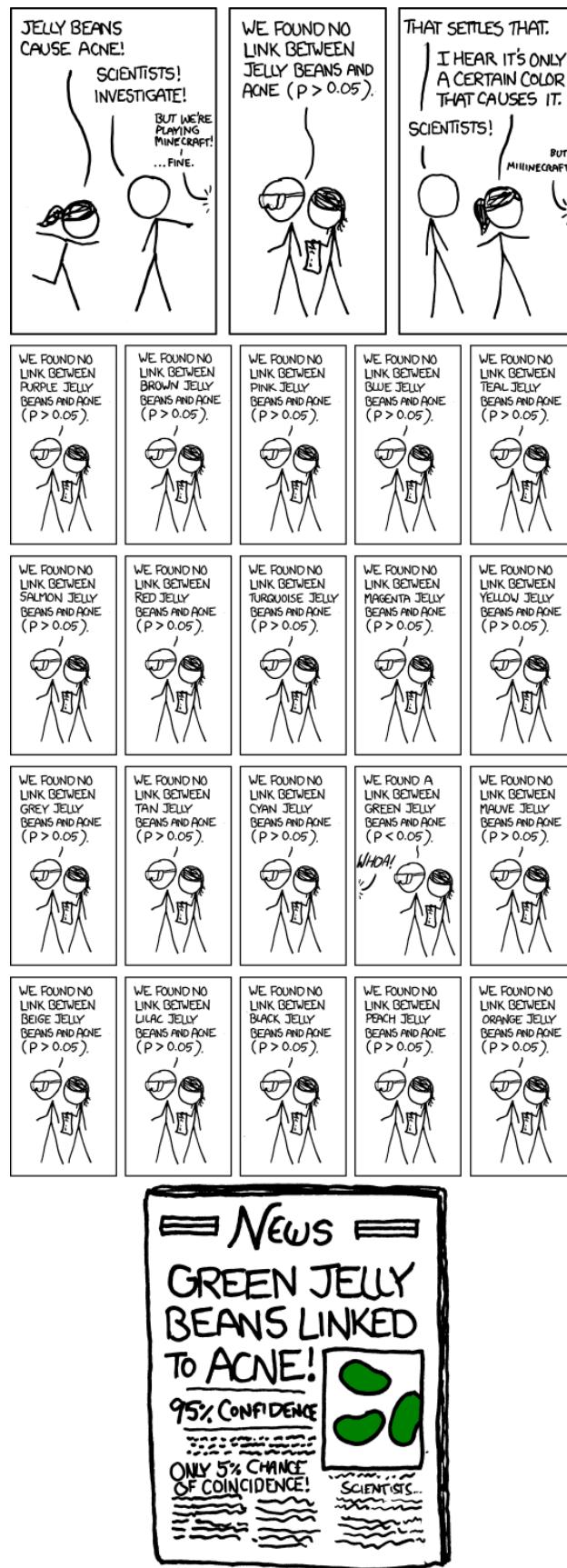


Figure 84: How the multiple comparisons problem can be real  
[<http://xkcd.com/882/>]

However, there is an important argument against adjustment for multiple comparisons, which is both philosophical and practical. It is the concern of *scope*. The argument for adjusting is based on the implications of making a number of inferences simultaneously. But ... why should this be restricted to the number of comparisons in one particular analysis, such as a single one-way analysis of variance? Suppose I write up my results, and Table 4 has an ANOVA with four groups and six pairwise comparisons, and Table 5 has another ANOVA, of a different response variable, also involving six pairwise comparisons. A conventional multiple comparison analytic strategy would make an adjustment for the six comparisons in Table 4, and a separate adjustment for the six comparisons in Table 5. It is reasonable to ask: don't we actually have 12 comparisons altogether? As soon as this question is conceded any merit, we are on a slippery slope to infinity. Why just these two tables? Why not the whole paper? The whole project? Every comparison I have ever, or will ever, make? While this objection starts to sound absurd, it does have compelling logical force and for that reason some statisticians recommend not adjusting for multiple comparisons routinely, while always making clear the number of inferences made.<sup>24</sup>

In practice, whether or not adjustments for multiple comparisons are made depends a lot on the reporting culture of specific research areas. It is also worth noting that within statistical science the application of multiple comparisons adjustments is quite patchy and inconsistent. For example, adjustments are not usually made in fitting regression models with several explanatory variables, even though exactly the same issue arises.

A good approach is to emphasize estimation: to consider the practical interpretation of the differences between the means that you have observed, and the precision with which they have been estimated.

---

<sup>24</sup>See, for example, Rothman KJ (1990). No adjustments are needed for multiple comparisons. *Epidemiology* 1:43 – 46, or Perneger TV (1998). What's wrong with Bonferroni adjustments. *BMJ* 316: 1236 – 1238.

## 10.5 Exercises

- 10.1 In a randomised controlled trial, 24 patients requiring bowel surgery are assigned to either the standard surgical technique or a new surgical method. Once a patient is anaesthetised, there are a number of preparatory procedures that must occur before the actual surgery commences. The time between being anaesthetised and the surgery commencing is referred to as the waiting time. It is important that the new technique does not substantially increase waiting time. For each patient, the waiting time in minutes was recorded. The data are given below, and are in `wait.mwx`.

new method	11.2	14.1	9.0	6.9	12.6	15.4	18.3	14.3	10.6	9.4	12.1	11.3
standard technique	6.1	4.7	8.3	9.0	10.5	9.2	12.5	6.4	11.6	9.7	12.2	8.6

- (a) Produce an appropriate visual display of the data. What do you conclude from your display?
  - (b) Use a t-test to find the  $P$ -value for the test of the null hypothesis of no difference between the average waiting time for the two methods. State your null and alternative hypotheses, and state your conclusions.
  - (c) Find a 95% confidence interval for the difference between the mean waiting time for the two methods.
  - (d) The head surgeon stated that he would be concerned if the new method increased waiting time by more than five minutes on average. What evidence does your analysis provide in addressing his concern?
- 10.2 Use the file `behav.mwx`. It contains cholesterol level measurements in mg/100 mL for 40 obese men. The men were classified as one of two behaviour pattern types. Type A behaviour is characterized as aggressive and urgent; type B behaviour is typically relaxed and non-competitive. The data are from a study that investigated if cholesterol levels in obese men are related to behaviour pattern type.
- (a) Obtain a suitable visual display of the data and sketch it below.
  - (b) Inspect the display and indicate what kinds of assumptions might be reasonable for analysis of these data.
  - (c) What null and alternative hypotheses should be tested?
  - (d) Conduct an appropriate statistical test. Report all assumptions that you have made, the test statistic and the  $P$ -value.
  - (e) Find the 95% confidence interval that should be reported along with the hypothesis test.
  - (f) What kind of study design is used here? Is it a randomised experiment, or an observational study?

- (g) Write a brief conclusion about the relationship of cholesterol levels to behaviour type in obese men, according to these data.
- (h)\* A journalist writes an article summarising the results of this study. The title reads: "Aggressive behaviour raises cholesterol in obese men". Comment on the appropriateness of the title in light of the study design and your analysis.
- 10.3 Pinot noir is one of the oldest grapes grown for making wine. A good bottle of pinot noir can be difficult to produce. The grape vine is susceptible to frost, viruses and birds. The wine can ferment violently and uncontrollably in production, and the flavour depends on the "flavour of the soil". Vincent Lakey, a wine-maker, was interested in using environmentally-friendly treatments of the soil under pinot noir vines. He compared a standard herbicide with two greener alternatives: straw mulch and compost.
- The experiment used these three treatments in each of six different areas of the vineyard, recorded in the variable called "block" in the data set, *pinot.mwx*.
- For the purposes of this problem, we ignore all the possible explanatory factors other than the treatment variable. There were three treatments used: 1: herbicide; 2: compost; 3: straw. We seek to examine the research question of interest:
- Does the treatment influence the size of the yield obtained at harvest?
- (a) The response variable is the weight of bunches harvested in kg (2001). Examine the distribution of the weight by treatment group, using an appropriate graph.
  - (b) What are the population parameters of interest?
  - (c) Explain the null hypothesis suitable for addressing Vincent's question and the alternative hypothesis.
  - (d) What assumptions are required for analysing the data using the model you have suggested? Describe these assumptions in concrete terms in relation to Vincent's study, rather than in abstract form.
  - (e) Fit a general linear model with one categorical explanatory variable (treatment group). What conclusion can be drawn about the test of the null hypothesis?
  - (f) Find 95% confidence intervals for comparing the mean yields for each pair of treatments: compost minus herbicide, straw minus herbicide, and straw minus compost.
  - (g) A claim is made that, for an experimental unit of the size used in Vincent's experiment, "You could get up to an average of 3 kg

more grapes using compost than herbicide." Comment on this claim in the light of the inferences you have obtained, based on the 2001 data, and the analysis you have carried out.

- 10.4 Consider data from smokers in the National Health and Nutrition Examination Survey. The MINITAB worksheet is NHANES.mwx. Consider the variable labelled Serum cholesterol (mg/100ml); this is the serum cholesterol, in mg/100 ml, measured in 1971.

Find the variable called 'Activity level'. This records answers to the question: "In your usual day, how active are you?", asked in 1971. The responses are coded: 0 for very active, 1 for moderately active, and 2 for inactive.

- (a) Produce an appropriate display of the serum cholesterol levels, according to the three different levels of activity. Edit your graph to give informative labels to the levels of activity. What do you conclude from your display?
- (b) Assuming the activity groups have equal variances ( $\circlearrowright$ ), carry out a test of the hypothesis that there are no differences in mean serum cholesterol levels of smokers between the three activity levels.  
[ Stat > ANOVA ▶ General Linear Model ▶ Fit General Linear Model; select cholesterol as the Responses; click in the Factors box; select active ]
- (c) Examine the equality of the variances. Does it appear reasonable to assume that the variances are equal? [ Inspect the 4-in-1 residual plots. Carry out a formal test. ]
- (d) Comment on the Normality of the standardised residuals.  
[ Again consider the 4-in-1 residual plots. ]
- (e) Find a set of simultaneous 95% confidence intervals for the difference between the means of each pair of activity levels.  
[ Stat > ANOVA ▶ General Linear Model ▶ Comparisons; select active as the Terms for comparisons; click to Compare levels for this item; choose the appropriate method for the simultaneous comparisons. ]
- (f)\* Complete the tables below to report the results of the analyses above:

Activity level	Serum cholesterol level (mg/100 ml)		
	Mean	Standard deviation	n
Inactive			
Moderately active			
Very active			
One-way ANOVA:	$F( , ) = , P$		

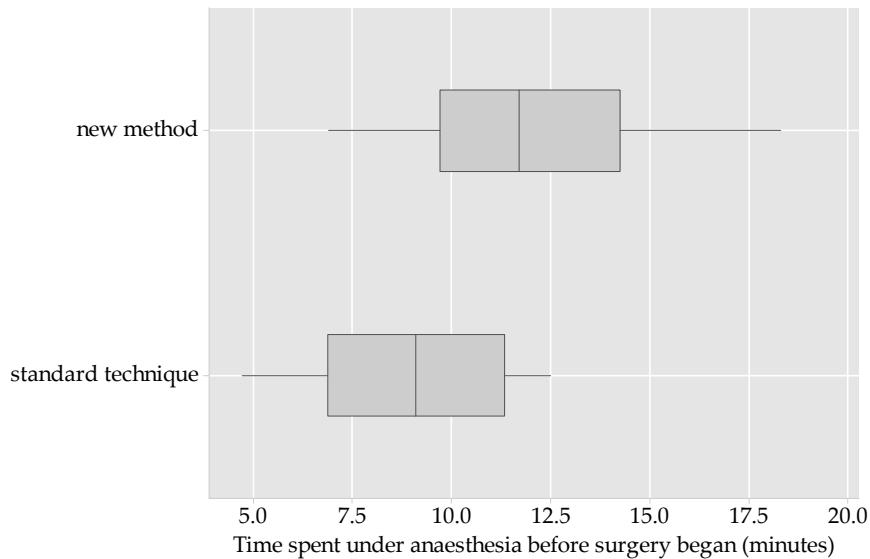
Comparison of activity levels	Difference in mean serum cholesterol (mg/100 ml)		
	Estimate	95% CI*	P-value*
Moderately active – Very active			
Inactive – Moderately active			
Inactive – Very active			

\* Confidence Intervals and P-values obtained using

- (g)\* Consider the interval plot for the differences of means produced when you obtained the comparisons. Make any improvements to the graph, guided by the principles of good graphics from Chapter 3.
- (h)\* Write a plain language summary of the findings of your analysis.

## 10.6 Answers

- 10.1 (a) It would appear that waiting time is generally longer, on average, for the new technique than for the standard technique. This is what was expected. The medians look no more than 3 minutes apart.



- (b) Here a two-sample t-test is appropriate.

### Two-Sample T-Test and CI: Time spent under ... hesia, Surgical method

#### Method

$\mu_1$ : mean of Time spent under anaesthesia when Surgical method = new method

$\mu_2$ : mean of Time spent under anaesthesia when Surgical method = standard technique

Difference:  $\mu_1 - \mu_2$

*Equal variances are assumed for this analysis.*

#### Descriptive Statistics: Time spent under anaesthesia

Surgical method	N	Mean	StDev	SE Mean
new method	12	12.10	3.11	0.90
standard technique	12	9.07	2.45	0.71

#### Estimation for Difference

Difference	Pooled	95% CI for
	StDev	Difference
3.03	2.80	(0.66, 5.40)

#### Test

Null hypothesis  $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2.65	22	0.014

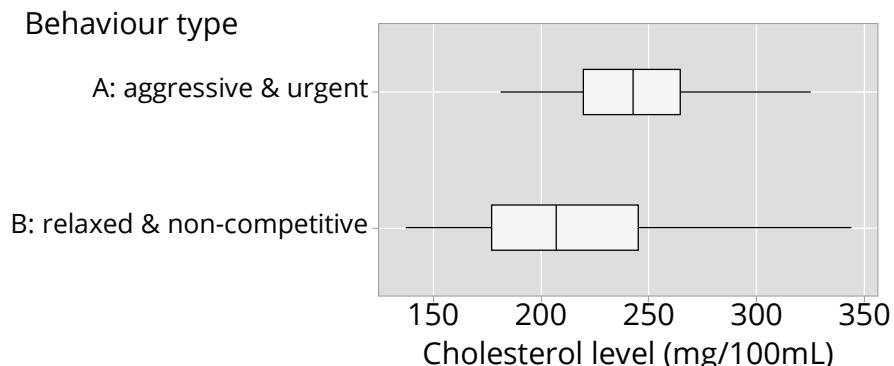
The null hypothesis tested by this procedure is that the average

waiting time for the two types of surgery was the same. The alternative hypothesis was that the average waiting time for two type of surgery would be different.

The estimated mean difference in waiting time is 3.0 minutes; the  $P$ -value is relatively small, suggesting the data are not consistent with the null hypothesis of no difference in waiting time.

- (c) A 95% confidence interval for the difference between the mean waiting time (new technique – standard) is (0.66 min, 5.40 min). The new technique appears to increase mean waiting time; the mean wait could be increased by forty seconds or it could be as large as five and a half minutes.
- (d) Although the estimated increase in average waiting time is 3.0 minutes, the 95% confidence interval indicates that the results we found were consistent with true mean increases in waiting time of five minutes or more with the new technique. On the basis of the data from this study, we cannot guarantee to the head surgeon that the average increase in wait will not be more than five minutes.

- 10.2 (a) Boxplots are appropriate here.



- (b) If we are considering a comparison of means, and wish to use a two-sample  $t$ -test, the assumptions we need to assess using a visual display of the data are: (i) are the samples in each group from an underlying normal distribution, and (ii) are the variances in the underlying populations equal. Looking at the boxplots, both these assumptions appear to be reasonable. You may also like to examine the dotplots to confirm this.
- (c) It is reasonable to use a two-sample  $t$ -test. The null hypothesis tested by this procedure is that there is no difference in the average cholesterol levels of type A and type B males. The alternative hypothesis is that the average cholesterol levels of type A and type B males are different.

- (d) The relevant output is:

**Two-Sample T-Test and CI: Cholesterol level (mg/100mL),**

**... viour type**

**Method**

$\mu_1$ : mean of Cholesterol level (mg/100mL) when Behaviour type = A: aggressive & urgent

$\mu_2$ : mean of Cholesterol level (mg/100mL) when Behaviour type = B: relaxed & non-competitive

Difference:  $\mu_1 - \mu_2$

*Equal variances are assumed for this analysis.*

**Descriptive Statistics: Cholesterol level (mg/100mL)**

Behaviour type	N	Mean	StDev	SE Mean
A: aggressive & urgent	20	245.1	36.6	8.2
B: relaxed & non-competitive	20	210.3	48.3	11

**Estimation for Difference**

Difference	Pooled	95% CI for
	StDev	Difference
34.8	42.9	(7.3, 62.2)

**Test**

Null hypothesis  $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
2.56	38	0.014

The estimated difference in mean cholesterol levels is 34.8 mg/100mL, with obese type A males having higher average cholesterol than obese type B males. A 95% confidence interval for the true difference in means (type A – type B) in mg/100mL is 7.3 to 62.2. This is consistent with the independent *t*-test result ( $t_{38} = 2.56, P = 0.014$ ).

The analysis assumes that the cholesterol levels of two groups of obese men are randomly sampled from populations which are normally distributed with the same variance. The observations are assumed to be independent.

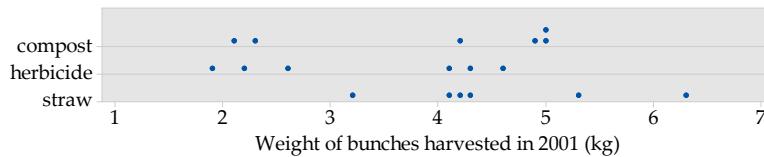
- (e) See the discussion above.
- (f) The study is an observational study. The groups are formed by classifying individuals according to behaviour patterns. Groups are not determined by the researcher.
- (g) In a comparison of cholesterol levels obtained from a sample of 20 obese type A men and 20 obese type B men, obese type A males were found to have higher average cholesterol by 34.8 mg/100mL. The 95% confidence interval for the difference in means (type A – type B) in mg/100mL is 7.3 to 62.2. The mean difference of 35 mg/100mL is likely to reflect a clinically important difference between the two types. However the confidence interval indicates that the precision of this estimate is relatively poor.

A useful summary table for the results is:

	Mean (standard deviation)		Difference in means (type A – type B)			$P$ -value
	Type A	Type B	Estimate	95% CI	Test statistic	
Cholesterol levels (mg/100mL)	245.0 (36.6)	210.3 (48.3)	34.8	7.3, 62.2	$t_{38} = 2.56$	0.014

- (h) The journalist's title implies that behaviour is the cause of raised cholesterol levels. Maybe it's around the other way: high cholesterol causes type A behaviour. Maybe both are caused by something else again. Causality is difficult to establish in observational studies, and cannot be claimed on the basis of statistical differences alone. A more appropriate title would describe differences rather than imply causality.

- 10.3 (a) Given the small number of observations in each treatment group, dotplots are appropriate.



- (b) If the treatment influences the size of the harvest, the mean weight of the bunches harvested should differ according to treatment. The parameters are the true mean weight of the bunches harvested under each treatment.
- (c) A suitable null hypothesis is that the true mean weight of the bunches harvested under each treatment do not differ ( $H_0 : \mu_{\text{straw}} = \mu_{\text{compost}} = \mu_{\text{herbicide}}$ ). An appropriate alternative is that there are differences between the treatments in the true mean weight of the bunches harvested.
- (d) A General Linear Model can be used to test the null hypothesis described above. The observations of the mean weight of bunches harvested must be independent. The three distributions from which the samples in each treatment are taken must have the same variance. The underlying distributions of harvest weights for each of the treatments should be Normal.

PINOT.MWX

**General Linear Model: Weight2001 versus Treatment****Method**

Factor (-1, 0, +1)  
coding

**Factor Information**

Factor	Type	Levels	Values
Treatment	Fixed	3	compost, herbicide, straw

**Analysis of Variance**

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Treatment	2	4.942	18.15%	4.942	2.471	1.66	0.223
Error	15	22.284	81.85%	22.284	1.486		
Total	17	27.226	100.00%				

**Model Summary**

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
1.21885	18.15%	7.24%	32.0888	0.00%	66.00	66.49

**Coefficients**

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	3.920	0.287	(3.308, 4.532)	13.64	0.000	
Treatment						
compost	0.010	0.406	(-0.856, 0.876)	0.02	0.981	1.33
herbicide	-0.647	0.406	(-1.513, 0.219)	-1.59	0.132	1.33
straw	0.637	0.406	(-0.229, 1.503)	1.57	0.138	*

**Regression Equation**

$$\text{Weight2001} = 3.920 + 0.010 \text{Treatment_compost} - 0.647 \text{Treatment_herbicide} + 0.637 \text{Treatment_straw}$$

(e)

The sample means in the three treatments vary from 3.3kg to 4.6kg. The *P*-value is 0.2, suggesting that the observed means are consistent with the true mean weights being the same. However, as the sample sizes are very small, the analysis may only be sensitive to large mean differences. Further, we need to consider the assumptions made in the analysis to check if the *P*-value is reliable.

- (f) The output with Fisher's intervals LSD and Tukey's intervals obtained from the General Linear Model procedure in Minitab is shown below:

**Tukey Pairwise Comparisons: Treatment****Tukey Simultaneous Tests for Differences of Means**

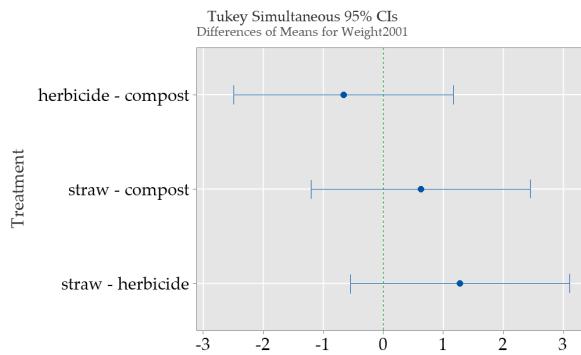
Difference of Treatment Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
herbicide - compost	-0.657	0.704	(-2.483, 1.169)	-0.93	0.628
straw - compost	0.627	0.704	(-1.199, 2.453)	0.89	0.654
straw - herbicide	1.283	0.704	(-0.543, 3.109)	1.82	0.196

Individual confidence level = 97.97%

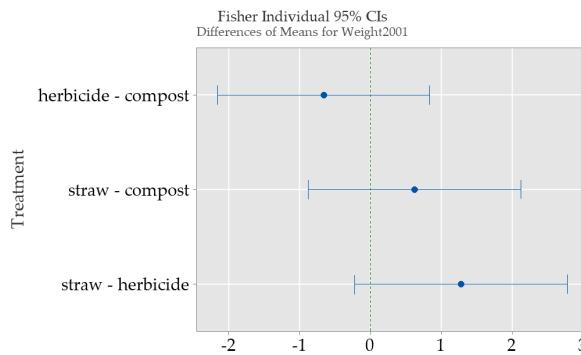
**Fisher Pairwise Comparisons: Treatment****Fisher Individual Tests for Differences of Means**

Difference of Treatment Levels	Difference of Means	SE of Difference	Individual 95% CI	T-Value	P-Value
herbicide - compost	-0.657	0.704	(-2.157, 0.843)	-0.93	0.366
straw - compost	0.627	0.704	(-0.873, 2.127)	0.89	0.387
straw - herbicide	1.283	0.704	(-0.217, 2.783)	1.82	0.088

Simultaneous confidence level = 88.31%



If an interval does not contain zero, the corresponding means are significantly different.



If an interval does not contain zero, the corresponding means are significantly different.

An example of an appropriate summary table is:

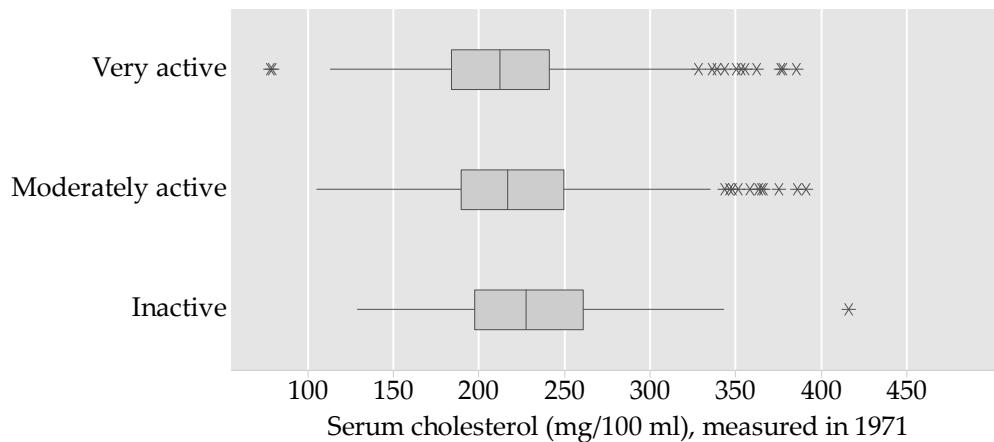
Comparison	Estimate	Difference in means
		95% confidence interval
Compost – Herbicide	0.66	-0.84, 2.16
Straw – Herbicide	1.28	-0.22, 2.78
Straw – Compost	0.63	-0.87, 2.13

These are the Fisher intervals, for example. Note that the estimates and confidence intervals are presented so that the mean differences are positive.

- (g) The 95% confidence interval for the true mean difference herbicide minus compost is  $(-2.48, 1.17)$  if you are using Tukey comparisons. It is  $(-2.16, 0.84)$  if you are using Fisher intervals. Neither of these are consistent with a true mean difference of 3 kg.

- 10.4 (a) The boxplots show large variation in the serum cholesterol levels of smokers and higher median serum cholesterol levels with decreasing activity levels.

The numeric labels should be replaced with text.



- (b) The output showing the test of the null hypothesis of no difference in mean serum cholesterol according to levels of activity is shown below. The  $P$ -value is small indicating that the pattern of observed means is surprising if there is no true difference in the mean cholesterol levels for the different activity groups. The formal report of the test statistic and the  $P$ -value is shown in part (f).

#### General Linear Model: Serum cholesterol (mg/100ml) versus Activity level

##### Method

```

Factor      (-1, 0, +1)
coding
Rows        20
unused

```

##### Factor Information

Factor	Type	Levels	Values
Activity level	Fixed	3	0, 1, 2

##### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Activity level	2	39661	1.11%	39661	19830	9.66	0.000
Error	1723	3536644	98.89%	3536644	2053		
Total	1725	3576304	100.00%				

##### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
45.3057	1.11%	0.99%	3549054	0.76%	18067.17	18088.96

##### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	222.31	1.38	(219.61, 225.01)	161.34	0.000	
Activity level						
0	-7.34	1.67	(-10.61, -4.06)	-4.39	0.000	1.00
1	-0.66	1.66	(-3.92, 2.61)	-0.40	0.692	1.00
2	8.00	2.41	(3.26, 12.73)	3.31	0.001	*

##### Regression Equation

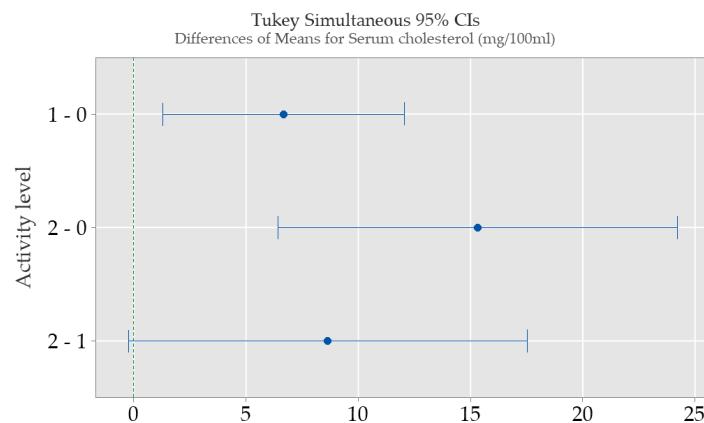
$$\begin{aligned} \text{Serum cholesterol (mg/100ml)} &= 222.31 - 7.34 \text{ Activity level}_0 \\ &\quad - 0.66 \text{ Activity level}_1 \\ &\quad + 8.00 \text{ Activity level}_2 \end{aligned}$$

- (c) Tukey's simultaneous 95% confidence intervals are shown below.

**Tukey Pairwise Comparisons: Activity level**  
**Tukey Simultaneous Tests for Differences of Means**

Difference of Activity level Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
1 - 0	6.68	2.30	(1.29, 12.06)	2.90	0.010
2 - 0	15.33	3.80	(6.43, 24.24)	4.03	0.000
2 - 1	8.65	3.80	(-0.23, 17.54)	2.28	0.059

Individual confidence level = 98.06%



If an interval does not contain zero, the corresponding means are significantly different.

(d)\* The completed tables are as follows:

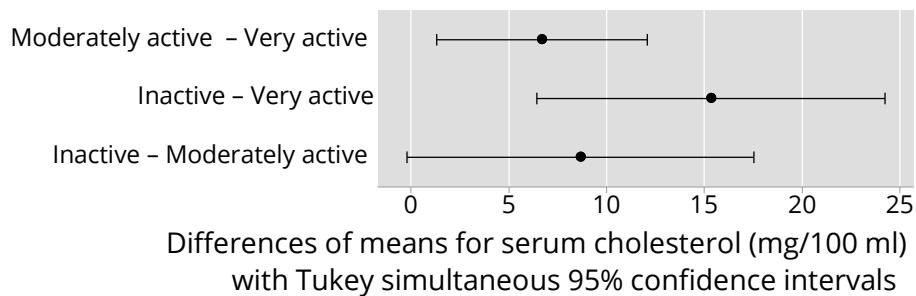
Serum cholesterol level (mg/100 ml)			
Activity level	Mean	Standard deviation	n
Inactive	230.3	45.9	174
Moderately active	221.7	46.5	785
Very active	215.0	43.9	767

One-way ANOVA:  $F(2, 1723) = 9.66, P < 0.001$

Difference in mean serum cholesterol (mg/100 ml)			
Comparison of activity levels	Estimate	95% CI*	P-value*
Moderately active – Very active	6.7	1.3, 12.1	0.010
Inactive – Moderately active	8.7	-0.2, 17.5	0.059
Inactive – Very active	15.3	6.4, 24.4	<0.001

\* Confidence Intervals and P-values obtained using Tukey's method

(e)\* Improvements to the graph, guided by the principles of good graphics from Chapter 3, are illustrated below.



(f)\* The analysis found mean differences in serum cholesterol levels (mg/100 ml) of smokers according to activity levels that were not consistent with an underlying model of no true mean differences ( $F_{(2,1723)} = 9.66, P < 0.001$ ). Mean serum cholesterol levels increased with decreasing levels of activity; for example, the mean serum cholesterol level for inactive smokers was 15.3 mg/100 ml higher than for very active smokers (95% confidence interval: 6.4 to 24.4 mg/100 ml).



# 11 Assumptions — numerical outcome and one categorical explanatory variable

In Chapter 10 we considered inferences for comparing means in a “standard” setting, with many assumptions. Often these assumptions may be fine, and the methods of Chapter 10 apply.

In this chapter we consider two issues, broadly. How do we assess whether the assumptions are justified? What strategies are available to deal with contexts where the assumptions are not valid?

## 11.1 Two sample $t$ test, unequal variances

In Chapter 10 we considered inferences on the difference between two means, assuming random samples from Normal distributions with the same variance,  $\sigma^2$ .

When the assumption of a common variance is not made, there is an alternative, approximate version of the  $t$  test that can be used to obtain a confidence interval for  $\mu_1 - \mu_2$  or to carry out a test of  $H_0 : \mu_1 - \mu_2 = 0$ .

In fact, we have already obtained the result for the confidence interval, in Chapter 5.

For the hypothesis test, we use the result previously discussed in the context of confidence intervals. We evaluate

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

and compare the observed test statistic,  $t$ , with the distribution of  $t_m$ , where  $m$  is given by a complicated formula;  $m$  will always be between the smaller of  $n_1 - 1$  and  $n_2 - 1$ , and  $n_1 + n_2 - 2$ .

In MINITAB: Stat > Basic Statistics ▶ 2-sample t . . . , and *do not* check the box labelled Assume equal variances.

### ▷ EXAMPLE. iBobbly (continued) (iBobbly.mwx)

For the iBobbly data, the test assuming a common variance  $\sigma^2$  gave  $t = 2.12$  and  $P = 0.038$ . The sample standard deviations in the two groups are very close (8.04 and 7.76), so the test statistic is the same to 2 decimal places whether we assume equal population variances or not:  $t = 2.12$ . There is (in this case) only a slight change to the degrees of freedom, from 59 to 58, so the  $P$ -value for the 2-sided alternative (when not assuming a common variance) turns out to be different only in the third decimal place,  $P = 0.039$ . This difference is immaterial.

## 11.2 Mann-Whitney test

One of the strategies for dealing with data that do not have the underlying “standard” structure is to use a distribution-free technique. We saw a couple of examples of this in Chapter 9, namely, the sign test and the Wilcoxon signed rank test. These were both for paired data.

There is a rank based test for the case of two independent samples: the Mann-Whitney test.

▷ **EXAMPLE. Insurance premiums** (claims.mwx)

A long time ago, an insurance company charged the same premium for car makes I and II. It suspected, however, that repair costs were not the same for the two makes. Random samples of claims gave the following data:

	Claims in dollars								
Make I	\$353	\$597	\$634	\$696	\$1913	\$649	\$593	\$658	\$2994
Make II	\$453	\$527	\$1725	\$228	\$568	\$523	\$568	\$155	

Are the underlying distributions of claims the same for each make?

In the example, essentially, we want to know whether the population means are the same or not; if there is evidence that they are different, there could be a reason to consider different premiums for the two makes.

We could consider using the 2-sample t-test to address the question about the means. However, the data do not appear to be Normally distributed:

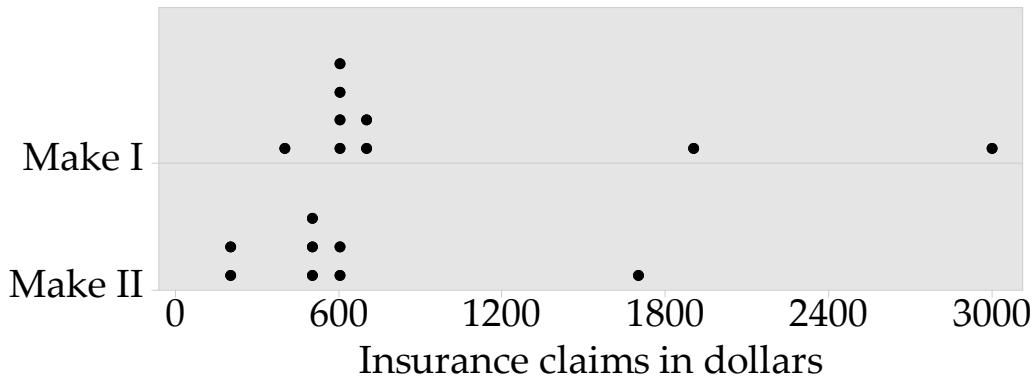


Figure 85: Dotplots of the insurance claims data by make of car.

When the sample sizes in each of two samples are small, and there is strong evidence that the data do not come from a Normal distribution, the Mann-Whitney test can be used to answer this question. This test is also known as the Wilcoxon rank-sum test. It is based on the relative magnitudes (ranks) of the observations in the two samples. It requires no assumptions other

than the independence of the observations.

$H_0$ : There is no difference between the two populations;

$H_1$ : There is a difference in location between the two populations.

Let  $m$  and  $n$  ( $m \leq n$ ) be the sizes of the two samples. The test makes use of the result that, when  $H_0$  is true, the probability that any of the ranks  $1, 2, \dots, m+n$  is associated with a particular population depends only on the values of  $m$  and  $n$ .

The procedure is as follows.

1. Determine the sample sizes  $m$  and  $n$ , where  $m \leq n$ .
2. Rank the observations from smallest to largest across the two samples taking signs into account (e.g.  $-3.1$  would have a lower rank than  $+0.5$ ). For tied observations assign the average of the tied ranks.
3. Determine  $W$ , the sum of the ranks associated with the smaller sample.
4. For small  $n$  (say,  $n \leq 10$ ) it is possible to compare  $W$  with the “null” distribution of the Mann-Whitney (or Wilcoxon rank-sum) distribution. For  $n > 10$  there is an approximation

$$W \xrightarrow{d} N\left(\frac{1}{2}m(m+n+1), \frac{1}{12}mn(m+n+1)\right).$$

MINITAB uses the Normal approximation for all values of  $m$  and  $n$ .

The Mann-Whitney test is most appropriate when it is not reasonable to assume a particular form of distribution for the two populations (usually Normal distributions), but where the data can be ranked.

#### ▷ EXAMPLE. Insurance premiums (continued)

$H_0$ : There is no difference between the distribution of claims for the two makes of car.

$H_1$ : There is a difference in location between the distribution of claims for the two makes of car.

	Ranks of claims									
Make I	3	10	11	14	16	12	9	13	17	
Make II	4	6	15	2	7.5	5	7.5	1		

$w$  =sum of the ranks for Make II = 48. We find that the  $P$ -value is 0.02.

In MINITAB: Stat > Nonparametrics ▶ Mann-Whitney ...

There is a corresponding confidence interval for the true difference in location, which assumes that the only difference between the two distributions is a location shift.

In this example, the 95% confidence interval for the true difference in location is (29, 541).

### 11.3 Assumptions of the model and model-checking

We now deal with assumptions when drawing inferences about means from several independent samples.

The one-way ANOVA model of Chapter 10 can be thought of in the following way:

$$Y_{ij} = \mu_i + E_{ij},$$

where the ‘random errors’  $E_{ij}$  are considered independent  $N(0, \sigma^2)$  random variables.

We can think of each observation  $Y_{ij}$  as being oriented around a mean  $\mu_i$ , and ending up at a point some distance from  $\mu_i$ , by making the ‘jump’ caused by  $E_{ij}$ .

This perspective is the one which most readily extends to situations with greater complexity.

We have already stated the ingredients of the model for one-way ANOVA. In summary:

- The data are assumed to come from  $k$  distributions with (possibly) different means  $\mu_i$ .
- The variances of each of the  $k$  distributions are assumed to be the same,  $\sigma^2$ .
- The distributions are assumed to be Normal.
- All observations are assumed to be independent.

There are a number of assumptions here. Are they correct, for the data we are analysing? How can we assess that? And does it matter?

Hard and fast rules are not really possible here. In essence, the assumptions matter because if they are not met, the inferences drawn may be incorrect, in a variety of ways. We might carry out the ANOVA and obtain a  $P$ -value for the test of  $H_0$ :  $P = 0.03$ , say. But if the assumptions of the ANOVA are incorrect, the true  $P$ -value could be quite different,  $P = 0.11$ , say. So the assumptions do matter.

Or other inferences, such as a confidence interval for the difference between the means of the two of the groups, may be wider than it should be. Or narrower than it should be.

But the impact of the failure of the assumptions is a complex mix of the extent of the failure (are the population variances only a little bit different, or very different?) and sample size, among other things.

Of the four assumptions listed above, the fourth—*independence*—has to be assessed, in general, from first principles. This means that we consider whether there are reasons why the observations should be dependent in some way. Common situations where independence is violated are that the

observations are in a time sequence, and observations close together are likely to be similar; observations which are ostensibly from different subjects are really from the same subject; observations are close in space and are affected by neighbouring observations; observations are from larger groupings (e.g. school classes, families etc.).

Ignoring statistical dependence in ANOVA can be a very big problem. Suppose you are assessing the effect of a treatment on hypertension, and hence you are measuring blood pressure. On each individual, just to be thorough, you take three measurements on each arm; six altogether. Suppose there are 20 individuals in treatment group A, another 20 in treatment group B, and 20 in the control group. It would be completely wrong to analyse these data using one-way ANOVA, with 120 observations in each of the three groups, because the six measurements on each individual would not be statistically independent. (These data can be analysed in a valid way, but not by a naive one-way ANOVA.)

The Normality and constant variance assumptions can be assessed in a number of ways, including by consideration of the **residuals**.

### 11.3.1 Residuals

$\mu_i$  is estimated by  $\bar{Y}_i$ , the mean of the sample from the  $i$ th population, and

$$\hat{E}_{ij} = Y_{ij} - \bar{Y}_i,$$

the estimates of the random errors, are referred to as the **residuals**.

The idea of a residual extends to more complex models; it is defined as the difference between the observed value, and the “fitted value”, after the model has been estimated. So if the model is a good fit, then the residuals should not be unusually large. But the problem with interpreting a residual as being large or not is that it depends on  $\sigma$ , the residual standard deviation. Further, while the residuals have zero mean they do not have constant variance:  $\text{var}(\hat{E}_{ij}) = \sigma^2 - \text{var}(\bar{Y}_i)$ . One way to understand this is to think of two possible extremes. If the sample size in one of the groups is huge, then the variance of  $\bar{Y}_i$  will contribute very little to the  $\text{var}(\hat{E}_{ij})$ , so that it will be essentially equal to  $\text{var}(Y_{ij})$ , which is  $\sigma^2$ . On the other extreme, if there is only one observation in group  $i$  then the sample mean for that group will be exactly equal to that observation (i.e.  $\bar{Y}_i = Y_{i1}$ ) and hence the residual for that point has to be zero, and  $\text{var}(\hat{e}_{i1}) = 0$ .

In order to make the interpretations of the residuals easier, it is common to look at the **standardized residuals**, which are defined by

$$E_{ij} = \frac{\hat{E}_{ij}}{\text{se}(\hat{E}_{ij})}$$

(where “se” stands for “standard error”) so that the standardized residuals have zero mean and a standard deviation of approximately one, if the model is correct.

$$\text{se}(\hat{E}_{ij}) = \sqrt{s^2 - \widehat{\text{var}}(\bar{Y}_i)} = s \sqrt{1 - \frac{1}{n_i}}.$$

### 11.3.2 Model-checking

We now consider a variety of approaches to checking the assumptions of constant variance and Normality. Among these two assumptions, the more important issue is constant variance in the groups, because the Central Limit Theorem makes the assumption of Normally distributed random errors relatively unimportant as the sample sizes grow large, if the constant variance assumption is met. The approaches are in two general categories:

- Formal tests of the assumptions;
- Informal assessments based on appropriate graphs, based on residuals.

The formal tests are often called ‘diagnostic tests’. We have an unusual disposition towards the results of diagnostic tests: *we are hoping for a large P-value*. This is because we are testing a null hypothesis which asserts that an assumption of the ANOVA is true. In general, we would like it to be true. If the data are consistent with the assumption tested, the diagnostic test will have a large *P*-value, and — at least as far as that test goes — we can proceed with the ANOVA. We should worry about small sample sizes, however: the diagnostic tests, like any statistical test, will not be very powerful in small data sets, so the departure from an assumption needs to be large before it is detectable.

We now list a variety of approaches to assessing the assumptions of Normality and constant variance.

- (a) Plot the data, by group. If the sample sizes are large, and any sensible graphical representation of the data shows that in each group there is symmetry at least, if not an approximate Normal shape, and close to the same spread, that will give us some initial reassurance. If that is not the case, it may alert us to what we should be concerned about.
- (b) For one-way ANOVA, it is possible to carry out a formal test of the assumption of constant population variance. That is, we test  $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ . This can be done in MINITAB: Stat > ANOVA ▶ Test for Equal Variances.... There are a number of options. Levene’s test does not assume normality within the groups. Bartlett’s test does. However, these formal tests do not extend to more complex models.

For the white pine data, Levene’s test gives  $P = 0.84$ , while Bartlett’s test gives  $P = 0.57$ . Neither test gives cause for concern about the

constant variance assumption, but the sample sizes are small and so the power of the tests will be correspondingly low.

The standardized residuals can be used to check the assumptions of the model by plotting them in various ways.

3. For one-way ANOVA, we plot the standardized residuals against  $i$  to examine the assumption of equal variances; we look to see if the ‘spreads’ are much the same for the different groups. More generally, in more complicated models, we plot the standardized residuals against the fitted or predicted values from the model, and look for an absence of pattern, or a random cloud of points.
4. We can do a histogram of the standardized residuals to try to assess whether they look like a sample from a Normal distribution, but it is hard to check Normality from this alone. What we do is to plot the cumulative distribution of the residuals on a scale which would be expected to show a straight line (approximately), if the Normality assumption is justified. This is called a “Normal probability plot”.
5. In MINITAB, the one-way ANOVA commands do not give standardized residuals. However, using the General Linear Model command they can be obtained and stored in a column: use Stat > ANOVA ▶ General Linear Model ... Storage, and check the box labelled Standardized residuals. Once the standardized residuals are in a column, they can be plotted against the group variable, and the test for Normality can be obtained: use Stat > Basic Statistics ▶ Normality Test. MINITAB gives a choice of tests; they have different properties. A good general test is the one labelled Ryan-Joiner. In MINITAB, using the General Linear Model menus, you can ask for Graphs and choose the standardized residuals. This can be used to obtain a Four in one plot; no test of Normality is provided however, but it gives a quick visual check.

The test of Normality described here is a completely general technique for assessing Normality, which can be used on any sample whose Normality you want to check, but you need to be aware that it will be hard to have statistically significant evidence against Normality when the sample size is small.

One minor point about the Normality assumption: it is vital to keep in mind that it is the Normality of the random errors that we are concerned about, not the Normality of the data set as a whole. If you simply ‘plot all the

data', the distribution may be affected by differences between the means, and a distinctly non-Normal pattern may be present. But that doesn't mean, necessarily, that the assumption of Normality is violated. You need to be thinking about the random errors, and hence considering the residuals.

## 11.4 Dealing with model violations

It is one thing to be aware of model assumptions and another to address them. What process should be followed if there is clear evidence of assumptions being violated? What if the variances in the groups do differ markedly, for example?

There are a number of possible strategies.

### 11.4.1 The Kruskal-Wallis test

One way to test the hypotheses is to use the Kruskal-Wallis test. This test is based on the relative magnitudes (ranks) of the observations from the  $k$  populations. It is an extension of the Mann-Whitney test. It requires no assumptions other than the independence of the observations.

Let  $n_i$  be the sample size from the  $i$ th population and let  $N = n_1 + \dots + n_k$  be the total number of observations. The test makes use of the result that, when  $H_0$  is true, the probability that any of the ranks,  $1, 2, \dots, N$ , is associated with the  $i$ th population depends only on the values of  $n_i$  (and  $N$ ).

#### Procedure

- Rank all of the observations from the  $k$  samples in a single series. For tied observations assign the average of the tied ranks.
- Determine  $R_i$ ,  $i = 1, 2, \dots, k$ , the sum of the ranks associated with the  $i$ th population.
- Calculate

$$H = \frac{12}{N(N+1)} \left( \sum \frac{R_i^2}{n_i} \right) - 3(N+1).$$

If there are many tied observations, a correction<sup>25</sup> needs to be made which has the effect of increasing the value of  $H$ .

- There are tables of the exact distribution of  $H$ , assuming that  $H_0$  is true. However, MINITAB uses the approximation

$$H \stackrel{d}{\approx} \chi_{(k-1)}^2, \quad \text{for all } k.$$

---

<sup>25</sup>Details about the correction can be found in Conover W.J. (1980). *Practical Nonparametric Statistics* 2nd ed. John Wiley & Sons.

The Kruskal-Wallis test is most appropriate when it is not reasonable to assume a particular form of distribution for the populations (usually Normal distributions) but where the observations can be ranked.

▷ **EXAMPLE. White pine** (*continued*)

$H_0$ : There are no differences in the moisture content distributions

$H_1$ :  $H_0$  is not true.

Conditions	Ranks					$R_i$
1	3	6.5	5	9	6.5	30
2	1	4	2			7
3	8	10	11			29

Applying the above result we find that

$$H = \frac{12}{11 \times 12} \left( \frac{30^2}{5} + \frac{7^2}{3} + \frac{29^2}{3} \right) - (3 \times 12) = 7.33, \quad P = 0.03;$$

and after adjusting for the tied ranks  $H = 7.37, P = 0.03$ .

In MINITAB: Stat > Nonparametrics ▶ Kruskal-Wallis.

### 11.4.2 Transformations

Residual plots or diagnostic tests may show patterns which suggest that the assumptions of the model are not right. A usual way to tackle this problem is to transform the response variable,  $Y$ .

The idea is to get to a situation where the assumptions of the model are met, albeit on a transformed scale. There are theoretical arguments to suggest that a transformation may help with stabilizing the variance in some cases. For example, if the response variable is a **count** then it may be that taking the square root of the response will render the usual assumptions more reasonable, although there are, in general, more sophisticated methods for dealing with outcomes such as counts and proportions directly.

More generally, we seek transformations of  $Y$  which give a more satisfactory “fit” to the assumptions of the model, than on the original scale. So we might transform  $Y$  by taking square roots, or the logarithm of the data, or the reciprocal. All these, and other possibilities, are in a useful family of transformations indexed by  $c$ :

- $Y^c$  if  $c > 0$ , e.g.  $Y^{\frac{1}{2}}$  (i.e.  $\sqrt{Y}$ );
- $\log(Y)$  if  $c = 0$ ;
- $-Y^c$  if  $c < 0$ , e.g.  $-Y^{-1}$  (i.e.  $-\frac{1}{Y}$ ).

The reason for the minus sign when  $c < 0$  is simply to preserve order; for example, if  $c = -1$  then we are taking reciprocals of the data, and (without

the minus sign) the originally largest observation becomes the transformed smallest. With the minus sign (which makes no essential difference to any of the inferences) the data are in the same order as originally.

The logarithmic transformation is one of the commonest. It is often found that the standard deviation of the residuals increases proportionally with the fitted values: the plot “fans out”. In this case the log transform will often make sense, and may be applicable anyway, because (for example) effects are operating on a percentage change basis rather than an absolute scale.

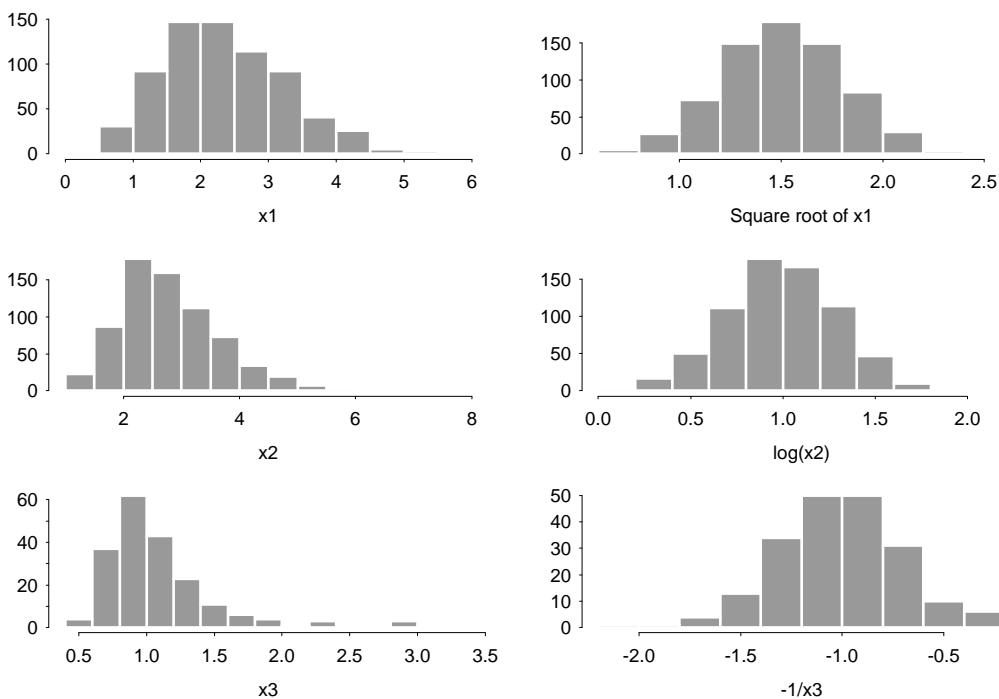


Figure 86: Three different sets of data with varying skewness, showing the effect of three transformations.

Transformations with  $c < 1$  for the response variable are most common (especially  $c = 0$ ) because they help in the situation where the residuals are skewed to the right, and this is a common finding. Smaller values of  $c$  are “stronger”, in the sense that the greater the (right) skewness, the smaller the value of  $c$  needs to be. So the square root transformation ( $c = 0.5$ ) may help with residuals that are not very skew, the log transformation ( $c = 0$ ) may be needed if the residuals are more skew, and the reciprocal transformation ( $c = -1$ ) might be required if the residuals are very skew indeed: see the examples in Figure 86.

Representing the results of an analysis in which a transformation has been used is not straightforward. This is because, in general, differences between

the means of the transformed data cannot be back-transformed to differences of means on the original scale. One simple approach is to carry out and report hypothesis tests (overall  $F$  test, pairwise comparisons) on the transformed scale, and simply report the point estimates of differences on the original scale.

### 11.4.3 One-way ANOVA without assuming equal variances

Recall that for the two sample  $t$ -test, there are two basic versions. One assumes equal variances in the two groups, and the other does not.

There is an equivalent situation with one-way ANOVA. There is an approximate test of  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  that does not assume equal variances. It is known as Welch's test.

In the One-Way menu, under Options, untick the box labelled Assume equal variances. An approximate  $F$  test is given. For the white pine data, this test gives  $F = 5.66$ , referenced to the  $F$  distribution with degrees of freedom 2 and 3.61, giving a  $P$  value of  $P = 0.077$ .

Note that in this case, whether or not the equal variance assumption is made makes a marked difference to the overall  $P$ -value;  $P = 0.008$  assuming equal population variances,  $P = 0.077$  without this assumption. The formal tests of this assumption did not show strong evidence against it.

This approach does not generalise to more complicated contexts, and that is one reason why there is a strong focus on the equal variance assumption; models with a rich structure of explanatory variables can be accommodated with a random error term that has constant variance.

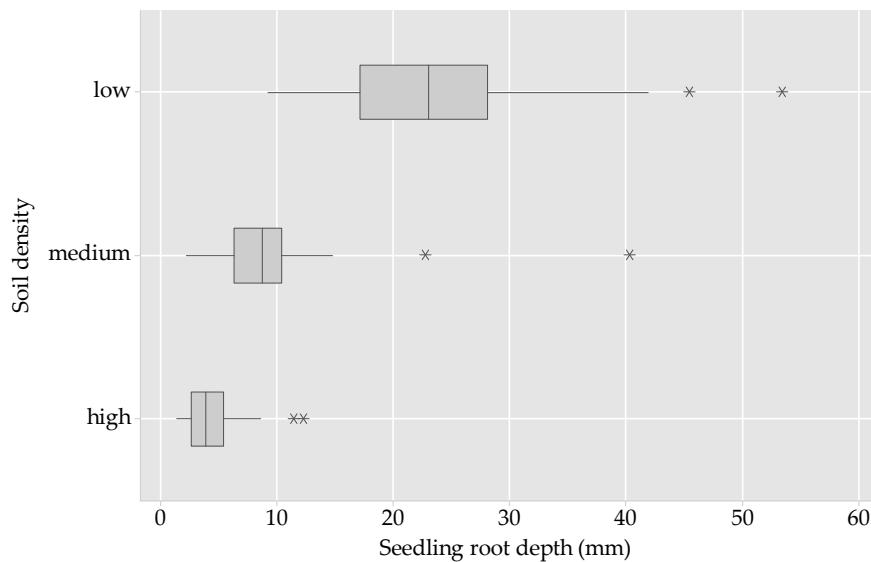
In MINITAB, a multiple comparisons procedure is provided that can be applied when the assumption of equal population variances is *not* made. It is known as the Games-Howell method. The interpretation of the results is as for Tukey's procedure.

## 11.5 Exercises

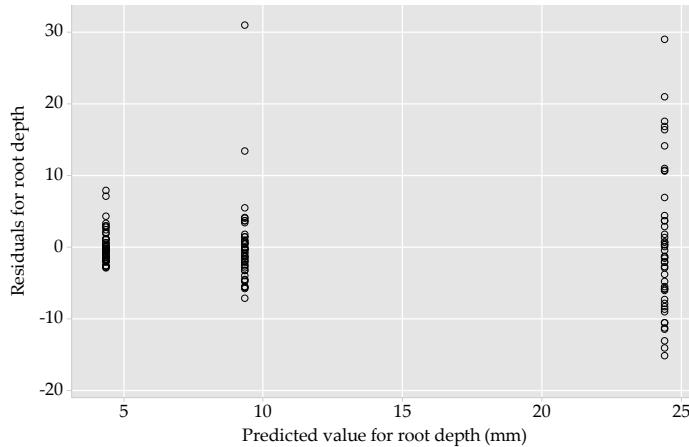
- 11.1 In Exercise 10.1, you carried out inferences for the waiting time prior to surgery, comparing a new method and the standard technique. What difference does it make to the 95% confidence interval, and the result of the *t*-test, if you do not assume equal population variances?
- 11.2 In Chapter 10 (exercise 10.3), you analysed a wine growing experiment that used three treatments in each of six different areas of the vineyard, using the data set, *pinot.mwx* to address the question: Does the treatment influence the size of the yield obtained at harvest?

You described the standard assumptions required for analysing the data using a linear model . Investigate the follow assumptions:

- (a) The three distributions from which the samples in each treatment are taken must have the same variance.
  - (b) The underlying distributions of harvest weights for each of the treatments should be Normal.
- 11.3 Consider data from smokers in the National Health and Nutrition Examination Survey. The MINITAB worksheet is *NHANES.mwx*. Consider the analysis you carried out of Serum cholesterol (mg/100ml) in exercise 10.4 where you fitted a linear model with ‘Activity level’ as a categorical explanatory variable.
- (a) Examine the equality of the variances. Does it appear reasonable to assume that the variances are equal? [Inspect the 4-in-1 residual plots. Carry out a formal test. ]
  - (b) Comment on the Normality of the standardised residuals.  
[ Again consider the 4-in-1 residual plots. ]
- 11.4 The data stored in the file *rtdepth.mwx* are the results of an experiment to examine the effect of soil bulk density on root growth. Seedlings were randomly allocated to plots with soil densities at three levels (low, medium and high). At the end of the study, root length (in mm) was measured. The figure below provides boxplots of root length for the three different soil densities.

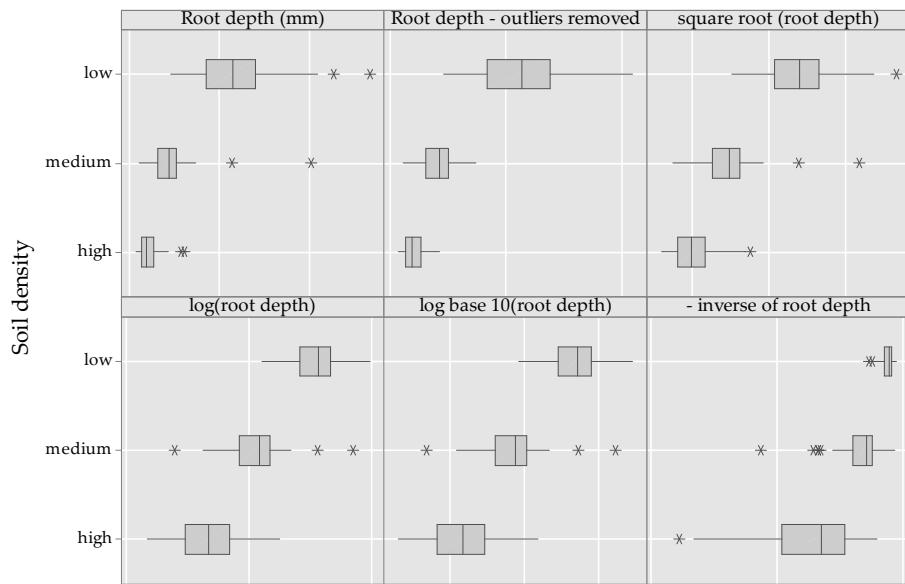


- (a) Consider using a linear model to test the hypothesis of no difference in mean root depth between the three levels of soil density. Based on the boxplots above, are there any assumptions that you are concerned about? Which one(s)?
- (b) The figure below shows a plot of standardized residuals against the fitted values when a linear model is used to analyse the data described above.



Based on the plot above, are there any assumptions of the linear model that you are concerned about? Which one(s)?

- (c) The figure below shows boxplots of the original data (in the top left panel) and several different attempts used by students to deal with problems identified above. ‘square root(root depth)’, for example, indicates the student created a new variable for analysis which is the square root of the original root depth measurements.



- (i) Which strategies are likely to successfully deal with the assumptions that were problematic in the original analysis?
- (ii) Four of the strategies involve a transformation and one removes the outliers. Is it reasonable to use the strategy of removing the outliers?
- (iii) Suppose that the outlier removal strategy resulted in box-plots with similar spread. Would it then be reasonable to use the strategy of removing the outliers?
- (iv)\* Fit the linear model using the strategy you prefer. Data corresponding to the various strategies used by students are provided in the data file. [Use the Stat > ANOVA > General Linear Model menu.]

## 11.6 Answers

11.1 For these data, the standard deviations were 3.11 minutes for the new method and 2.45 for the standard technique, which are quite close.

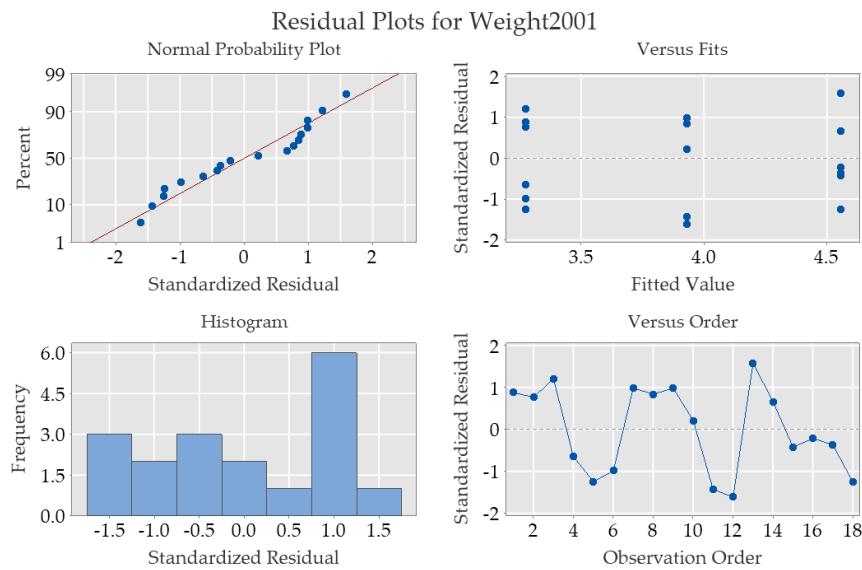
The analysis done in Chapter 10, assuming a common  $\sigma^2$ , gave a difference of means of 3.03 minutes (new minus ‘standard’), a 95% CI of (0.66 to 5.40) minutes, and a  $P$ -value of  $P = 0.014$ .

If we do not assume a common variance, the 95% CI is (0.65 to 5.42) minutes and the  $P$ -value is  $P = 0.015$ .

There there are no meaningful differences between the two sets of results.

11.2 (a) From the dotplots produced for Chapter 10, the variability in the three samples looks similar; there is no evidence to suggest the assumption of similar variances in the underlying populations is not reasonable.

(b) We can also examine the residuals from the General Linear Model fit to assess the assumptions. The plot below does not suggest any problems with the residuals.



- 11.3 (a) The formal test of homogeneity of variance for the three activity groups is shown below.

You can find this under Stat > Anova > Test for equal variances.

#### Test for Equal Variances: Serum cholesterol (mg/100ml) versus Activity level

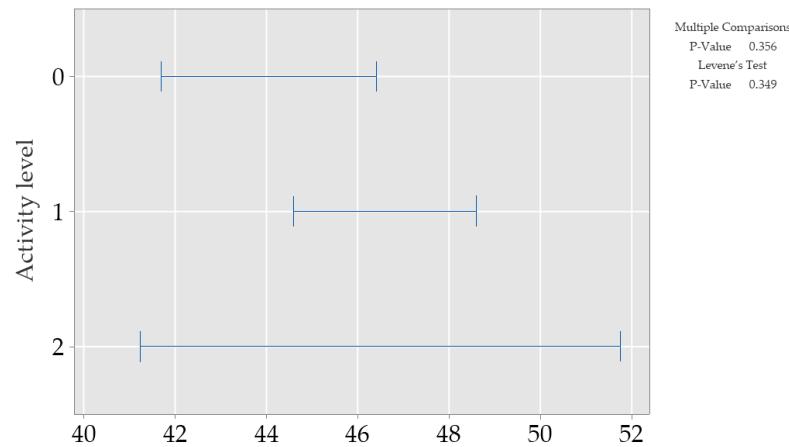
##### Method

Null hypothesis	All variances are equal
Alternative hypothesis	At least one variance is different
Significance level	$\alpha = 0.05$

##### Tests

Method	Test	
	Statistic	P-Value
Multiple comparisons	—	0.356
Levene	1.05	0.349

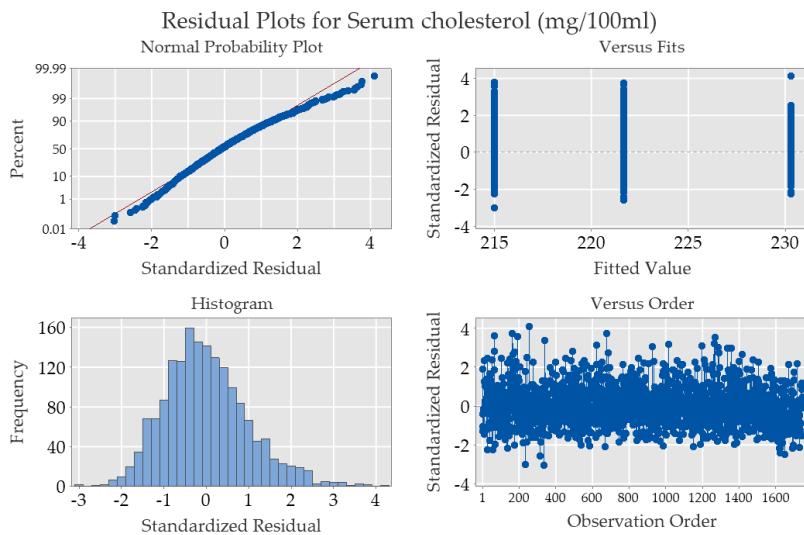
Test for Equal Variances: Serum cholesterol (mg/100ml) vs Activity level  
Multiple comparison intervals for the standard deviation,  $\alpha = 0.05$



If intervals do not overlap, the corresponding std devs are significantly different.

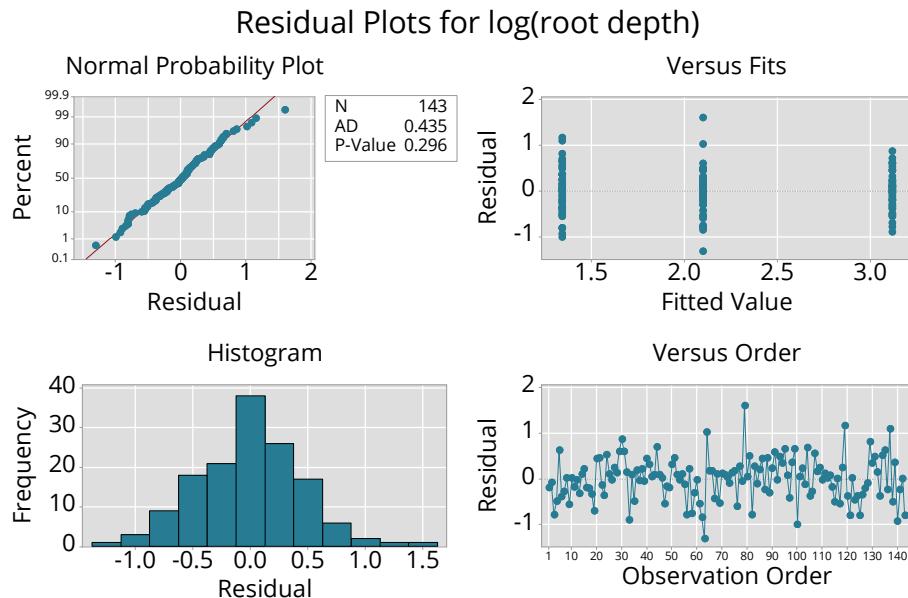
The data are consistent with the assumption of homogeneity of variance, according to this analysis.

- (b) The standardised residuals do not look consistent with an underlying Normal distribution. However, the analysis is robust to violations of this assumption, given the sample size.



- 11.4 (a) A striking problem in the boxplots is the difference in the variability between the groups. The linear model assumes that the variance of the three underlying populations from which the samples are taken are the same. The three samples obtained do not look consistent with this assumption.
- (b) The plot of standardized residuals against the fitted values reflects the same problem identified in the boxplots; the residuals vary with the predicted values. This is the kind of pattern we expect to see when the assumption of constant variance is not satisfied.
- (c) (i) The logarithmic transformations (either natural log or log base ten) appear to result in boxplots with much more consistent variability. These are likely to successfully deal with the problem of the assumption of constant variance.
- (ii) The strategy of removing outliers does not solve the problem of lack of constant variance. It is not a reasonable strategy to use. The linear model requires constant variance, but we should not manipulate the data to fit the assumptions of the model. Rather we should find an appropriate model for the data observed.
- (iii) Even if the strategy of removing outliers does solve the problem of lack of constant variance, it is not a reasonable strategy to use. The linear model requires constant variance, but we should not remove data to fit the assumptions of the model. Rather we should find an appropriate model for the data observed.
- (iv) The analysis using the natural log transformation is presented here. First we check if the problem with the lack of constant

variance has been resolved. The plot below of the residuals versus the fitted values suggests that it has.



Some relevant output is:

### General Linear Model: log(root depth) versus Soil density

#### Method

Factor coding (-1, 0, +1)

#### Factor Information

Factor	Type	Levels	Values
Soil density	Fixed	3	high, medium, low

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Soil density	2	76.99	71.14%	76.99	38.4937	172.52	0.000
Error	140	31.24	28.86%	31.24	0.2231		
Total	142	108.22	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.472359	71.14%	70.72%	32,5873	69.89%	196.57	208.13

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	2.1855	0.0395	(2.1074, 2.2637)	55.29	0.000	
Soil density						
high	-0.8463	0.0552	(-0.9555, -0.7372)	-15.33	0.000	1.32
medium	-0.0862	0.0564	(-0.1977, 0.0253)	-1.53	0.129	1.32
low	0.9325	0.0561	(0.8217, 1.0434)	16.63	0.000	*

#### Regression Equation

$$\text{log(root depth)} = 2.1855 - 0.8463 \text{ Soil density\_high} - 0.0862 \text{ Soil density\_medium} + 0.9325 \text{ Soil density\_low}$$

Note that the order of the soil density variable has been changed to High, Medium, Low; this provides positive mean differences.

Here is the output describing pairwise comparisons:

### Tukey Simultaneous Tests for Differences of Means

Soil density <u>Levels</u>	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
medium - high	0.7601	0.0965	(0.5315, 0.9887)	7.88	0.000
low - high	1.7789	0.0960	(1.5516, 2.0062)	18.54	0.000
low - medium	1.0187	0.0980	(0.7867, 1.2508)	10.40	0.000

Individual confidence level = 98.08%

An appropriate summary table is:

Density	Log(seeding root depth mm)		
	Mean	Standard deviation	n
Low	3.12	0.40	47
Medium	2.10	0.50	46
High	1.34	0.51	50

One-way ANOVA:  $F_{(2,140)} = 172.5, P < 0.001$

Comparison of densities	Estimate	Difference in means	
		95% confidence interval*	P-value
Low – Medium	1.02	0.79, 1.25	<0.001
Low – High	1.78	1.55, 2.01	<0.001
Medium – High	0.76	0.53, 0.99	<0.001

\*Estimated using Tukey's HSD

The table provides the summary statistics and sample sizes, describes the results (test statistic and  $P$  value), and gives Tukey's pairwise comparisons. The estimates of the effects of different soil densities are all on the log scale. Note that the estimates of differences are presented as positive mean differences.



## 12 Inference — numerical outcome and two categorical explanatory variables

In this chapter we meet data structures with a numerical outcome and two explanatory variables, and, specifically, two factors. Remember that a factor is a categorical explanatory variable. Such a data structures arise in more than one context, and the inferences we draw, and how we analyse the data varies with the context.

The presence of the two explanatory factors is a jump in complexity, leading to consideration of the important and general concept of ‘interaction’.

The types of studies we consider here are restricted to designed experiments, in which the researcher has control over the choice of levels in the experiment and the number of observations obtained for each level or combination of levels.

Two types of designs are considered:

- Randomized blocks design;
- A factorial design with two factors.

We will deal systematically with the design of experiments in Chapter 16, so this is a foretaste of issues that arise in that Chapter.

### 12.1 Randomized blocks design

It is useful to introduce the ideas with an example.

▷ **EXAMPLE. Metal hardness** (metal.mwx)

The following data were obtained from an experiment to compare four types of tip for measuring the hardness of metal. Five strips of metal were used and the depth of penetration measured with each tip used on each strip, using a standardised protocol of measurement for all tips and strips. The units are 0.002 mm, so a value of 10.0 means that the tip penetrated to a depth of 0.02 mm.

Strip	Type of tip				Mean
	1	2	3	4	
1	9.3	9.4	9.2	9.7	9.4
2	9.4	9.3	9.4	9.6	9.4
3	9.6	9.8	9.5	10.0	9.7
4	10.0	9.9	9.7	10.2	10.0
5	9.7	9.7	9.5	9.9	9.7
Mean	9.6	9.6	9.5	9.9	9.6

We envisage that the strip of metal used may have an effect on the outcome, and we want to allow for that in the analysis; intuitively, we think that we can make more efficient inferences about differences between the tips (the key question) if we can make the tip comparisons within each strip. This is the idea behind the inferences described here.

The context here is a natural extension of previous material in Chapter 9, as we shall see.

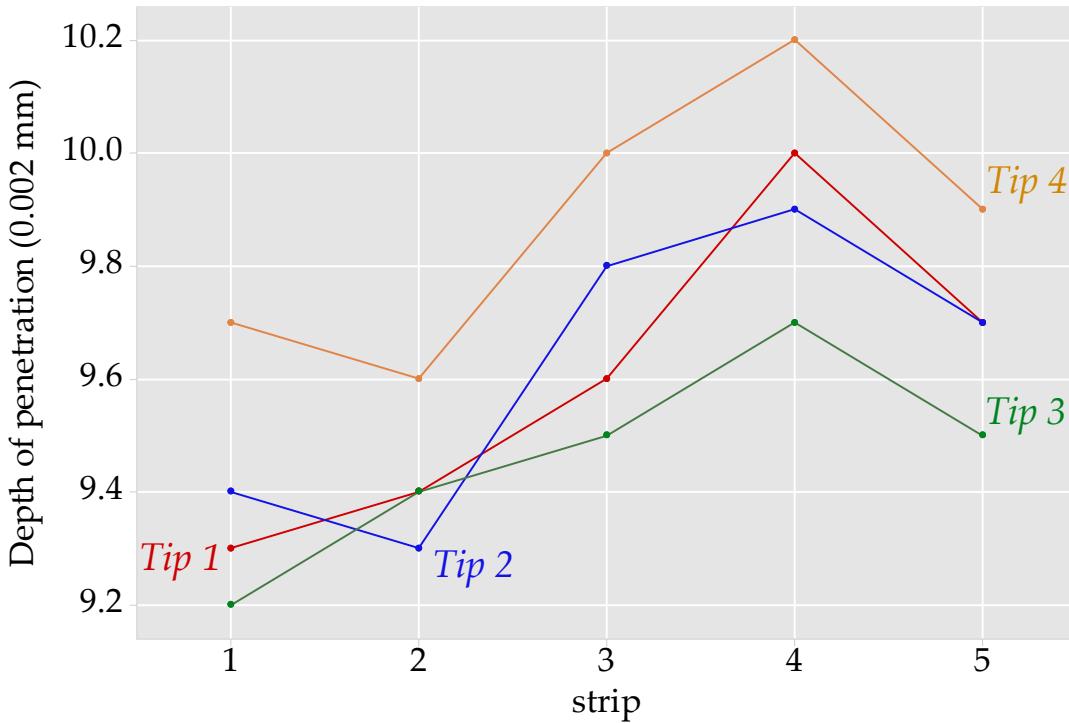


Figure 87: Data from an experiment comparing four types of tip for measuring the hardness of metal.

In general terms, when the data have this structure, we refer to the key factor of interest as the **treatment** factor, and the variable for which we use to increase the precision of the comparisons as the **block** factor. This nomenclature came from agriculture in which the ideas were originally applied. In that context a “block” was an actual area of ground (in, say, a crop yield trial) and the “treatment” referred to something like different fertilisers.

### 12.1.1 Two-way analysis of variance

We may choose to make simple and ‘standard’ linear model assumptions for the data. As always, we need to ask whether these assumptions are reasonable.

We denote the data by  $Y_{ij}$ , where  $i$  indexes the block and  $j$  the treatment.

We assume the linear model

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}$$

is appropriate, where  $Y_{ij}$  denotes the measurement from block  $i$  and treatment  $j$ ,  $\mu$  denotes the ‘overall’ mean,  $\alpha_i$  denotes the effect of the  $i$ th block,  $\beta_j$  denotes the effect of the  $j$ th treatment and the  $E_{ij}$ ’s, the random errors, are a random sample from  $N(0, \sigma^2)$ .

This model is sometimes referred to as the strictly additive model since it is assumed that the differences between the (true) treatment means is exactly the same for all blocks.

Look again at Figure 87. There is plenty of overlap in the distributions of hardness for the four tips. However, if we take into account the strips, we see that for each strip, there are generally consistent and marked differences between pairs of tips. For example, for each of the five strips, tip 4 gives the highest values.

We obtain a form of analysis of variance that takes both strips and tips into account. It shares some features with the analysis of variance we saw in Chapter 10 for one explanatory factor, and has some new ones.

The formal test of the treatment factor is based on a comparison of the between treatments (tips) variation and an estimate of the ‘error’ variance which takes account of both the variation between treatments (tips) and between blocks (strips).

The analysis is an extension of the  $t$ -test for paired samples; if there were only two tips, we would think of the data as paired samples. Here, we consider them as matched samples.

For the above model:

- $\hat{\mu}$  (the estimate of  $\mu$ ) =  $\bar{y}_{..}$ , the mean of all the observations;
- $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ , the mean of the observations from the  $i$ th block minus the overall mean;
- $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$ , the mean of the observations from the  $j$ th treatment minus the overall mean;
- $\hat{e}_{ij}$  (a residual) =  $\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$

The residuals can be used to check the underlying assumptions as described in Chapter 11.

$H_0 : \beta_1 = \beta_2 = \dots = \beta_c$  (i.e. there are no differences between the treatments);  
 $H_1$ :  $H_0$  is not true.

It is also possible to test for differences between the  $\alpha_i$  (blocks), although often this is of secondary interest, or no interest at all.

The results are set out in the form of an analysis of variance table as follows:

### Analysis of variance

Source	df	SS	MS	F
Blocks	$r - 1$	Bl.SS	$\text{Bl.MS} = \frac{\text{Bl.SS}}{r-1}$	$\frac{\text{Bl.MS}}{\text{Res.MS}}$
Treatments	$c - 1$	Tr.SS	$\text{Tr.MS} = \frac{\text{Tr.SS}}{c-1}$	$\frac{\text{Tr.MS}}{\text{Res.MS}}$
Residual	$(r - 1)(c - 1)$	Res.SS	$\text{Res.MS} = \frac{\text{Res.SS}}{(r-1)(c-1)}$	
Total	$rc - 1$	Tot.SS		

“SS” = sum of squares; “MS” = mean square; “Bl.” = Block”; “Tr.” = Treatment; “Res.” = Residual.

To test for differences between treatments compare the value of  $F$ -ratio for treatments  $\frac{\text{TreatmentMS}}{\text{ResidualMS}}$  with the  $F$ -distribution on  $(c - 1)$  and  $(r - 1)(c - 1)$  degrees of freedom.

To test for differences between the blocks compare the value of  $F$ -ratio for blocks  $\frac{\text{BlockMS}}{\text{ResidualMS}}$  with the  $F$ -distribution on  $(r - 1)$  and  $(r - 1)(c - 1)$  degrees of freedom. Usually, the differences between the blocks are not of interest, since we expect the blocks to vary in their response.

The  $F$ -test(s) are most appropriate when the model described above can reasonably be assumed.

▷ **EXAMPLE. Metal hardness (continued)**

### Analysis of variance

Source	df	SS	MS	F	P
Strips	4	0.843	0.211	29.75	< 0.001
Tips	3	0.460	0.153	21.65	< 0.001
Residual	12	0.085	0.0071		
Total	19	1.388			

We can obtain the analysis of variance in MINITAB in different ways. The existence of these different ways is partly due to the gradual development of MINITAB.

- Stat > ANOVA ▶ Balanced ANOVA, and then enter the two explanatory factors in the Model box;
- Stat > ANOVA ▶ General Linear Model ▶ Fit General Linear Model, then enter the two explanatory factors in the Factors box. It is a good idea to get into the habit of using this menu option, because of its generality.

From the overall hypothesis tests we conclude that there are highly statis-

tically significant differences between the strips and highly statistically significant differences between the tips.

Note that this difference between the tips would not be nearly so clear-cut if we did not include the strips in the model. Here is the (one-way) analysis of variance obtained when we do ignore strips:

Analysis of variance

Source	df	SS	MS	F	P
Tips	3	0.460	0.153	2.64	0.085
Residual	16	0.928	0.058		
Total	19	1.388			

Now, the  $P$ -value for the tips is much larger. So we do benefit a lot from being able to make comparisons within strips. This is exactly analogous to the matched-pairs case: the inference is much more precise when we allow for the pairing, than when we ignore it.

As usual, this overall test for (any) differences in location between the tips does not offer much detail.

To carry out multiple comparisons, we can extend Tukey's method to this case. The smallest statistically significant difference for the treatment means is, as before,

$$Q(\alpha; k, \nu) \frac{s}{\sqrt{n}}.$$

Now  $s$  is obtained from the residual mean square, i.e.  $s^2 = \text{ResidualMS}$ ,  $n$  is the number of observations for each "treatment",  $k$  is the number of treatments, and  $\nu$  is the residual degrees of freedom.

This can be obtained automatically in MINITAB using Stat > ANOVA ▶ General Linear Model and selecting Comparisons. Remember this option will become available after you have fitted the model. The results are shown below.

## Comparisons for depth

### Tukey Pairwise Comparisons: tip

#### Grouping Information Using the Tukey Method and 95% Confidence

tip	N	Mean	Grouping
4	5	9.88000	A
2	5	9.62000	B
1	5	9.60000	B C
3	5	9.46000	C

Means that do not share a letter are significantly different.

#### Tukey Simultaneous Tests for Differences of Means

Difference of tip Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
2 - 1	0.0200	0.0532	(-0.1381, 0.1781)	0.38	0.981
3 - 1	-0.1400	0.0532	(-0.2981, 0.0181)	-2.63	0.089
4 - 1	0.2800	0.0532	(0.1219, 0.4381)	5.26	0.001
3 - 2	-0.1600	0.0532	(-0.3181, -0.0019)	-3.01	0.047
4 - 2	0.2600	0.0532	(0.1019, 0.4181)	4.88	0.002
4 - 3	0.4200	0.0532	(0.2619, 0.5781)	7.89	0.000

Individual confidence level = 98.83%

MINITAB produces an elegant graph of the confidence intervals for the pairwise differences, as shown in Figure 88.

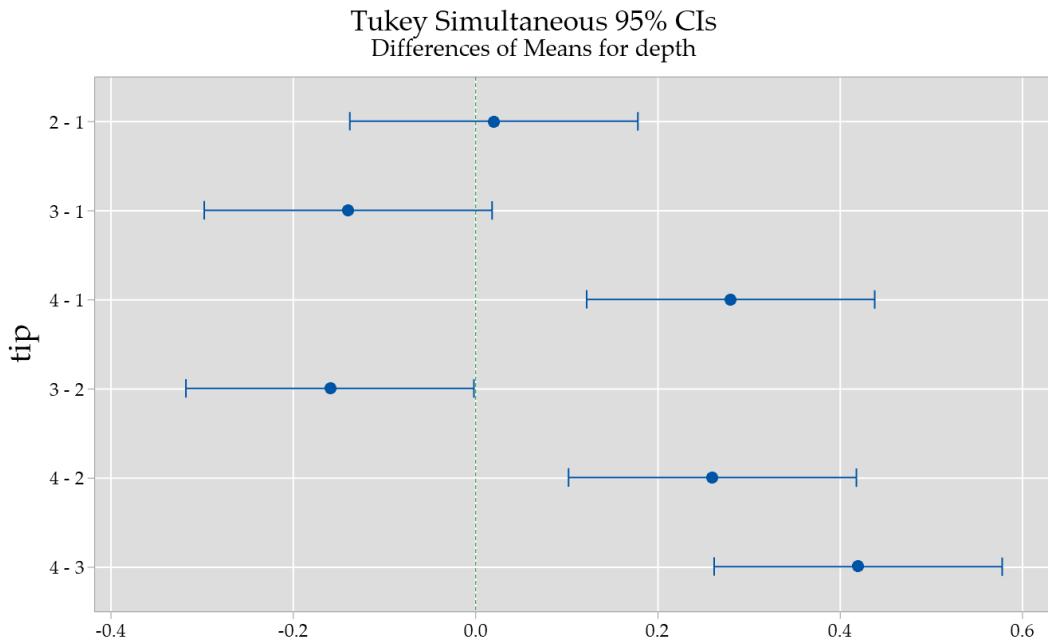


Figure 88: Estimates and 95% confidence intervals for differences between the tip means, allowing for strips, and adjusting for multiple comparisons using Tukey's method.

### 12.1.2 The Friedman test

For data with a randomised blocks structure there is a ‘distribution-free’ test that can be used in small samples, without making an assumption of Normality.

The Friedman test<sup>26</sup> is based on the rankings of the data within each matched sample (or block).

$H_0$ : within each block there are no differences between the treatments;  
 $H_1$ :  $H_0$  is not true.

The test makes use of the result that, when  $H_0$  is true, all possible orderings of the data, within each block, are equally likely.

#### Procedure

- (a) Set out the observations in a two-way table with  $r$  rows (the blocks) and  $c$  columns (the treatments).
- (b) Rank the observations in each row from 1 to  $c$ . For tied observations assign the average of the tied ranks.
- (c) Calculate  $R_j, j = 1, \dots, c$ , the sum of the ranks for the  $j^{\text{th}}$  treatment.
- (d) Calculate
 
$$F = \frac{12}{rc(c+1)} \left( \sum_j R_j^2 \right) - 3r(c+1).$$
 If there are many tied observations, a correction<sup>27</sup> needs to be made.
- (e) For small  $c$  ( $3 \leq c \leq 6$ ), there are tables of the “null” distribution of the Friedman statistic.
- (f) For  $c > 6$ , and large  $r$ , there is an approximation
 
$$F \stackrel{d}{\approx} \chi_{(c-1)}^2.$$
 MINITAB uses the  $\chi^2$  approximation always.

The Friedman test is most appropriate when matched samples (blocks) are available and when it is not reasonable to assume a particular form of distribution for the populations (usually Normal distributions). All that is required is that it be possible to use the observations to rank the data within each of the matched samples.

---

<sup>26</sup>The Friedman after whom the test is named is the famous monetarist economist, Milton Friedman, who created the test in 1937.

<sup>27</sup>Details about the correction can be found in Conover W.J. (1980). *Practical Nonparametric Statistics* 2nd ed. John Wiley & Sons

## ▷ EXAMPLE. Metal hardness (continued)

$H_0$ : there are no differences between the four types of tip.

$H_1$ :  $H_0$  is not true.

Strip	Ranks			
	Type of Tip			
	1	2	3	4
1	2	3	1	4
2	2.5	1	2.5	4
3	2	3	1	4
4	3	2	1	4
5	2.5	2.5	1	4
$R_j$	12	11.5	6.5	20

$F = \frac{12}{100}(12^2 + 11.5^2 + 6.5^2 + 20^2) - 75 = 11.22$ . We find that  $P = 0.011$ . As with many of these tests there is an adjustment for tied ranks, and in this case the adjusted test result is  $P = 0.009$ .

In MINITAB: Stat > Nonparametrics ▶ Friedman.

## 12.2 Factorial experiments with two factors

Another context in which there is a numerical outcome and two categorical explanatory variables (factors) is a so-called “factorial experiment” with two experimental factors.

The difference between this and the randomised blocks design is that in a factorial experiment each of the two explanatory factors reflects a treatment variable, and we are directly interested in the effects of each of these on the response variable.

If the purpose of an experiment is to investigate the effect of two or more factors then it is more informative and more efficient to use various combinations of the factors in the one experiment than to investigate factors one at a time while holding the others constant. Using various combinations of the factors enables us to investigate possible ‘interaction’ between factors as well as the ‘main effect’ of the factors.

Factorial experiments may involve more than two factors, and the levels of any factor may be more than two, leading to more and more complex designs of this type. Having a factorial structure in an experimental design is a very general idea. Here we consider experiments with two factors only.

## ▷ EXAMPLE. Weight gain of pigs (vitamin.mwx)

The following data were obtained from an experiment to investigate the effects of vitamin B<sub>12</sub> and antibiotics on the weight gain of pigs. All of the pigs used were of similar ages but from different litters, the four combinations

of the two factors were each used on three pigs and the response measured was the weight gain (in kg) over a four-week period.

$B_{12}$	Antibiotics	
	No	Yes
No	1.30, 1.19, 1.08	1.05, 1.00, 1.05
Yes	1.26, 1.21, 1.19	1.52, 1.56, 1.55

In an experiment of this sort there are two broad research questions that we can address.

- (a) We may want to know about the effect of one treatment variable, averaged across the levels of the other treatment variable(s). This is known as the **main effect** of a treatment variable.
- (b) Another research question is this: “Is the effect of one of the treatment variables the same at each level of the other treatment variable(s)?” This is about the **interaction** between two or more treatment variables.

### 12.2.1 Interaction

The interaction between two explanatory variables is also sometimes known as ‘effect modification’. While we are seeing it here in a designed experiment, it is a widely applicable concept.

There is no interaction between the two factors if the effect of one of the factors is the *same* at each level of the other factor.

The interaction refers to the extent to which the effects of the two factors are not strictly additive. It is always useful to look at an “interaction plot”, to see possible evidence of non-additivity.

In MINITAB: Stat > ANOVA ▶ Interactions Plot. The interactions plot for the weight gain data is shown in Figure 89.

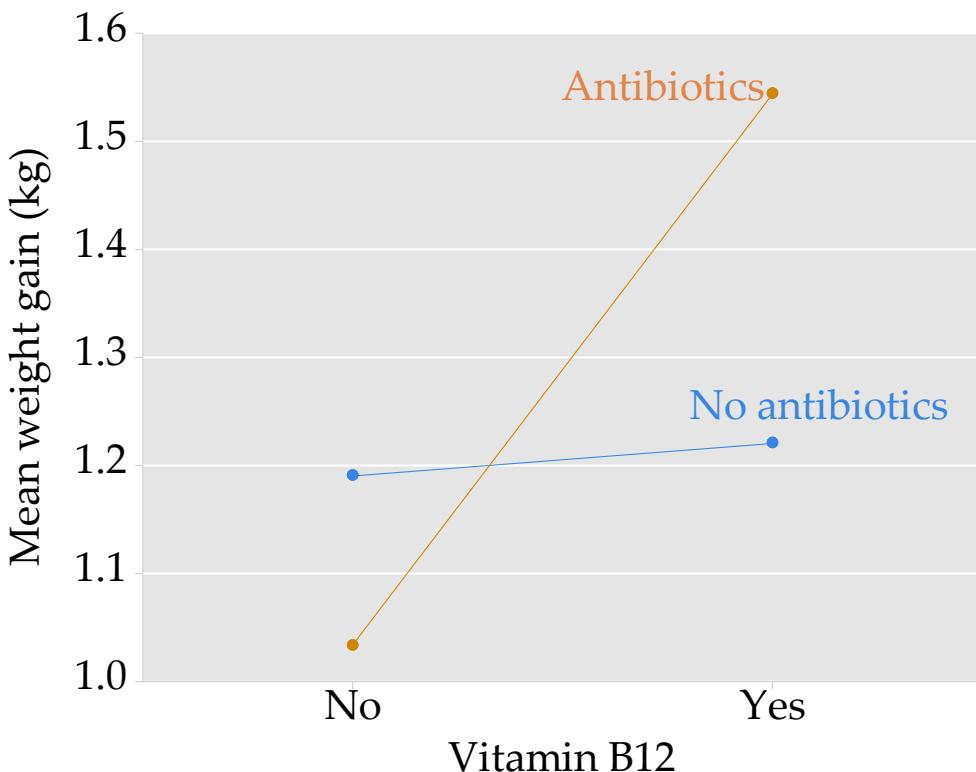


Figure 89: *Interaction graph for weight gain of pigs*

If there is no interaction the lines should be (almost) parallel (they would be exactly parallel if there were no random errors) and by observing where there are substantial departures from parallelism it is, at times, possible to give an interpretation of the interaction.

If two factors interact then obviously they both have an effect, but it may not make much sense to try to interpret the main effects. It is possible for the interaction to be statistically significant but for one or both main effects to be non-significant.

For experiments with  $m$  factors, one can have interaction terms involving up to a maximum of  $m$  factors. However the interpretation gets more and more complicated as the number of factors increases. For example, for an experiment with three or more factors, the 3-factor interaction denotes the changes in the 2-factor interaction pattern for two of the factors over the levels of the third factor.

From the graph, for the weight gain data, it is evident that vitamin B<sub>12</sub> has little, if any, effect on weight gain without antibiotics, but that it has quite a dramatic, positive effect with antibiotics.

### 12.2.2 Analysis of a factorial design

The linear model assumed here is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$$

where

- $Y_{ijk}$  denotes the  $k$ th observation using level  $i$  of factor A (vitamin B<sub>12</sub> in the above example) and level  $j$  of factor B (antibiotics in the above example),
- $\mu$  denotes the overall mean,
- $\alpha_i$  denotes the *main* or average effect of the  $i$ th level of factor A,
- $\beta_j$  denotes the *main* effect of the  $j$ th level of factor B,
- $\gamma_{ij}$  denotes the *interaction* between the  $i$ th level of factor A and the  $j$ th level of factor B, and
- $E_{ijk}$ , the random errors, are assumed to be independent observations from  $N(0, \sigma^2)$ .

The analysis that is appropriate for such experiments is an extension of the analysis of variance techniques described in Section 12.1.1.  $F$ -tests are again used to test for the effect of the factors and for the ‘interaction’ between the factors.

The interaction term needs to be added into the model in MINITAB: General Linear Model > Model.

Analysis of variance

Source	df	SS	MS	F	P
Antibiotics	1	0.0208	0.0209	5.68	0.04
B <sub>12</sub>	1	0.2187	0.2187	59.65	< 0.001
Interaction	1	0.1728	0.1728	47.13	< 0.001
Residual	8	0.0293	0.0037		
Total	11	0.4417			

How should we carry out further inferences when there is strong evidence of interaction, as is the case in the weight gain example? A full answer to this is beyond the scope of the subject. However, note that the combinations of the levels of the factors define meaningful treatment groups, and it is possible, and sometimes helpful, to regard these combinations as making up a single explanatory (treatment) variable. We can then carry out a simple analysis that does not use the factorial structure.

For the weight gain data, we do this by asking for comparisons between the groups defined by the combinations. After fitting the model, Stat > ANOVA > General Linear Model > Comparisons; in the box labelled Choose terms for comparisons, tick the interaction term, Antibiotics\*Vitamin B12. Arising from this analysis, the output below shows the pairwise comparisons

between the four groups, using Tukey's adjustment for multiple comparisons. Note carefully the labelling of the groups in Figure 90: "No Yes" means "No" for the first variable, Antibiotics, and "Yes" for the second, Vitamin B12.

### Grouping Information Using the Tukey Method and 95% Confidence

Antibiotics*Vitamin B12		N	Mean	Grouping
Yes Yes		3	1.54333 A	
No Yes		3	1.22000 B	
No No		3	1.19000 B C	
Yes No		3	1.03333 C	

Means that do not share a letter are significantly different.

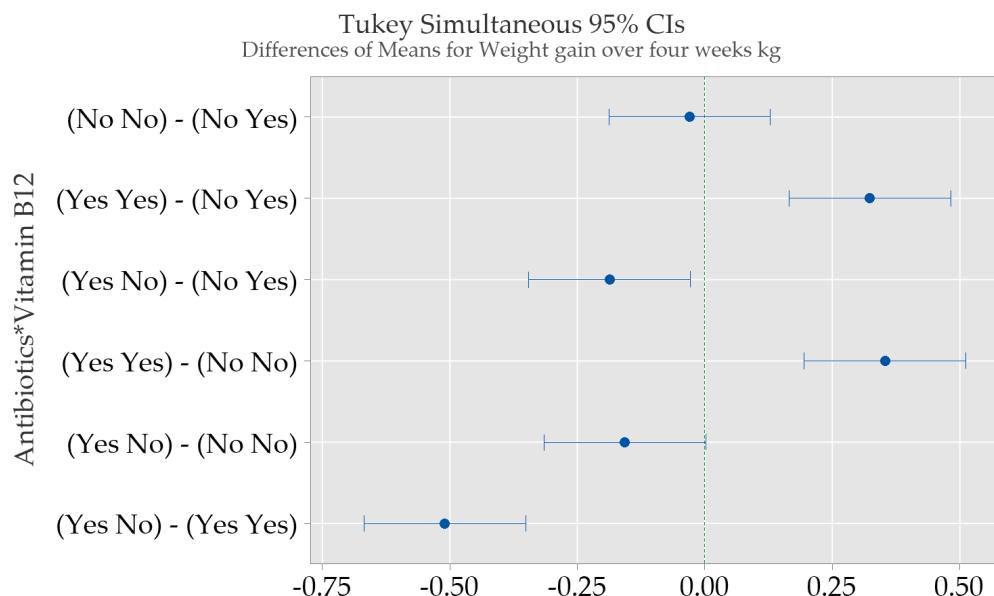


Figure 90: Estimates and 95% confidence intervals for differences between the four treatment combinations in the weight gain experiment, adjusting for multiple comparisons using Tukey's method.

In the absence of an important interaction effect, it is easier to interpret the results by refitting the linear model without the interaction term.

## 12.3 Exercises

- 12.1 An experimenter is interested in the effects of different types of conditioning in mice. She has data on 15 wild type mice, 3 from each of five different strains. Three mice from each strain were assigned at random to one of three conditioning programs. After conditioning, the mice were required to run a maze and the following times to complete the maze (in seconds) were obtained, stored as **maze.mwx**.

Conditioning program	strain				
	A	B	C	D	E
1	49	60	48	53	51
2	50	59	56	57	51
3	63	72	66	69	61

- (a) Assuming a model with Normally distributed errors, carry out an analysis of these data, including the use of multiple comparisons where appropriate.

[A useful plot may be obtained using Graph > Scatter plot > With Groups; select Maze completion time (seconds) — this puts Maze completion time (seconds) under Y variables; select Program — this puts Program under X variables; click in Categorical variables for grouping, select Strain; click OK.]

An analysis of variance can be carried out using Stat > ANOVA > General Linear Model > Fit General Linear Model; select Maze completion time (seconds) as the Responses; click in the Factors box; select Program and Strain; click Options and then under Means select All terms in the model, then click OK; click Results and then tick the box labelled Means; click OK twice.

Use the 4-in-1 residual plots to examine the assumptions of constant variance and Normality of standardized residuals.]

- (b) Which explanatory variable, conditioning program or strain, is of primary interest to the researcher?  
 (c) What difference does it make whether or not you allow for strain in assessing the effect of the program?

- 12.2 In Chapters 10 and 11 you have been considering Vincent Lakey's experiment using a General Linear Model with one categorical explanatory variable (one-way analysis of variance). In fact, Vincent's experiment used blocking. Each treatment was used in each of six blocks.

- (a) Produce a suitable plot that shows reveals both the treatment and block effects on the weight of bunches harvested in 2001. Does it appear from the plot that taking account of blocking will produce a more sensitive inference of the treatment effect? Explain.  
 (b) Fit a General Linear Model (two-way analysis of variance) to the weight of bunches harvested in 2001, including both block and treatment in the model.

- (c) Now find, again, 95% confidence intervals for comparing the mean yields for each pair of treatments: compost minus herbicide, straw minus herbicide, and straw minus compost, and compare your answers to those obtained in exercise 10.3
- 12.3 Michael Eysenck (1974)<sup>28</sup> investigated the role of age and type of processing on memory. He used a word memory task. He randomly assigned 50 younger participants and 50 older participants to one of five learning groups.

Here are the tasks that each of the five learning groups were asked to do while reading a list of 27 words:

---

Counting	Count the number of letters in each word (lowest level of processing).
Rhyming	Think of a rhyme for each word.
Adjective	Think of an adjective to modify each word.
Imagery	Think of a visual image for each word.
Intentional	Try to memorize each word.

---

Each participant read the word list three times and then was asked to recall as many words as possible. Eysenck recorded the number of correct words recalled.

The data file is `Eysenck_1974.mwx`.

The data are as follows:

Age group	Counting	Rhyming	Adjective	Imagery	Intentional
Younger	8	10	14	20	21
Younger	6	7	11	16	19
Younger	4	8	18	16	17
Younger	6	10	14	15	15
Younger	7	4	13	18	22
Younger	6	7	22	16	16
Younger	5	10	17	20	22
Younger	7	6	16	22	22
Younger	9	7	12	14	18
Younger	7	7	11	19	21
Older	9	7	11	12	10
Older	8	9	13	11	19
Older	6	6	8	16	14
Older	8	6	6	11	5
Older	10	6	14	9	10
Older	4	11	11	23	11
Older	6	6	13	12	14
Older	5	3	13	10	15
Older	7	8	10	19	11
Older	7	7	11	11	11

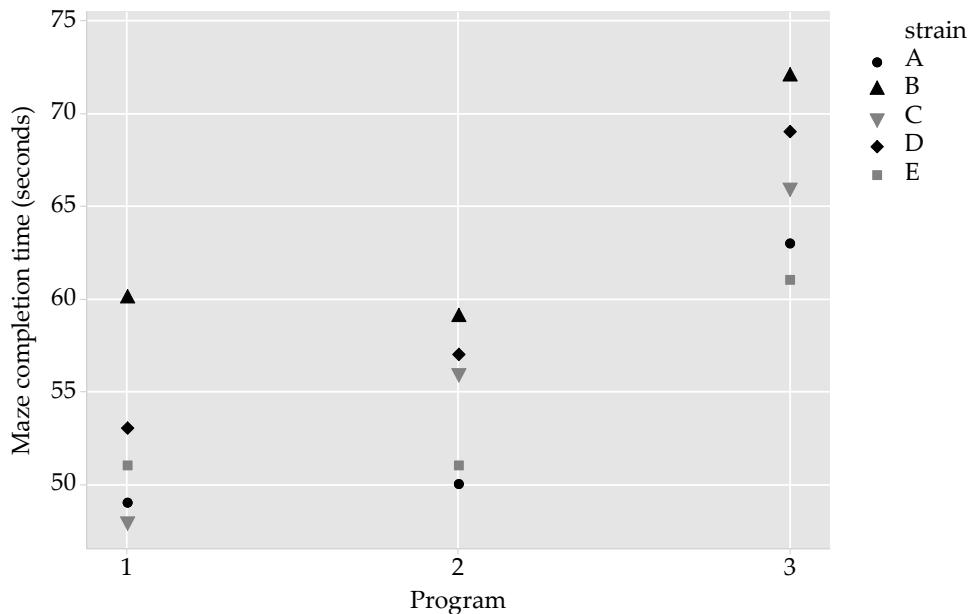
- (a) Obtain an appropriate visual display of the data.

<sup>28</sup>Eysenck, M.W. (1974). Age differences in incidental learning. *Developmental Psychology*, 10, 936-941.

- (b) Do you expect to find an interaction of age group and type of task? Explain why or why not.
- (c) Assuming a model with Normally distributed errors, carry out an analysis of the experiment. [ To include an interaction term in your model use Age Group, Process and Age Group\*Process in the Model terms. Remember to obtain the least squares means. ]
- (d) Write up an appropriate statistical summary of the results.
- (e) Describe the findings in a way that a lay-person could understand.

## 12.4 Answers

- 12.1 (a) The plot shows the mean scores of each strain and each program.



From the plot it appears that program 3 produces higher scores than the other programs, that there are differences between strains, and that the assumption of additivity is reasonable. We cannot assess the interaction between programs and strains because we have only one replicate for each treatment combination.

Here is some relevant output:

MAZE.MWX																																									
<b>General Linear Model: Maze completion time (seconds) versus Program, Strain</b>																																									
<b>Method</b>																																									
Factor (-1, 0, +1) coding																																									
<b>Factor Information</b>																																									
<table border="1"> <thead> <tr> <th>Factor</th> <th>Type</th> <th>Levels</th> <th>Values</th> </tr> </thead> <tbody> <tr> <td>Program</td> <td>Fixed</td> <td>3</td> <td>1, 2, 3</td> </tr> <tr> <td>Strain</td> <td>Fixed</td> <td>5</td> <td>A, B, C, D, E</td> </tr> </tbody> </table>		Factor	Type	Levels	Values	Program	Fixed	3	1, 2, 3	Strain	Fixed	5	A, B, C, D, E																												
Factor	Type	Levels	Values																																						
Program	Fixed	3	1, 2, 3																																						
Strain	Fixed	5	A, B, C, D, E																																						
<b>Analysis of Variance</b>																																									
<table border="1"> <thead> <tr> <th>Source</th> <th>DF</th> <th>Seq SS</th> <th>Contribution</th> <th>Adj SS</th> <th>Adj MS</th> <th>F-Value</th> <th>P-Value</th> </tr> </thead> <tbody> <tr> <td>Program</td> <td>2</td> <td>560.53</td> <td>70.83%</td> <td>560.53</td> <td>280.267</td> <td>65.69</td> <td>0.000</td> </tr> <tr> <td>Strain</td> <td>4</td> <td>196.67</td> <td>24.85%</td> <td>196.67</td> <td>49.167</td> <td>11.52</td> <td>0.002</td> </tr> <tr> <td>Error</td> <td>8</td> <td>34.13</td> <td>4.31%</td> <td>34.13</td> <td>4.267</td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td>14</td> <td>791.33</td> <td>100.00%</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>		Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value	Program	2	560.53	70.83%	560.53	280.267	65.69	0.000	Strain	4	196.67	24.85%	196.67	49.167	11.52	0.002	Error	8	34.13	4.31%	34.13	4.267			Total	14	791.33	100.00%				
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value																																		
Program	2	560.53	70.83%	560.53	280.267	65.69	0.000																																		
Strain	4	196.67	24.85%	196.67	49.167	11.52	0.002																																		
Error	8	34.13	4.31%	34.13	4.267																																				
Total	14	791.33	100.00%																																						
<b>Model Summary</b>																																									
<table border="1"> <thead> <tr> <th>S</th> <th>R-sq</th> <th>R-sq(adj)</th> <th>PRESS</th> <th>R-sq(pred)</th> <th>AICc</th> <th>BIC</th> </tr> </thead> <tbody> <tr> <td>2.06559</td> <td>95.69%</td> <td>92.45%</td> <td>120</td> <td>84.84%</td> <td>94.90</td> <td>76.57</td> </tr> </tbody> </table>		S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC	2.06559	95.69%	92.45%	120	84.84%	94.90	76.57																										
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC																																			
2.06559	95.69%	92.45%	120	84.84%	94.90	76.57																																			

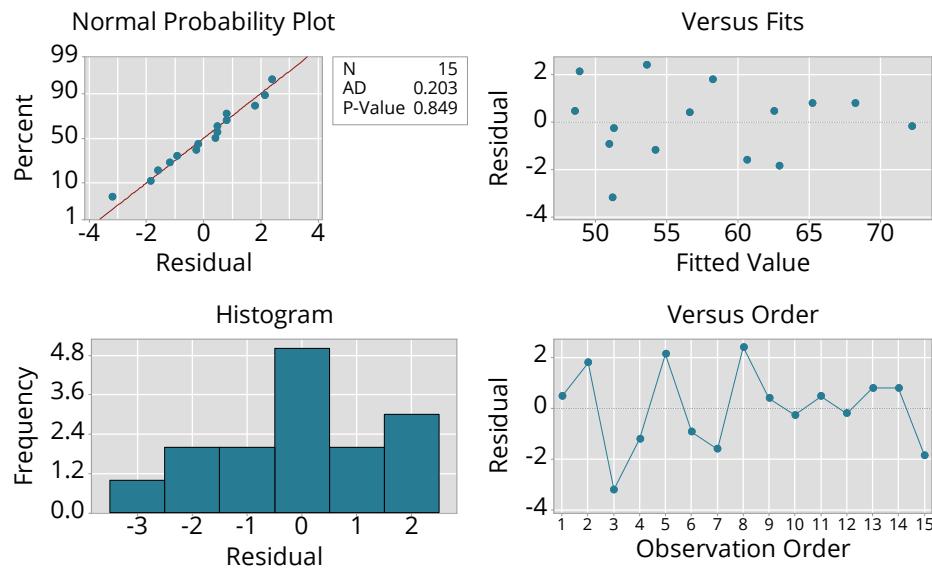
To obtain the least squares means using the General linear model menu, click on the Options, and under Means, choose All terms in the model.

### Means

Term	Fitted Mean	SE Mean
<b>Program</b>		
1	52.200	0.924
2	54.600	0.924
3	66.200	0.924
<b>Strain</b>		
A	54.00	1.19
B	63.67	1.19
C	56.67	1.19
D	59.67	1.19
E	54.33	1.19

Before interpreting the analysis of variance, we need to check if the model assumptions are sound by looking at the distribution of the residuals. A Normal probability plot of the standardised residuals is shown below. There is no evidence to suggest that the normality assumption is not met. However, with a sample size this small, we may not be able to pick up departures from normality.

Residual Plots for Maze completion time (seconds)



Now consider the comparisons between the pairs of different programs; here are the multiple comparisons from Tukey's HSD method:

## Comparisons for Maze completion time (seconds)

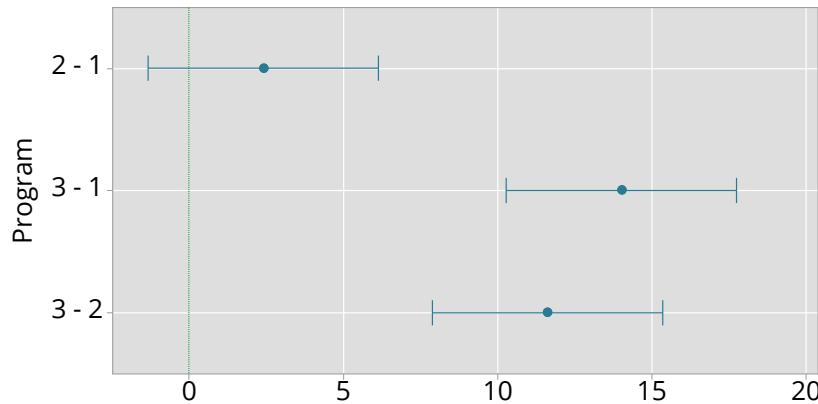
### Tukey Pairwise Comparisons: Program

### Tukey Simultaneous Tests for Differences of Means

Difference of Program Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
2 - 1	2.40	1.31	(-1.33, 6.13)	1.84	0.219
3 - 1	14.00	1.31	(10.27, 17.73)	10.72	0.000
3 - 2	11.60	1.31	(7.87, 15.33)	8.88	0.000

*Individual confidence level = 97.87%*

**Tukey Simultaneous 95% CIs  
Differences of Means for Maze completion time (seconds)**



From the analysis, the mean differences between programs and the mean differences between strains that are not consistent with null hypotheses of no mean differences (programs: ( $F_{2,8} = 65.7, P < 0.001$ ), strains: ( $F_{4,8} = 11.5, P = 0.002$ )).

The difference in mean time to complete the maze is less than 2.4 seconds for programs 2 and 1 (95% CI for program 2 – program 1:  $-1.3, 6.1$ ); the mean completion time is substantially longer for program 3: 11.6 seconds slower than program 2 (95% CI for program 3 – program 2:  $7.9, 15.3$ ) and 14.0 seconds longer than program 1 (95% CI for program 3 – program 1:  $10.3, 17.7$ ).

A summary table giving descriptive statistics, the analysis of variance results and the pairwise comparisons appears below.

Program	Maze completion times in seconds		
	Mean	Standard deviation	n
1	52.2	4.8	5
2	54.6	3.9	5
3	66.2	4.4	5

**ANOVA**For program:  $F_{(2,8)} = 65.7, P < 0.001$ ; for strain:  $F_{(4,8)} = 11.5, P = 0.002$ .

Comparison of programs	Estimate	Difference in means	
		95% confidence interval*	P-value
2 – 1	2.4	-1.3, 6.1	0.219
3 – 1	14.0	10.3, 17.7	<0.001
3 – 2	11.6	7.9, 15.3	<0.001

\*Estimated using Tukey's HSD

- (b) We are primarily interested in differences between the three conditioning programs. Blocking by strain should increase the efficiency of the analysis. Hence the summary above focussed on the program differences.
- (c) The analysis, without strain:

MAZE.MWX

**General Linear Model: Maze completion time (seconds) versus Program**

**Method**

Factor (-1, 0, +1)  
coding

**Factor Information**

Factor	Type	Levels	Values
Program	Fixed	3	1, 2, 3

**Analysis of Variance**

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Program	2	560.5	70.83%	560.5	280.27	14.57	0.001
Error	12	230.8	29.17%	230.8	19.23		
Total	14	791.3	100.00%				

**Model Summary**

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
4.38558	70.83%	65.97%	360.625	54.43%	95.57	94.40

A summary table giving descriptive statistics, the analysis of variance results and the pairwise comparisons appears below.

Program	Maze completion times in seconds		
	Mean	Standard deviation	n
1	52.2	4.8	5
2	54.6	3.9	5
3	66.2	4.4	5

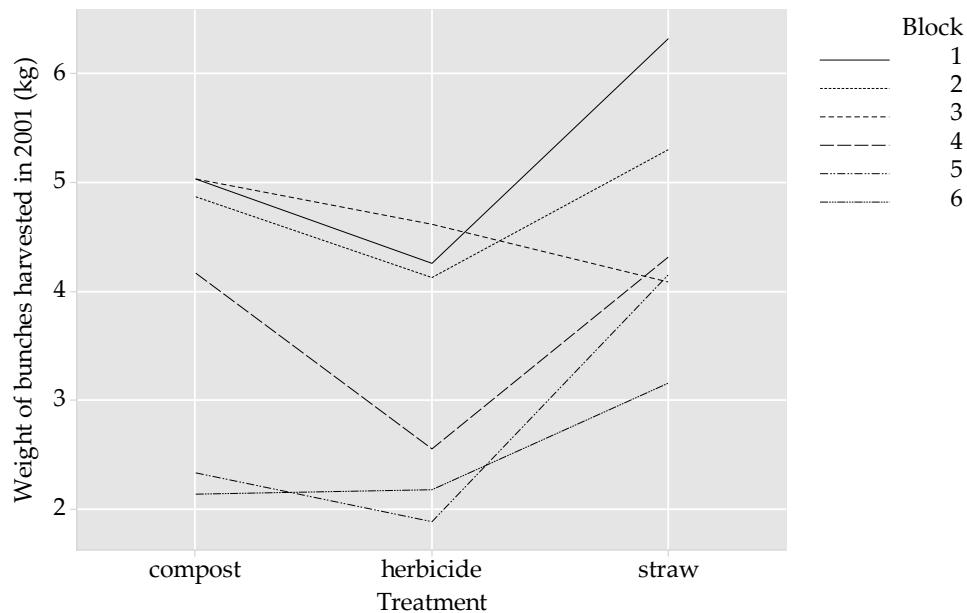
ANOVA for program:  $F_{(2,12)} = 14.6, P < 0.001$ .

Comparison of programs	Estimate	Difference in means	
		95% confidence interval*	P-value
2 – 1	2.4	-5.0, 9.8	0.671
3 – 1	14.0	-6.6, -21.4	0.001
3 – 2	11.6	-4.2, -19.0	0.003

\*Estimated using Tukey's HSD

Compare the summary table above with the one for part (a) to see what is the same and what is different. In the second table, the confidence intervals for differences between the programs are much wider than in the first, reflecting the loss of precision when strain is excluded from the model.

- 12.2 (a) The line plot below helps show points that are quite close together:

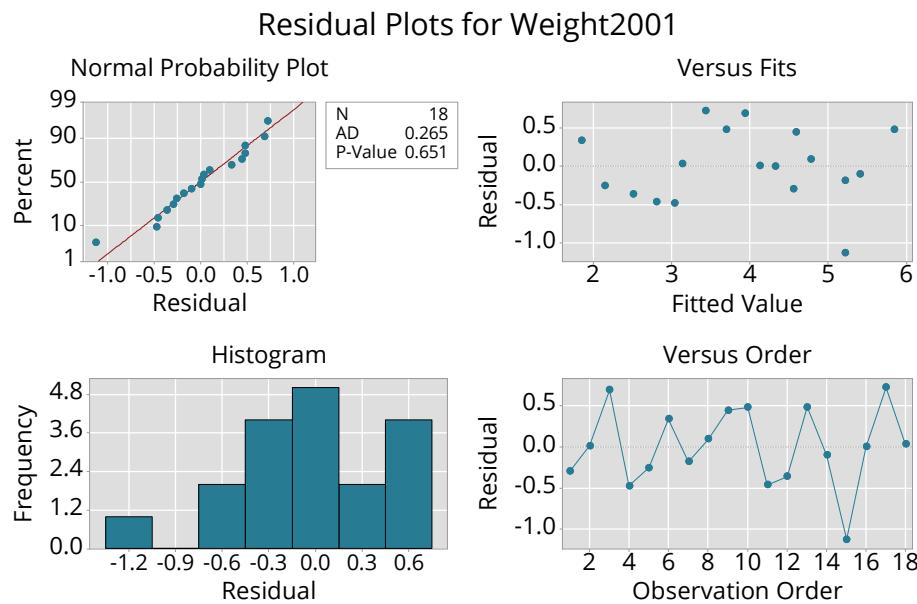


There appear to be block effects, so it is likely that the inference would be more sensitive if block was included as a factor.

- (b) The output is shown below:

PINOT.MWX							
<b>General Linear Model: Weight2001 versus Treatment, Block</b>							
<b>Method</b>							
<b>Factor coding</b> (-1, 0, +1)							
<b>Factor Information</b>							
Factor	Type	Levels	Values				
Treatment	Fixed	3	compost, herbicide, straw				
Block	Fixed	6	1, 2, 3, 4, 5, 6				
<b>Analysis of Variance</b>							
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Treatment	2	4.942	18.15%	4.942	2.4709	6.50	0.016
Block	5	18.480	67.88%	18.480	3.6961	9.72	0.001
Error	10	3.803	13.97%	3.803	0.3803		
Total	17	27.226	100.00%				

The check of the assumptions:



- (c) Pairwise comparisons using Tukey's HSD simultaneous confidence intervals:

### Comparisons for Weight2001

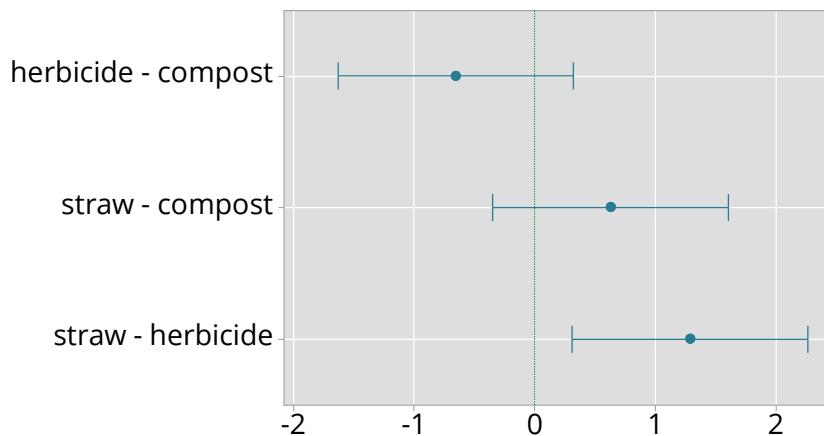
#### Tukey Pairwise Comparisons: Treatment

#### Tukey Simultaneous Tests for Differences of Means

Difference of Treatment Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
herbicide - compost	-0.657	0.356	(-1.634, 0.320)	-1.84	0.205
straw - compost	0.627	0.356	(-0.350, 1.604)	1.76	0.232
straw - herbicide	1.283	0.356	(0.306, 2.260)	3.60	0.012

Individual confidence level = 97.93%

Tukey Simultaneous 95% CIs  
Differences of Means for Weight2001



A suitable summary table for the analysis:

Program	Weight of bunches harvest in 2001 (kg)		
	Mean	Standard deviation	n
Herbicide	3.27	1.19	6
Compost	3.93	1.35	6
Straw	4.56	1.10	6

ANOVA  
For treatment:  $F_{(2,10)} = 6.49, P = 0.016$ ; for block:  $F_{(5,10)} = 9.73, P = 0.001$ .

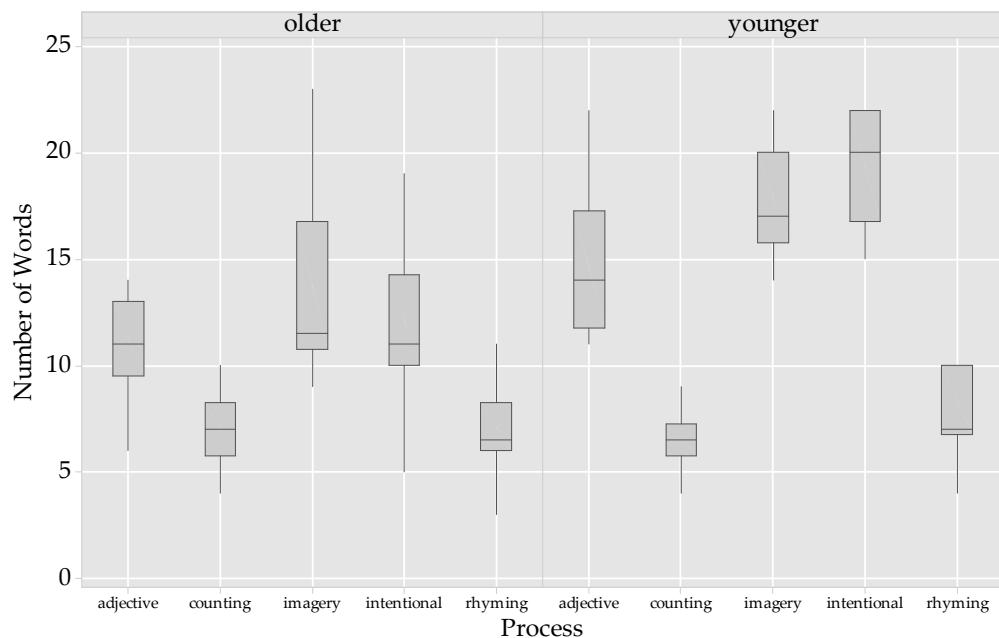
Comparison	Estimate	Difference in means*	
		95% confidence interval	P-value
Compost – Herbicide	0.66	-0.32, 1.63	0.205
Straw – Herbicide	1.28	0.31, 2.26	0.012
Straw – Compost	0.63	-0.35, 1.60	0.232

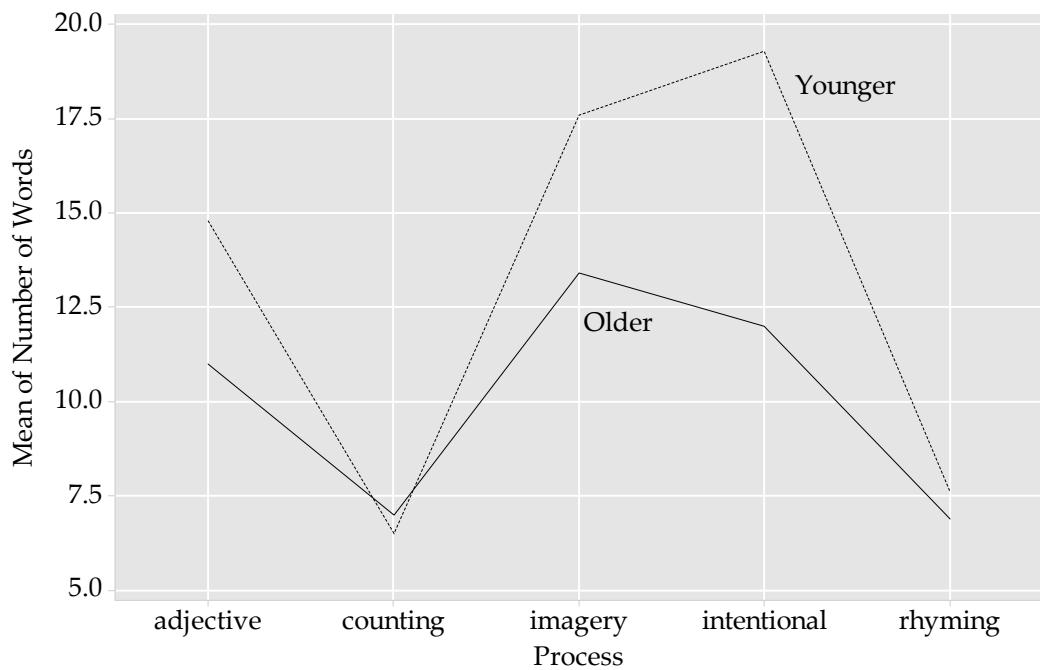
\*Based on Tukey's HSD

The estimates for the mean differences are consistent with those obtained in exercise 10.3; the confidence intervals are narrower, as is expected. There is improved precision because of the inclusion of the block factor.

Note that the estimates are presented in the table so that the mean differences are positive.

### 12.3 (a) Here are two useful graphs.





The pattern is different according to age.

- (b) In the figures above, the pattern of means for different levels of processing looks different for the two age groups. For example, mean differences between the Adjective, Imagery and Intentional groups appear to be greater in the younger participants than in the older participants. This appears to be an interaction.
- (c) Here is some relevant output:

#### Descriptive Statistics: Number of Words

##### Results for Age Group = older

##### Statistics

Variable	Process	N	Mean	StDev
Number of Words	adjective	10	11.000	2.494
	counting	10	7.000	1.826
	imagery	10	13.40	4.50
	intentional	10	12.00	3.74
	rhyming	10	6.900	2.132

##### Results for Age Group = younger

##### Statistics

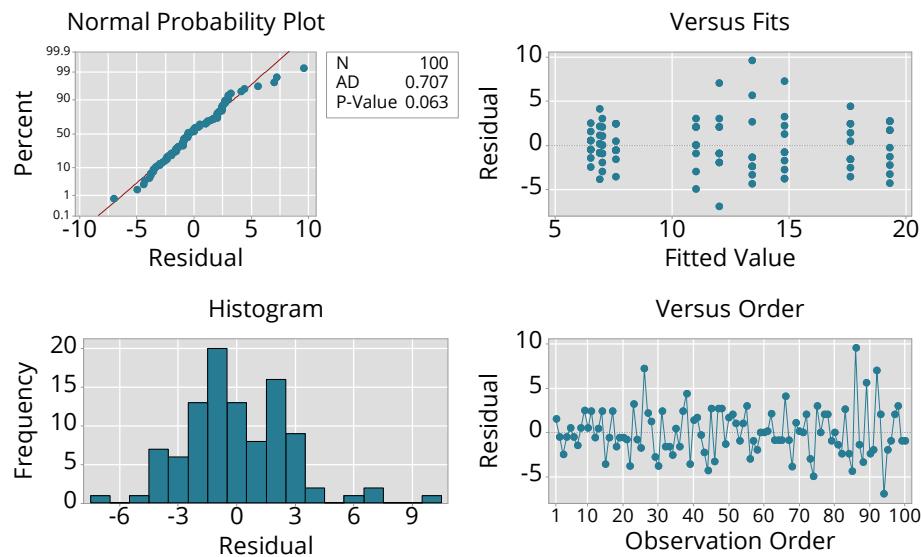
Variable	Process	N	Mean	StDev
Number of Words	adjective	10	14.80	3.49
	counting	10	6.500	1.434
	imagery	10	17.600	2.591
	intentional	10	19.300	2.669
	rhyming	10	7.600	1.955

Output from General Linear Model:

General Linear Model: Number of Words versus Age Group, Type of processing										
Method										
Factor coding (-1, 0, +1)										
Factor Information										
Factor	Type	Levels	Values							
Age Group	Fixed	2 younger, older								
Type of processing	Fixed	5 adjective, counting, imagery, intentional, rhyming								
Analysis of Variance										
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value				
Age Group	1	240.2	9.01%	240.2	240.250	29.94				
Type of processing	4	1514.9	56.79%	1514.9	378.735	47.19				
Age Group*Type of processing	4	190.3	7.13%	190.3	47.575	5.93				
Error	90	722.3	27.07%	722.3	8.026					
Total	99	2667.8	100.00%							

Examining assumptions:

### Residual Plots for Number of Words



Estimates and standard errors:

### Means

Term	Fitted Mean	SE Mean
Age Group		
older	10.060	0.401
younger	13.160	0.401
Process		
adjective	12.900	0.633
counting	6.750	0.633
imagery	15.500	0.633
intentional	15.650	0.633
rhyming	7.250	0.633
Age Group*Process		
older adjective	11.000	0.896
older counting	7.000	0.896
older imagery	13.400	0.896
older intentional	12.000	0.896
older rhyming	6.900	0.896
younger adjective	14.800	0.896
younger counting	6.500	0.896
younger imagery	17.600	0.896
younger intentional	19.300	0.896
younger rhyming	7.600	0.896

- (d) Fifty younger participants and 50 older participants were assigned at random to one of five learning groups. In each learning group, participants were given a particular task to do while reading a list of 27 words. The outcome of interest was the correct number of words recalled. Similar average numbers of words were correctly recalled by both age groups when counting or rhyming tasks were involved. Younger groups recorded higher averages on the other three types of tasks, and appeared to benefit more, on average, than older groups when doing an intentional task of attempting to memorise the words.

A General Linear Model was fitted to the number of words recalled using age group and type of task as explanatory factors. The interaction of age and task suggested that the observed patterns of means was not consistent with a null hypothesis that the mean difference between older and younger people did not vary according to different levels of processing ( $F_{4,90} = 5.9, P < 0.001$ ). A suitable table appears below:

## ANOVA

Age group	$F_{(1,90)} = 29.9, P < 0.001$
Level of processing	$F_{(4,90)} = 47.2, P < 0.001$
Age group by Level of processing	$F_{(4,90)} = 5.9, P < 0.001$

Age group	Level of processing	Estimate	Mean
			95% confidence interval
Younger	Counting	6.5	4.7, 8.3
Younger	Rhyming	7.6	5.8, 9.4
Younger	Adjective	14.8	13.0, 16.6
Younger	Imagery	17.6	15.8, 19.4
Younger	Intentional	19.3	17.5, 21.1
Older	Counting	7.0	5.2, 8.8
Older	Rhyming	6.9	5.1, 8.7
Older	Adjective	11.0	9.2, 12.8
Older	Imagery	13.4	11.6, 15.2
Older	Intentional	12.0	10.2, 13.8

The mean number of words correctly recalled when counting or rhyming is around 7. Average recall on the other three tasks for older participants is between 11.0 and 13.6. Younger participants do better on the other three tasks, with means ranging from 14.8 to 19.3. The 95% confidence interval for each age by task group in the study appears in the table above. The confidence intervals were derived from estimates and standard errors in the output above.

- (e) We summarise the recall in each group with the average number of words recalled. To describe the patterns of responses to the word recall task, we need to consider both the age group and task involved together. For example, age does not appear to have a strong impact on the average number of words recalled when counting or rhyming tasks are used; in these cases, around 7 words on average are recalled. For the other three tasks, the younger age group does better than the older age group, on average, but by varying amounts. For tasks involving adjectives or imagery, the mean difference is about 4 words, but on the intentional task, younger participants recall over 7 words more on average than older participants.

# 13 Inference — numerical outcome and numerical explanatory variables

In this chapter we consider another particular case of the linear model, which is usually known as **linear regression**.

The use of this word is not helpful at all, really. It arises from Galton's analysis of parents' heights and the height of their children, and the tendency of parents of extreme height (very short or very tall) to have children with heights that are less extreme. This phenomenon is known as "regression to the mean". When Galton was looking at these data, he used the model-fitting algorithm that is used here, which is how the name "regression" came to be used for such models. The term is very widely used in statistics, applying to many contexts that include the linear model and go well beyond it. Hence there is "linear regression", "logistic regression", "Poisson regression", "ordinal logistic regression", "Cox's proportional hazards regression" and so on.

In this chapter we are concerned with linear regression, which — in the standard case — involves a numerical outcome and one or more numerical explanatory variables.

## 13.1 Simple linear regression: one predictor

In "simple" linear regression there is one numerical explanatory variable. In this case we can think about the data and the problem graphically.

We introduce the idea with an example.

▷ **EXAMPLE. Evaporation coefficient of burning fuel** (evap.mwx)

The following data were obtained on the air velocity ( $\text{cm s}^{-1}$ ) and evaporation coefficient ( $\text{mm}^2 \text{s}^{-1}$ ) of burning fuel droplets in an impulse engine.

Air velocity	20	60	100	140	180	220	260	300	340	380
Evaporation coefficient	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

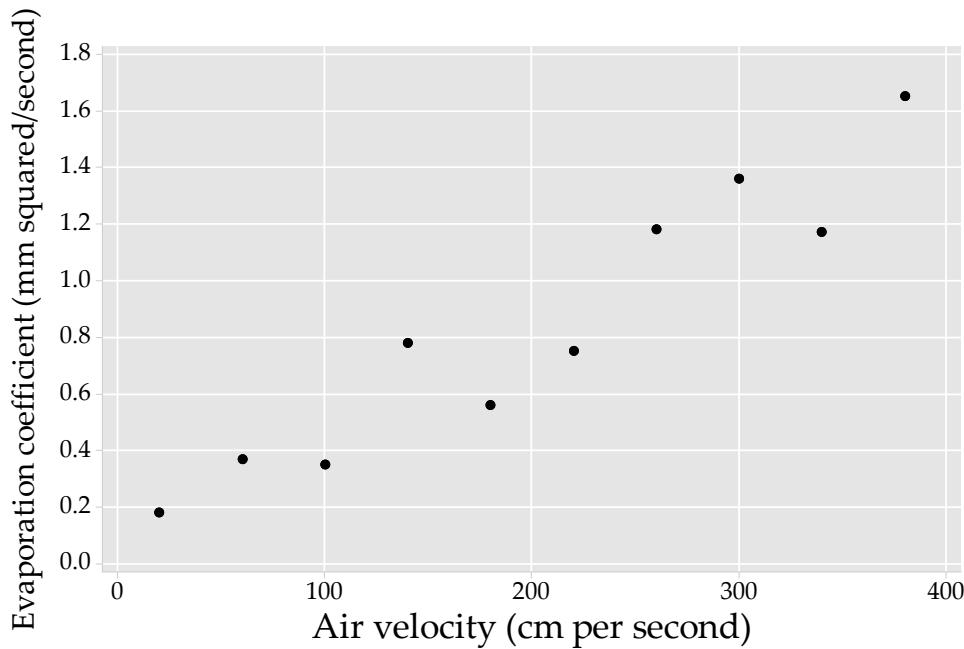


Figure 91: Plot of evaporation coefficient vs air velocity

We usually use  $Y$  to denote the response or outcome variable, and  $x$  to denote the explanatory variable.

One way to summarize the relationship between  $Y$  and  $x$ , and to predict  $Y$  from  $x$  is to fit a straight line to the data. We think of  $Y$  as a random variable and  $x$  as fixed or known, sometimes actually determined. The observed data consist of pairs  $(x_i, y_i), i = 1, 2, \dots, n$ , and we assume that the pairs are independent of each other.

We assume a linear model of the form

$$Y_i = \alpha + \beta x_i + E_i$$

where the  $E_i$ s are random errors. These random errors are assumed to be independent and Normally distributed with mean zero and constant variance,  $\sigma^2$ .

This model says that there is an underlying linear relationship between  $x$  and  $Y$ ; in individual pairs there are deviations following a Normal distribution.

Obviously, this is just another special case of the linear model. One important difference is that a numerical explanatory variable  $x$  may be able to take any possible value within a range, so that we are not restricted to the values observed in the data. When we have a categorical explanatory variable, such as ‘treatment’, that takes the values ‘active treatment’ and ‘placebo’, for example, we cannot make predictions for yet another treatment level, not used in the experiment (‘different active treatment’). For a numerical

explanatory variable, on the other hand, the information from the data and the fitted model may allow us to make predictions about the response variable,  $Y$ , for values of  $x$  not observed in the data, as we shall see.

### 13.1.1 Estimation — the method of least squares

This model is a clear instance in which the parameter estimates are not obvious. When we wanted to estimate  $\mu$  from a single random sample, we regarded it as more or less self-evident that the sample mean,  $\bar{x}$ , should be used as the estimate. We estimated population proportions using sample proportions, and so on. Here, however, to estimate the parameters of this model, we need to fit a line to the data. How should this be done? We may feel that we can do a reasonable job by eye, and we probably could. But this would not be satisfactorily objective and reproducible. We need a systematic approach that will lead to good estimates of  $\alpha$  and  $\beta$ .

The parameters  $\alpha$  and  $\beta$  are (usually) estimated by the method of least squares which amounts to finding the values  $\hat{\alpha}$  and  $\hat{\beta}$  such that the sum of the squared (vertical) deviations from the line is as small as possible. That is,  $\hat{\alpha}$  and  $\hat{\beta}$  are values for  $\alpha$  and  $\beta$  such that

$$\sum (y_i - \alpha - \beta x_i)^2 \text{ is a minimum.}$$

This gives

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

It is also necessary to estimate  $\sigma^2$ , the variance of the random errors, or deviations from the model, and

$$s^2 \text{ (or } \hat{\sigma}^2) = \frac{\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n - 2}$$

is an unbiased estimate of  $\sigma^2$ .

In MINITAB: Stat > Regression ▶ Regression.

As an aside: there is a software design issue here. MINITAB now has two modules that fit a “general linear model”: a model with categorical explanatory variables (“factors”) and numerical explanatory variables (“covariates”). The first is General Linear Model. This has the orientation of analysis of variance, to which covariates may be added. The second is Regression, which has the orientation of a regression, to which factors may be added. Both modules fit the same fundamental model, but because of the way they have been written, different information and options are provided. We are using the Regression module because it has some useful features in this context.

The full extent of the linear model is an extension of both analysis of vari-

ance and regression, which is beyond the scope of SRW.

▷ **EXAMPLE. Evaporation coefficient of burning fuel** (*continued*)

$$n = 10; \hat{\alpha} = 0.069; \hat{\beta} = 0.00383; s = 0.159.$$

### 13.1.2 Confidence intervals

Often, the key question inferentially relates to  $\beta$ ; we want to know how large  $\beta$  is, so we would like an estimate and confidence interval for  $\beta$ , because it captures how the response variable changes per unit increase in  $x$ .

We might also be interested in making an inference about the mean value of  $Y$  for a given  $x$ , i.e.  $\alpha + \beta x$ .

If the errors are Normally distributed, it turns out that the estimators  $\hat{\alpha}$  and  $\hat{\beta}$  have Normal distributions themselves:

$$\begin{aligned}\hat{\alpha} &\stackrel{d}{=} N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\right), \\ \hat{\beta} &\stackrel{d}{=} N\left(\beta, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)\end{aligned}$$

and the distribution of  $S^2$  is given by

$$\frac{(n - 2)S^2}{\sigma^2} \stackrel{d}{=} \chi_{(n-2)}^2.$$

This allows us to make inferences about  $\alpha$ ,  $\beta$  and  $\alpha + \beta x_0$ , the mean value of  $Y$  when  $x = x_0$ . Quite often  $\alpha$  has no special importance in its own right, being merely the mean value of  $Y$  when  $x = 0$ .

A 95% confidence interval for  $\beta$  is given by:

$$\hat{\beta} \pm t_{(n-2)}(0.975) \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}.$$

The slope  $\beta$  is estimated more efficiently when the  $x$ -values are well-spaced out; in fact, if you are sure that the linear model is correct, and you have control over the  $x$  values, the optimal design is to place half of the points at the smallest possible  $x$  value and half at the largest possible  $x$  value.

▷ **QUESTION:** Why would you choose *not* to use this “optimal” strategy?

A 95% confidence interval for the mean value of  $Y$  when  $x$  takes the particular value  $x_0$ , i.e.  $\alpha + \beta x_0$ , is given by:

$$(\hat{\alpha} + \hat{\beta}x_0) \pm t_{(n-2)}(0.975) \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}.$$

The confidence interval gets wider the further that  $x_0$  is away from  $\bar{x}$ . This interval is a confidence interval in the usual sense; it is an inference for an unknown parameter. So it can be made narrow by increasing the sample size. It expresses our uncertainty in the *average* value of  $Y$  for the specific value  $x_0$ .

Substituting  $x_0=0$  into this formula gives the 95% confidence interval for  $\alpha$ :

$$\hat{\alpha} \pm t_{(n-2)}(0.975) \sqrt{s^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)}$$

The regression line should not be used to make predictions for  $x$  outside the range of values in the observed data. Almost always, the assumed linear model is only a convenient approximation, and therefore we may have little basis for believing that the relationship holds for  $x$ -values more distant.

Nor should the regression of  $y$  on  $x$  be used to predict  $x$  from  $y$ ; regression is not a symmetric relation between two variables and so it is essential to have the appropriate regression line for the proposed prediction.

To obtain the confidence intervals for the regression parameters in MINITAB, use Stat > Regression ▶ Regression ▶ Fit Regression Model and click on the Results button. In the drop down box labelled Display of results, choose Expanded tables.

If the distribution of the errors is not Normal then the above results can often still be used as a good approximation, due to the Central Limit Theorem.

### 13.1.3 Prediction intervals

A prediction interval is an interval within which we predict a *single* future value of  $Y$  will lie, for a given value of  $x$ . Since  $\text{var}(Y|x) = \sigma^2$ , we cannot make a prediction interval as narrow as we please: as the sample size increases, the width of a confidence interval tends to zero whereas the width of a prediction interval tends to  $2 \times 1.96\sigma$ .

A 95% prediction interval for  $Y$  given  $x = x_0$ :

$$(\hat{\alpha} + \hat{\beta}x_0) \pm t_{(n-2)}(0.975) \sqrt{s^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right)}$$

To obtain the confidence intervals and prediction intervals in MINITAB, for given values of  $x$ : *after you have fitted the regression model*, use Stat > Regression ▶ Regression ▶ Predict.

### 13.1.4 Hypothesis testing

As usual, there are hypothesis tests corresponding to the confidence intervals found earlier. In hypothesis testing terms there is particular interest in the null hypothesis  $H_0 : \beta = 0$ , since if  $\beta = 0$  then in terms of a linear relationship,  $Y$  is constant as  $x$  changes.

#### t-test:

As mentioned in Section 13.1.2, if the random errors can be assumed to be Normally distributed, then  $\hat{\beta}$  is Normally distributed. We can use this result to carry out hypothesis testing on  $\beta$ . In particular, we can test if  $\beta$  is zero:

$$\frac{\hat{\beta}}{\text{se}(\hat{\beta})} \stackrel{d}{=} t_{(n-2)}, \quad \text{if } \beta \text{ is zero.}$$

#### ANOVA:

Source	df	SS	MS	F
Regression	1	regression SS	regression MS = regression SS/1	regression MS/residual MS
Residual	$n - 2$	residual SS	residual MS = residual SS/( $n - 2$ )	
Total	$n - 1$	Total SS		

A 2-sided test using the t-test above is equivalent to the F-test in this ANOVA.

#### ▷ EXAMPLE. Evaporation coefficient of burning fuel (continued)

MINITAB output:

#### Regression Analysis: Evaporation coefficient versus Air ... per second) Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1.9351	90.53%	1.9351	1.93507	76.49	0.000
Air velocity (cm per second)	1	1.9351	90.53%	1.9351	1.93507	76.49	0.000
Error	8	0.2024	9.47%	0.2024	0.02530		
Total	9	2.1374	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0.159052	90.53%	89.35%	0.304162	85.77%

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value
Constant	0.069	0.101	(-0.164, 0.302)	0.69	0.512
Air velocity (cm per second)	0.003829	0.000438	(0.002819, 0.004838)	8.75	0.000
Term	VIF				
Constant					
Air velocity (cm per second)	1.00				

#### Regression Equation

$$\text{Evaporation coefficient} = 0.069 + 0.003829 \text{ Air velocity (cm per second)}$$

After fitting the regression, if we wanted to predict the response  $Y$  when  $x = 200$ , we use Stat ▶ Regression ▶ Regression ▶ Predict ... to get the following:

### Prediction for Evaporation coefficient

#### Regression Equation

Evaporation coefficient =  $0.069 + 0.003829$  Air velocity (cm per second)

#### Settings

Variable	Setting
Air velocity (cm per second)	200

#### Prediction

Fit	SE Fit	95% CI	95% PI
0.835000	0.0502967	(0.719016, 0.950984)	(0.450323, 1.21968)

Note the following in the MINITAB output:

- The estimated line is given.
- The estimates and standard errors of  $\alpha$  and  $\beta$  are given, and the  $t$ -test and  $P$ -value are provided. Confidence intervals for  $\alpha$  and  $\beta$  are provided as part of the output but as outlined above *you need to ask for them under Results*; they are not provided by default.
- The estimate of  $\sigma$ , the standard deviation of the random errors, is given as  $S$ ; here  $s = 0.159$ .
- R-sq =  $R^2$  is explained below.
- In this case, a confidence interval and a prediction interval were requested for  $x = 200$ . The estimated value of  $Y$  when  $x = 200$  is 0.8350, the 95% confidence interval for the mean value of  $Y$  when  $x = 200$ , namely,  $\alpha + \beta \times 200$ , is (0.719, 0.951), and a 95% prediction interval for a future observation of  $Y$  when  $x = 200$  is (0.450, 1.220).

## 13.2 Multiple regression

### 13.2.1 Extension of simple linear regression

In practice, there are often several explanatory variables,  $x_1, x_2, \dots, x_p$ , rather than only one. So we may want to consider how they jointly affect the outcome,  $Y$ . Sometimes  $Y$  is called the “dependent” variable, and the  $xs$  are

called the “independent” variables.<sup>29</sup> Better names are the “response variable” for  $Y$  and the “explanatory variables” for the  $xs$ . These names capture more appropriately the role of the variables in the statistical model.

The linear model we consider is:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + E_i,$$

where  $E_i$  denotes random errors. As usual, the random errors are assumed to be independent and Normally distributed with mean zero and constant variance,  $\sigma^2$ .

The estimates,  $\hat{\beta}_j$ 's, of the coefficients can be obtained by the method of least squares. That is, they are the values for the  $\beta_j$ 's such that

$$\sum (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2$$

is a minimum.

It follows from the assumed Normality of the errors that the distributions of the estimators of the  $\beta_j$ 's are also Normally distributed. Hence it is possible to find confidence intervals and prediction intervals, and to carry out hypothesis testing on the coefficients analogous to the straight line regression case, based on the  $t$ -distribution.

Many new issues arise in multiple regression, or, more generally, in models with several potential explanatory variables. One of the most important is “variable selection”. Suppose that we are considering a regression model and there are 10 possible explanatory variables. Then, considering only models that use these variables (and not derived functions of them as well), the number of possible models is  $2^{10} = 1024$ . That is a lot of models to consider.

There are many statistical approaches to selecting a suitable model and we do not look at these in detail. Usually, some non-statistical considerations are relevant.

### 13.2.2 $R^2$

It is useful to consider how well the regression model performs, in terms of explaining the variation in the response variable,  $Y$ , as a function of the explanatory variables, the  $xs$ .

A measure that is used for this purpose is

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \left( \frac{\text{residual SS}}{\text{total SS}} \right).$$

---

<sup>29</sup>“Independent” is a rather poor choice here, since the  $x$  variables are not, in general, expected to be statistically independent of each other or of  $Y$ . It is used just to mean “not the dependent variable”.

The SS (sums of squares) in these expressions come from the ANOVA. The total SS represents the overall variation in the response variable,  $Y$ . The Residual SS, represents the variation from the fitted line; it is equal to the sum of the squared deviations of the  $y_i$  from  $\hat{\alpha} + \hat{\beta}x_i$ .

$R^2$  lies between 0 and 1 always, but it is usually expressed on a percentage scale, so that when  $R^2 = 0.92$  we say that  $R^2$  is 92%, meaning that 92% of the variation in  $Y$  (the outcome or dependent variable) is explained by the regression. A large  $R^2$  is desirable.

When there is just one explanatory variable,  $x$ , and we have a simple linear regression model,  $r = \pm\sqrt{R^2}$  is in fact the Pearson's correlation between  $Y$  and  $x$ . This gives some tangible meaning to  $r$ . Note that if  $r$  is small,  $R^2$  is even smaller. For example, if  $r = 0.3$  then  $R^2 = 0.09$ , or 9%; if the correlation between  $x$  and  $y$  is as low as 0.3, from a regression point of view,  $x$  only "explains" 9% of the variation in  $y$ .

It might be thought that  $R^2$  is a good criterion to use to determine which of a set of possible models is the most suitable. A rule that comes to mind is to choose the model leading to the largest  $R^2$ . A problem with this, however, is that if an extra explanatory variable is added to an existing model, the value of  $R^2$  is *guaranteed* to increase, regardless of whether the added variable actually has any explanatory power.

To get around this deficiency in  $R^2$ , a related measure called the "adjusted  $R^2$ " is used, which makes an adjustment for the number of predictors in the model.

$$R_{\text{adj}}^2 = 1 - \left( \frac{\text{residual SS}/(n - p - 1)}{\text{total SS}/(n - 1)} \right),$$

where  $p$  is the number of explanatory variables.

The difference between the  $R^2$  and  $R_{\text{adj}}^2$  will be small if the number of observations is much larger than the number of predictors.

$R_{\text{adj}}^2$  is one of several measures of fit that has a built-in penalty for including more variables in the model — or, more generally, model complexity — and is therefore more reasonable to use to select a model statistically. Other measures in the same class are AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion).

### 13.2.3 Adjusting for other predictors

#### ▷ EXAMPLE. Young people and the environment

A study was carried out on Year 10 students to investigate their attitudes towards environmental issues. The MINITAB worksheet `environ.mwx`<sup>30</sup> con-

---

<sup>30</sup>Permission to use the data was kindly granted by Professor David Yencken, Faculty of Architecture, Planning and Building, University of Melbourne.

tains scores relating to three aspects of the study:

- degree of support for an environmental paradigm as opposed to a technological paradigm (**support**);
- level of environmental knowledge (**know**);
- degree of past involvement in improving the environment (**involve**).

For all three aspects above, the higher the score, the higher the degree or level.

Suppose we think of support as the response variable, and we consider the research questions:

- Does know predict support?
- Does involve predict support?
- Do know and involve, together, predict support?

Before fitting models it is a good idea — as always — to plot the data. With only three variables, a matrix plot is feasible.

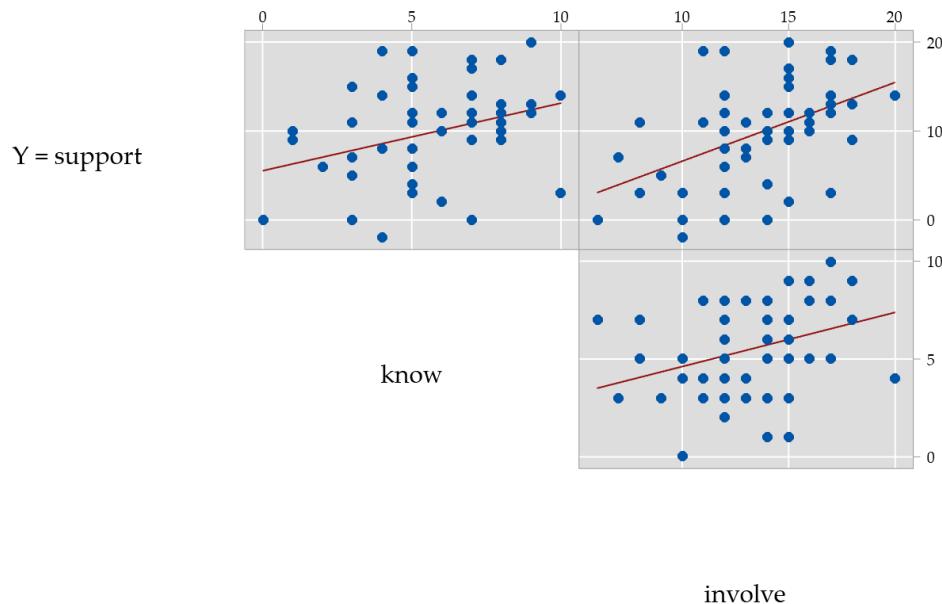


Figure 92: Plot showing the associations between the three variables in the environment data set.

The following is some relevant MINITAB output; use Stat > Regression ► Regression.

## Regression Analysis: support versus know

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	166.6	10.24%	166.6	166.59	5.71	0.021
know	1	166.6	10.24%	166.6	166.59	5.71	0.021
Error	50	1459.5	89.76%	1459.5	29.19		
Lack-of-Fit	9	191.2	11.76%	191.2	21.24	0.69	0.717
Pure Error	41	1268.3	78.00%	1268.3	30.93		
Total	51	1626.1	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
5.40276	10.24%	8.45%	1574.86	3.15%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	5.54	1.94	(1.65, 9.43)	2.86	0.006	
know	0.760	0.318	(0.121, 1.399)	2.39	0.021	1.00

### Regression Equation

$$\text{support} = 5.54 + 0.760 \text{ know}$$

## Regression Analysis: support versus involve

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	378.5	23.28%	378.5	378.51	15.17	0.000
involve	1	378.5	23.28%	378.5	378.51	15.17	0.000
Error	50	1247.6	76.72%	1247.6	24.95		
Lack-of-Fit	12	346.7	21.32%	346.7	28.89	1.22	0.306
Pure Error	38	900.9	55.40%	900.9	23.71		
Total	51	1626.1	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
4.99513	23.28%	21.74%	1339.12	17.65%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2.23	3.17	(-8.59, 4.13)	-0.70	0.485	
involve	0.884	0.227	(0.428, 1.340)	3.89	0.000	1.00

### Regression Equation

$$\text{support} = -2.23 + 0.884 \text{ involve}$$

## Regression Analysis: support versus know, involve

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	418.7	25.75%	418.68	209.34	8.50	0.001
know	1	166.6	10.24%	40.17	40.17	1.63	0.208
involve	1	252.1	15.50%	252.09	252.09	10.23	0.002
Error	49	1207.4	74.25%	1207.40	24.64		
Lack-of-Fit	39	1015.4	62.44%	1015.40	26.04	1.36	0.315
Pure Error	10	192.0	11.81%	192.00	19.20		
Total	51	1626.1	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
4.96395	25.75%	22.72%	1347.54	17.13%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2.96	3.20	(-9.39, 3.47)	-0.93	0.359	
know	0.400	0.313	(-0.230, 1.030)	1.28	0.208	1.15
involve	0.773	0.242	(0.287, 1.259)	3.20	0.002	1.15

### Regression Equation

$$\text{support} = -2.96 + 0.400 \text{ know} + 0.773 \text{ involve}$$

The above output indicates that support is predicted to increase by 0.76 for each unit increase in know, when know is used as a predictor on its own. However, when involve is also taken into account, the predictive role of know changes: now support is predicted to increase by a lesser amount, 0.40, for each unit increase in know. The confidence intervals and *P*-values for know are correspondingly quite different, depending on whether involve is also in the model or not.

In multiple regression, the estimated coefficients are adjusted for all other predictors in the model. Because of the possible associations among the predictor variables, it may not be easy to assign meaning to the coefficients in a multiple regression. It is even possible for a predictor that has a high positive correlation with the response variable to have a negative coefficient in a multiple regression equation.

#### 13.2.4 Checking assumptions: residual analysis

The residuals are given by

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}), \quad i = 1, \dots, n;$$

and the standardized residuals can be used in various ways to check the model:

- Plot of standardized residuals vs. fitted values, and plots of standardized residuals vs. individual predictors.

This is a rough check on constant variance: Serious departure from random scatter appearance implies that the constant variance assumption is invalid.

- Normality plot of standardized residuals: Is there any serious departure from Normality?
- About 95% of the standardized residuals should be between  $-2$  and  $2$ . Note that MINITAB gives messages regarding large standardized residuals.

### 13.2.5 Polynomial regression

Polynomial regression is a special case of multiple regression where some of the explanatory variables are defined as polynomial functions of the basic explanatory variable:  $x_r = x^r$ . That is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + E_i.$$

A particular application of this occurs when there is one predictor,  $x$ , and the plot of  $y$  against  $x$  shows a pattern which appears to be more of a curve than a line. We may fit a model of the form:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

This can be done in MINITAB by first creating a column which contains the square of the values of  $x$ , and then fitting the regression on  $x$  and  $x^2$ .

More generally, other functions of the explanatory variables can be considered, in principle. The model is said to be “linear” not because of the relationship between  $Y$  and the  $xs$ , but because of the linear way the unknown parameters are related to the response variable,  $Y$ .

### 13.3 Exercises

- 13.1 A data set containing 6 columns of data was created by an English statistician Frank Anscombe. The scatterplots arising from these data are sometimes called the “Anscombe quartet”. The data are stored as `anscombe.mwx`.

x1	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

- (a) Carry out 4 simple linear regressions:  $y_1$  on  $x_1$ ,  $y_2$  on  $x_1$ ,  $y_3$  on  $x_1$  and  $y_4$  on  $x_4$ . What do you notice about the results?

[ To regress  $y_1$  on  $x_1$  use Stat > Regression ▶ Regression ▶ Fit regression model ... select C2 as the Response; select C1 as the Continuous predictor; click **OK**. ]

- (b) Now look at the four scatterplots of the data with the corresponding fitted line. Anscombe concocted these data to make a point. What was the point?

[ Use Graph > Scatterplot > With regression, enter  $y_1$  and  $x_1$  and so on to  $y_4$  and  $x_4$ . Click Multiple Graphs... and then choose In separate panels of the same graph and tick each of Same Y and Same X under Same Scales for Graphs.]

- (c) What are the observed and predicted values at  $x_4 = 19$ ? Change the  $y_4$  value for this datum to 10 and refit the regression. What are the observed and predicted values at  $x_4 = 19$  now? [ Repeat the regression of  $y_4$  on  $x_4$  using Stat > Regression > Regression; select c6 as the Response; select c5 as the Continuous Predictor; click Fits under Storage to store the fitted (or predicted) values; click OK.

Now change the value of  $y_4$  from 12.5 to 10 by clicking on the cell that you want to change in the Data Window, then typing in the new value. Repeat the regression. ]

- (d) Looking at the plots, what would you conclude about the appropriateness of simple linear regression in each case?

- 13.2 The Sydney to Hobart yacht race is one of the world’s great ocean racing classics. It has been held every year since 1945. This problem looks at the winning times in this race: the data file: `Sydney_Hobart_2019.mwx`

contains the name of the winning boat (line honours), the year and the winning time in days, for the years 1945 to 2019.

- (a) Plot the winning times against year. Add gridlines to enhance the graph.
- (b) Describe the general pattern shown in the plot. What is the correlation between the winning time and year?
- (c) What conceptual issues arise in the fitting of a regression model to these data?
- (d) Fit a regression model. [Use Stat > Regression ► Regression ► Fit Regression Model.]
- (e) On the graph, add the fitted regression line.  
[In the graph window, right click and select Add > Regression fit.]
- (f) On average, by how much have the winning times decreased each year, according to the straight line regression? Over 74 years?
- (g) Find a 95% confidence interval for the average annual reduction in the winning time, in hours, according to the regression.
- (h) The data file has values for 2020 and 2054 for the explanatory variable Year, but the winning times are not included. Find the predicted values and 95% prediction intervals for these years.  
[After you carry out the regression, Use Stat > Regression ► Regression ► Predict and enter 2020 and 2054 for the Year, then click OK.]
  - (i) What is the predicted winning time in 2020, using these data, and the 95% prediction interval for the winning time?
  - (ii) What is the predicted winning time in 2054 (in hours), using these data?

Comment on these predictions.

- (i) The decrease in winning times does look quite linear, for the data so far.

One way to see what evidence there is of non-linearity is to add a loess curve to the data.

[In the graph window, right click and select Add > Smoother and click OK.]

### 13.3 You explored the cheese data set in Chapter 2; here you carry out statistical inference with those data.

“Taste” is the response variable, and the explanatory variables are the transformed concentrations of acetic acid and hydrogen sulfide ( $H_2S$ ) and the untransformed concentration of lactic acid. The data are stored as `cheese.mwx`.

- (a) Fit a separate regression of taste on each of the three explanatory variables, and state your conclusions.

- (b) Fit a multiple regression of taste on the three explanatory variables, and state your conclusions. In particular, are all three explanatory variables needed? If not, determine the “best” model.
- (c) Look again at the correlations between the four variables [exercises in chapter 2] and use them to interpret your findings in (a) and (b).

13.4 Consider data from smokers in the National Health and Nutrition Examination Survey. The MINITAB worksheet is NHANES.mwx.

Consider predicting the serum cholesterol of smokers, in mg/100 ml, measured in 1971, from a number of variables: age at which started smoking, weight, cigarettes per day and years of smoking. Find the potential explanatory variables in the data set.

- (a) Which of the explanatory variables do you expect to be associated with serum cholesterol levels? What kind of association do you expect?
- (b) Examine the variables cigarettes per day and years of smoking in the data set. Do you expect these variables to be associated? Investigate the relationship with a scatterplot.
- (c) Produce a single plot showing the relationship between the outcome and each of the explanatory variables. [ Graph > Matrix Plot, choose Each Y versus each X, Simple. Specify cholesterol as the Y variable, and your chosen explanatory variables as the X variables. ]
- (d) Consider a simple linear regression predicting serum cholesterol in smokers from years of smoking. What is the value of the regression coefficient for years of smoking? Explain the regression coefficient.
- (e) Create a new variable which converts years of smoking into decades; this will be years of smoking divided by 10.  
[ Calc > Calculator; give a name to the new variable in the box Store result in variable. You could use, for example, ‘Decades of smoking’. The Expression should be ‘Years of smoking’/10.]
- (f) Carry out a simple linear regression predicting serum cholesterol in smokers from decades of smoking. What is the value of the regression coefficient for decade of smoking? Explain this regression coefficient.
- (g) Carry out simple linear regressions predicting serum cholesterol for each of the following variables separately: age started smoking, weight in 1971 and cigarettes per day. Enter appropriate results into the table below to summarise the results of the simple linear regression models you have fitted.

Regression models of Serum cholesterol level (mg/100 ml) in smokers

<i>Explanatory variable</i>	<i>Regression coefficient from Simple linear regression</i>			<i>Regression coefficient from Multiple linear regression</i>		
	Estimate	95% CI	P-value	Estimate	95% CI	P-value
Age started smoking						
Weight in kilograms						
Cigarettes per day						
Decades of smoking						

- (h) Fit a multiple linear regression predicting serum cholesterol for smokers from age started smoking, weight in 1971, cigarettes per day and time smoking in decades. Add the results to the table above. Comment on the differences between the simple linear regressions and the multiple linear regression.
- (i) Consider the residual plot from the multiple regression analysis above. Do you have any concerns about the assumptions of the model?

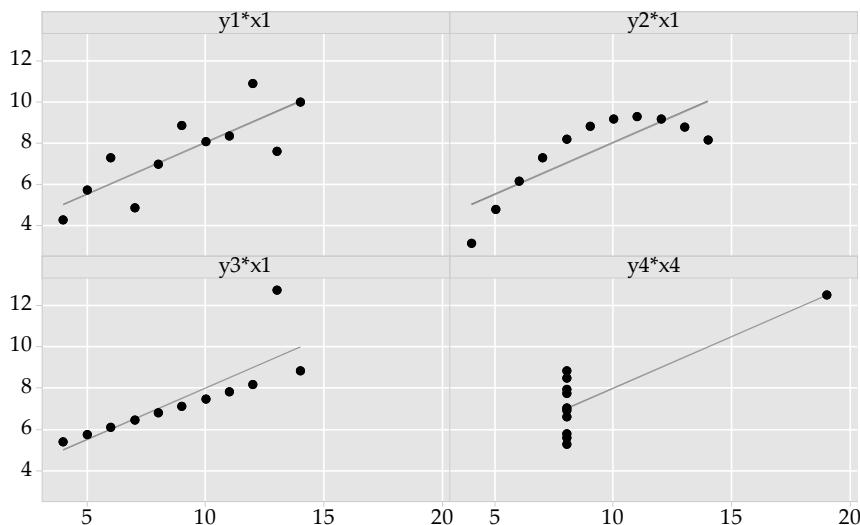
## 13.4 Answers

- 13.1 (a) The regression equations are all the same (and the  $R^2$  values are all the same). The following is the regression of  $y_1$  on  $x_1$ .

The regression equation is  $y_1 = 3.00 + 0.500x_1$

- (b) The point that Anscombe wanted to make was that it is important to examine the scatterplots before calculating regression lines and correlations. By looking at just the regression analyses, we would not have seen how different the data sets were.

Scatterplot of  $y_1$  vs  $x_1$ ,  $y_2$  vs  $x_1$ ,  $y_3$  vs  $x_1$ ,  $y_4$  vs  $x_4$

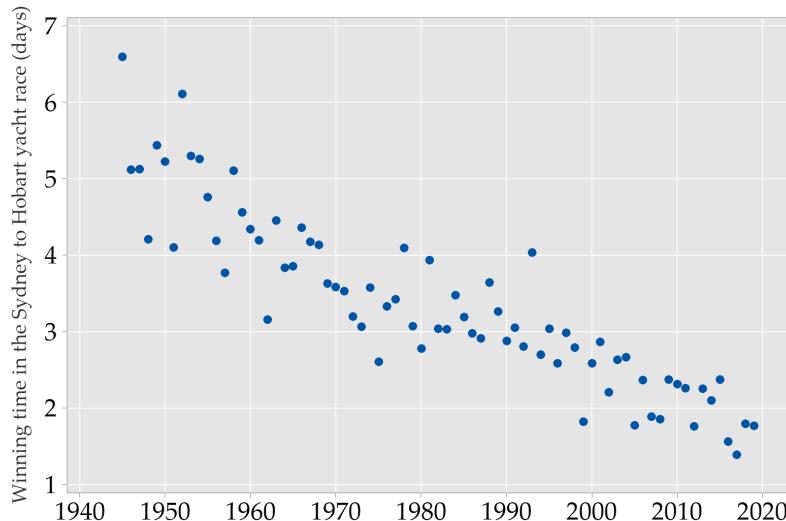


- (c) The observed value at  $x_4 = 19$  is 12.5, which is the same as the predicted value. Changing  $y_4$  from 12.5 to 10 and refitting the regression line results in a predicted value of 10, which is the same as the observed again.

From the plot, we can see that the point (19, 12.5) is used to fit the regression line, resulting in the observed being the same as the fitted.

- (d) Data set 1 ( $y_1$  on  $x_1$ ) looks reasonable for the usual assumptions and so the regression is meaningful and appropriate. Set 2 ( $y_2$  on  $x_1$ ) is curvilinear and therefore linear regression is not appropriate. Set 3 ( $y_3$  on  $x_1$ ) lies almost on an exact straight line except for one observation which looks like an outlier and should therefore be investigated further before carrying out the regression. Set 4 ( $y_4$  on  $x_4$ ) looks very unusual. The  $x$  values are identical except for one. With only two  $x$  values represented there is no way of knowing if the relationship is linear or non-linear.

- 13.2 (a) Here is the graph:



- (b) There is a negative and roughly linear relationship between the winning time and the year; winning times get faster as year increases. The correlation is  $-0.90$ .
- (c) Time is a special explanatory variable. We cannot “go back and get more observations”. So in a sense each outcome is unique, and there can’t be a distribution of  $y$  values for each time, from which we sample. Nevertheless, in some circumstances the usual regression model may be a useful way to model the variation we see in a time series. For many time series, we should expect an association between measurements close in time, even after allowing for the overall trend. This seems unlikely in this case, and in fact, there is little evidence of it.
- (d) Some relevant output:

SYDNEY\_HOBART\_2019.MWX

**Regression Analysis: Time (days) versus Year**

**Method**

Rows 2  
unused

**Regression Equation**

Time (days) = 95.99 - 0.04673 Year

**Coefficients**

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	95.99	5.18	(85.66, 106.32)	18.51	0.000	
Year	-0.04673	0.00262	(-0.05195, -0.04152)	-17.87	0.000	1.00

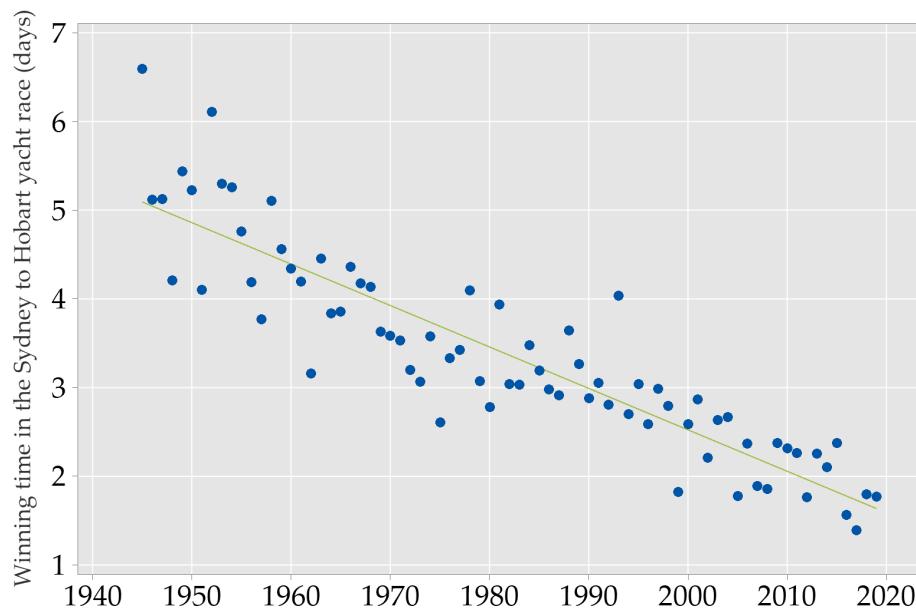
**Model Summary**

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.490384	81.39%	81.13%	18.6611	80.22%	110.27	116.88

**Analysis of Variance**

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	76.77	81.39%	76.77	76.7661	319.22	0.000
Year	1	76.77	81.39%	76.77	76.7661	319.22	0.000
Error	73	17.55	18.61%	17.55	0.2405		
Total	74	94.32	100.00%				

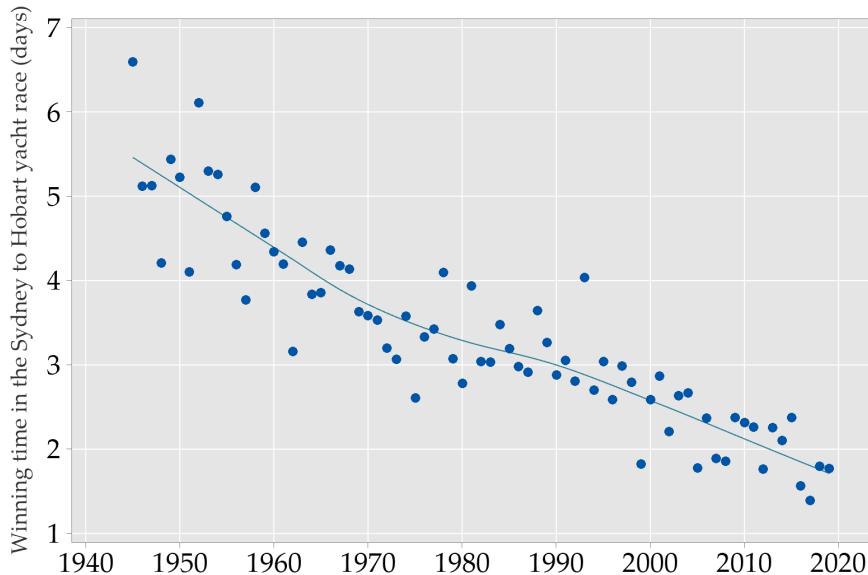
- (e) With the regression line:



- (f) The regression coefficient for year is  $-0.047$ ; this suggests an average decrease of about  $0.05$  of a day (or about  $70$  minutes) each year. Over  $74$  years, the average decline is  $3.46$  days or about  $83$  hours.
- (g) The  $95\%$  confidence interval can be seen in the output.  
It is  $(-0.052, -0.042)$ .
- (h)
  - (a) For  $2020$ , the predicted winning time is  $1.59$  days; the prediction interval is:  $0.58$  to  $2.59$ . Extrapolating just a little amount in this context is reasonable.
  - (b) In  $2054$ , the predicted winning time is  $-0.0004$  days; the prediction interval is:  $-1.05$  to  $1.05$ . This is  $-0.01$  hours (prediction interval:  $-9.42$  to  $9.40$  hours) or  $-0.6$  minutes (prediction interval:  $-565$  to  $564$  minutes).

Clearly predicting far into the future does not make sense; the winning time is predicted to be negative (and the prediction interval includes negative times) which is not possible in this context. This is because the linear model cannot continue to be reasonable for ever.

- (i) Graph with the smoother:



- 13.3 (a) Regressing taste on acetic acid:  
The regression equation is

$$\text{taste} = -61.5 + 15.6 \text{ acetic}$$

### Regression Analysis: taste versus acetic

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2314.1	30.20%	2314.1	2314.1	12.11	0.002
acetic	1	2314.1	30.20%	2314.1	2314.1	12.11	0.002
Error	28	5348.7	69.80%	5348.7	191.0		
Lack-of-Fit	27	5102.3	66.58%	5102.3	189.0	0.77	0.736
Pure Error	1	246.4	3.22%	246.4	246.4		
Total	29	7662.9	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
13.8212	30.20%	27.71%	6111.26	20.25%

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-61.5	24.8	(-112.4, -10.6)	-2.48	0.020	
acetic	15.65	4.50	(6.44, 24.86)	3.48	0.002	1.00

#### Regression Equation

$$\text{taste} = -61.5 + 15.65 \text{ acetic}$$

#### Fits and Diagnostics for Unusual Observations

Obs	taste	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
30	5.50	35.14	3.96	(27.04, 43.25)	-29.64	-2.24	-2.43	0.0819664	0.22
Obs	DFITS								
30	-0.724828 R								

*R Large residual*

Regression of taste on hydrogen sulfide:

The regression equation is taste =  $-9.79 + 5.78 \text{ H}_2\text{S}$

### Regression Analysis: taste versus H<sub>2</sub>S

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4376.7	57.12%	4376.7	4376.7	37.29	0.000
H <sub>2</sub> S	1	4376.7	57.12%	4376.7	4376.7	37.29	0.000
Error	28	3286.1	42.88%	3286.1	117.4		
Lack-of-Fit	27	3170.6	41.38%	3170.6	117.4	1.02	0.670
Pure Error	1	115.5	1.51%	115.5	115.5		
Total	29	7662.9	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
10.8334	57.12%	55.58%	3688.08	51.87%

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-9.79	5.96	(-21.99, 2.42)	-1.64	0.112	
H <sub>2</sub> S	5.776	0.946	(3.839, 7.714)	6.11	0.000	1.00

#### Regression Equation

$$\text{taste} = -9.79 + 5.776 \text{ H}_2\text{S}$$

#### Fits and Diagnostics for Unusual Observations

Obs	taste	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS
12	57.20	35.89	2.71	(30.33, 41.45)	21.31	2.03	2.16	0.0628038	0.14	0.559389
15	54.90	29.21	2.12	(24.87, 33.56)	25.69	2.42	2.67	0.0383376	0.12	0.532955
<hr/>										
12	R									
15	R									
<i>R Large residual</i>										

Regressing taste on lactic acid:

The regression equation is taste =  $-29.9 + 37.7 \text{ lactic}$

## Regression Analysis: taste versus lactic Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3800.4	49.59%	3800.4	3800.40	27.55	0.000
lactic	1	3800.4	49.59%	3800.4	3800.40	27.55	0.000
Error	28	3862.5	50.41%	3862.5	137.95		
Lack-of-Fit	26	3666.6	47.85%	3666.6	141.02	1.44	0.492
Pure Error	2	195.8	2.56%	195.8	97.92		
Total	29	7662.9	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
11.7450	49.59%	47.79%	4375.64	42.90%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-29.9	10.6	(-51.5, -8.2)	-2.82	0.009	
lactic	37.72	7.19	(23.00, 52.44)	5.25	0.000	1.00

### Regression Equation

$$\text{taste} = -29.9 + 37.72 \text{ lactic}$$

### Fits and Diagnostics for Unusual Observations

Obs	taste	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS
15	54.90	27.48	2.22	(22.94, 32.02)	27.42	2.38	2.61	0.0356111	0.10	0.502229
<hr/>										
<hr/>										

*R Large residual*

Independently each of the explanatory variables appears to be a useful predictor of taste, with  $R^2$  values of 30% or more.

- (b) Fitting a multiple regression of taste on all three variables:  
The regression equation is

$$\text{taste} = -28.9 + 0.33 \text{ acetic} + 3.91 \text{ H}_2\text{S} + 19.7 \text{ lactic}$$

## Regression Analysis: taste versus acetic, H<sub>2</sub>S, lactic

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	3	4994.5	65.18%	4994.48	1664.83	16.22	0.000
acetic	1	2314.1	30.20%	0.55	0.55	0.01	0.942
H <sub>2</sub> S	1	2147.0	28.02%	1007.66	1007.66	9.82	0.004
lactic	1	533.3	6.96%	533.32	533.32	5.20	0.031
Error	26	2668.4	34.82%	2668.41	102.63		
Total	29	7662.9	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
10.1307	65.18%	61.16%	3402.24	55.60%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-28.9	19.7	(-69.4, 11.7)	-1.46	0.155	
acetic	0.33	4.46	(-8.84, 9.49)	0.07	0.942	1.83
H <sub>2</sub> S	3.91	1.25	(1.35, 6.48)	3.13	0.004	1.99
lactic	19.67	8.63	(1.93, 37.41)	2.28	0.031	1.94

### Regression Equation

$$\text{taste} = -28.9 + 0.33 \text{ acetic} + 3.91 \text{ H}_2\text{S} + 19.67 \text{ lactic}$$

### Fits and Diagnostics for Unusual Observations

Obs	taste	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D	DFITS
15	54.90	29.45	3.04	(23.20, 35.70)	25.45	2.63	3.02	0.0900034	0.17	0.948341
<hr/>										
<i>Obs</i>										
15 R										
<i>R Large residual</i>										

The multiple regression accounts for 61% of the variability in taste. However, not all of the explanatory terms are needed. The *P*-value for acetic acid is 0.9 which implies that it can (possibly should) be removed from the model. Removing acetic acid:  
The regression equation is

$$\text{taste} = -27.6 + 3.95 \text{ H}_2\text{S} + 19.9 \text{ lactic}$$

# Regression Analysis: taste versus H<sub>2</sub>S, lactic Acid Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	4993.9	65.17%	4993.9	2496.96	25.26	0.000
H2S	1	4376.7	57.12%	1193.5	1193.52	12.07	0.002
lactic	1	617.2	8.05%	617.2	617.18	6.24	0.019
Error	27	2669.0	34.83%	2669.0	98.85		
Total	29	7662.9	100.00%				

## Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
9.94236	65.17%	62.59%	3135.44	59.08%

## Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-27.59	8.98	(-46.02, -9.16)	-3.07	0.005	
H2S	3.95	1.14	(1.62, 6.28)	3.47	0.002	1.71
lactic	19.89	7.96	(3.56, 36.22)	2.50	0.019	1.71

## Regression Equation

$$\text{taste} = -27.59 + 3.95 \text{ H}_2\text{S} + 19.89 \text{ lactic}$$

## Fits and Diagnostics for Unusual Observations

The removal of acetic acid has virtually no effect on the fit of the regression model. In the 'reduced' model, both of the explanatory variables remain statistically significant.

- (c) In Chapter 2 you found the correlations between taste and each of the explanatory variables; the regression equations in part (i) above are consistent with the correlations, as expected.

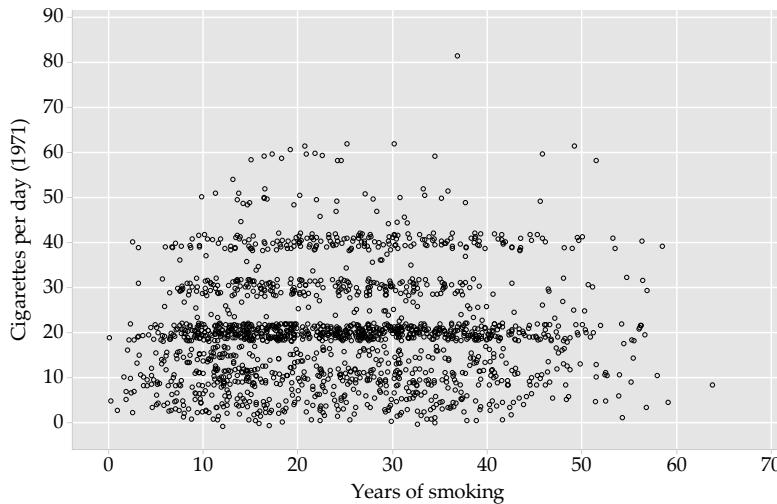
The high correlations between the three explanatory variables help to explain why all three are not needed in the multiple regression model. Given the values of  $H_2S$  and lactic acid, we can obtain quite a reasonable estimate of the value of acetic acid, and once  $H_2S$  and lactic acid are allowed for, the actual concentration of acetic acid has little to tell us about taste.

A useful summary table of the final model is:

Regression:  $F_{2,27} = 25.3$ ,  $P < 0.001$ ,  $R^2 = 0.63$

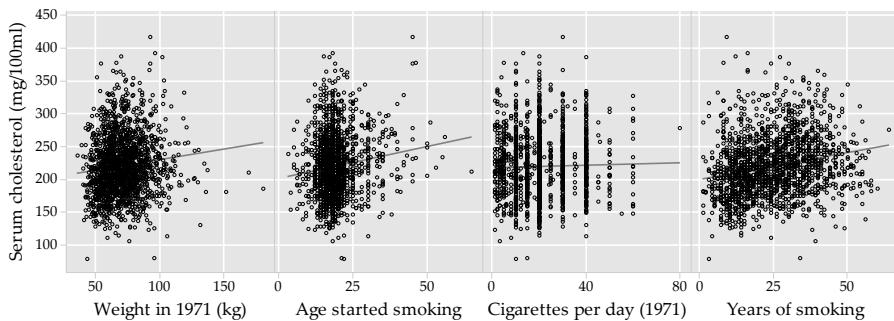
Explanatory variable	Estimate	Regression coefficient		<i>P</i> -value
		95% confidence interval		
log of Concentration of hydrogen sulphide	3.95	1.62, 6.28		0.002
Concentration of lactic acid	19.89	3.56, 36.22		0.019

- 13.4 (a) Serum cholesterol levels are expected to be associated with age, weight and smoking; in each case, the association is positive, with higher serum cholesterol with increasing age, weight and amount of smoking. It is always useful to consider the likely pattern of results you expect before you start your analysis.
- (b) Cigarettes per day and years of smoking need not be strongly associated, and are both potentially useful explanatory variables. Smokers often have a particular 'habit' that is not related to the duration of smoking. The scatterplot below confirms this. There is little evidence of an association.



Note that the points on this graph have been jittered to make overlapping points visible.

- (c) Produce a single plot showing the relationship between the outcome and each of the explanatory variables.



Note that when you produce this graph, if you hover the cursor over one of the lines, brief details of the fitted line will be displayed; this can be useful when quickly exploring data.

- (d) The output is shown below. The regression coefficient, to two decimal places, is 0.81. This means that for each additional year of smoking, the regression model predicts an increase of 0.81 mg/100ml in serum cholesterol.

## Regression Analysis: Serum cholesterol (mg/100ml) ... ears of smoking

### Method

Rows unused 20

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	168045	4.70%	168045	168045	85.00	0.000
Years of smoking	1	168045	4.70%	168045	168045	85.00	0.000
Error	1724	3408259	95.30%	3408259	1977		
Lack-of-Fit	58	162601	4.55%	162601	2803	1.44	0.018
Pure Error	1666	3245658	90.75%	3245658	1948		
Total	1725	3576304	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
44.4629	4.70%	4.64%	3416447	4.47%

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	199.49	2.42	(194.74, 204.25)	82.27	0.000	
Years of smoking	0.8149	0.0884	(0.6415, 0.9882)	9.22	0.000	1.00

#### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 199.49 + 0.8149 \text{ Years of smoking}$$

- (e) Check the last column in the data file to make sure the result has been calculated.
- (f) The output is shown below; it is identical to the output in part (d), except for information about the regression coefficient. The regression coefficient, to one decimal place, is 8.1. This means that for each additional decade of smoking, the regression model predicts an increase of 8.1 mg/100ml in serum cholesterol. It may be more useful to think of changes in cholesterol levels in terms of decades, rather than years, of smoking. This example illustrates how an explanatory variable can be linearly rescaled to allow interpretation of the regression coefficient on a practically meaningful scale.

## Regression Analysis: Serum cholesterol (mg/100ml) ... ades of smoking

### Method

Rows unused 20

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	168045	4.70%	168045	168045	85.00	0.000
Decades of smoking	1	168045	4.70%	168045	168045	85.00	0.000
Error	1724	3408259	95.30%	3408259	1977		
Lack-of-Fit	58	162601	4.55%	162601	2803	1.44	0.018
Pure Error	1666	3245658	90.75%	3245658	1948		
Total	1725	3576304	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
44.4629	4.70%	4.64%	3416447	4.47%

#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	199.49	2.42	(194.74, 204.25)	82.27	0.000	
Decades of smoking	8.149	0.884	(6.415, 9.882)	9.22	0.000	1.00

#### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 199.49 + 8.149 \text{ Decades of smoking}$$

- (g) The simple linear regressions predicting serum cholesterol for smokers from each of the the other three variables are shown below.

## Regression Analysis: Serum cholesterol (mg/100ml) ... ight in 1971 (kg)

### Method

Rows unused 20

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	46938	1.31%	46938	46938	22.93	0.000
Weight in 1971 (kg)	1	46938	1.31%	46938	46938	22.93	0.000
Error	1724	3529366	98.69%	3529366	2047		
Lack-of-Fit	533	1172524	32.79%	1172524	2200	1.11	0.073
Pure Error	1191	2356843	65.90%	2356843	1979		
Total	1725	3576304	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
45.2460	1.31%	1.26%	3538073	1.07%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	196.28	4.98	(186.52, 206.05)	39.41	0.000	
Weight in 1971 (kg)	0.3267	0.0682	(0.1929, 0.4605)	4.79	0.000	1.00

### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 196.28 + 0.3267 \text{ Weight in 1971 (kg)}$$

## Regression Analysis: Serum cholesterol (mg/100ml) ... started smoking

### Method

Rows unused 20

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	59164	1.65%	59164	59164	29.00	0.000
Age started smoking	1	59164	1.65%	59164	59164	29.00	0.000
Error	1724	3517140	98.35%	3517140	2040		
Lack-of-Fit	47	88865	2.48%	88865	1891	0.92	0.619
Pure Error	1677	3428275	95.86%	3428275	2044		
Total	1725	3576304	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
45.1675	1.65%	1.60%	3526013	1.41%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	201.34	3.55	(194.37, 208.31)	56.66	0.000	
Age started smoking	0.956	0.178	(0.608, 1.305)	5.39	0.000	1.00

### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 201.34 + 0.956 \text{ Age started smoking}$$

## Regression Analysis: Serum cholesterol (mg/100ml) ... es per day

(1971)

### Method

Rows unused 20

### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1885	0.05%	1885	1885	0.91	0.340
Cigarettes per day (1971)	1	1885	0.05%	1885	1885	0.91	0.340
Error	1724	3574419	99.95%	3574419	2073		
Lack-of-Fit	36	78170	2.19%	78170	2171	1.05	0.391
Pure Error	1688	3496249	97.76%	3496249	2071		
Total	1725	3576304	100.00%				

### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
45.5338	0.05%	0.00%	3583045	0.00%

### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	217.74	2.20	(213.42, 222.05)	98.93	0.000	
Cigarettes per day (1971)	0.0886	0.0929	(-0.0936, 0.2708)	0.95	0.340	1.00

### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 217.74 + 0.0886 \text{ Cigarettes per day (1971)}$$

The results are tabulated below.

Regression models of Serum cholesterol level (mg/100 ml) in smokers

<i>Explanatory variable</i>	Regression coefficient from Simple linear regression			Regression coefficient from Multiple linear regression		
	Estimate	95% CI	P-value	Estimate	95% CI	P-value
Age started smoking	0.96	0.61, 1.31	<0.001	1.58	1.00, 1.93	<0.001
Weight in kilograms	0.33	0.19, 0.46	<0.001	0.33	0.20, 0.46	<0.001
Cigarettes per day	0.09	-0.09, 0.27	0.340	0.15	-0.03, 0.33	0.098
Decades of smoking	8.15	6.42, 9.88	<0.001	9.88	8.13, 11.62	<0.001

(h) The results shown below are added to the table above.

Note that the regression coefficient for weight in 1971 is the same in both analyses (to 4 decimal places, in fact, which is remarkable). All of the other three variables have slightly larger regression coefficients in the multiple regression than in the simple linear regressions. This is because two out of the three associations between these variables (specifically, the associations between age started smoking and cigarettes per day, and between age started smoking and decades of smoking), are negative. *As an aside: it is impossible for a set of three or more variables to have all pair-*

wise associations strongly negative. On the other hand, it is possible, and quite common, for a set of variables to have all pairwise associations positive, or even strongly positive.)

### Regression Analysis: Serum cholesterol (mg/100ml) ... es per day (1971)

#### Method

Rows unused 20

#### Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	4	351210	9.82%	351210	87803	46.85	0.000
Age started smoking	1	59164	1.65%	144905	144905	77.33	0.000
Weight in 1971 (kg)	1	56069	1.57%	46374	46374	24.75	0.000
Decades of smoking	1	230845	6.45%	230574	230574	123.04	0.000
Cigarettes per day (1971)	1	5132	0.14%	5132	5132	2.74	0.098
Error	1721	3225094	90.18%	3225094	1874		
Lack-of-Fit	1719	3214921	89.90%	3214921	1870	0.37	0.934
Pure Error	2	10173	0.28%	10173	5086		
Total	1725	3576304	100.00%				

#### Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
43.2893	9.82%	9.61%	3245643	9.25%

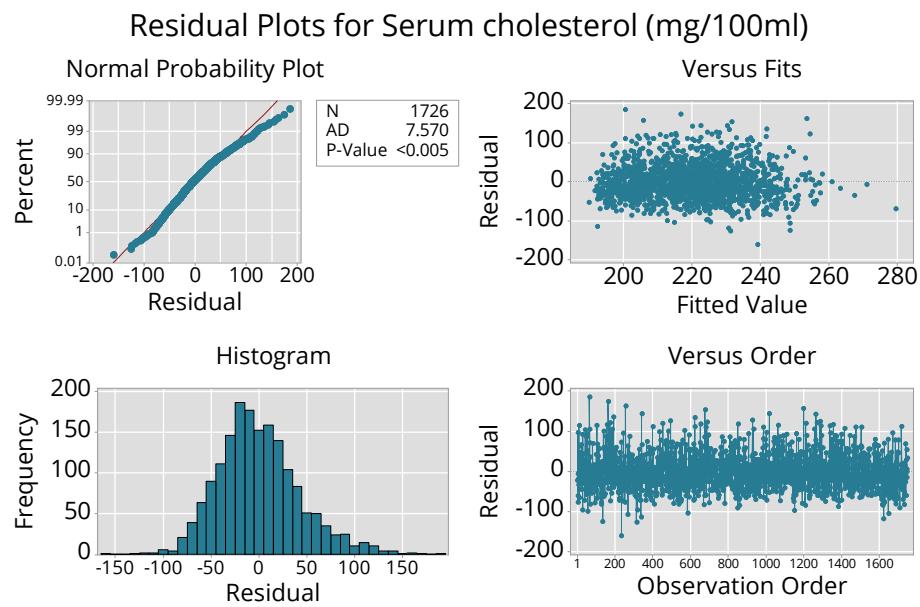
#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	138.91	6.97	(125.23, 152.58)	19.93	0.000	
Age started smoking	1.575	0.179	(1.223, 1.926)	8.79	0.000	1.11
Weight in 1971 (kg)	0.3267	0.0657	(0.1979, 0.4555)	4.97	0.000	1.01
Decades of smoking	9.877	0.890	(8.130, 11.623)	11.09	0.000	1.07
Cigarettes per day (1971)	0.1490	0.0900	(-0.0276, 0.3256)	1.65	0.098	1.04

#### Regression Equation

$$\text{Serum cholesterol (mg/100ml)} = 138.91 + 1.575 \text{ Age started smoking} + 0.3267 \text{ Weight in 1971 (kg)} + 9.877 \text{ Decades of smoking} + 0.1490 \text{ Cigarettes per day (1971)}$$

- (i) The residual plots from the multiple regression analysis are shown below. Although the residuals show some deviation from a normal distribution, this is not a concern. The plot of residuals against fitted values does not show a systematic pattern; hence the assumption of constant variance is reasonable.



# 14 Inference — binary outcome

## 14.1 Introduction

The inferences for proportions we have considered up until now have involved simple contexts, such finding a confidence interval for a comparison of two proportions.

We now consider the statistical modelling of proportions. The linear model often applied to numerical outcomes with Normally distributed errors are usually inappropriate for proportions for the following reasons.

- (a) Unlike the case of Normally distributed errors, the variance of the data is intrinsically related to its mean. Think of the simplest model for a count of binary data:  $Y \stackrel{d}{=} \text{Bi}(n, \theta)$ . For this binomial distribution,  $E(Y) = n\theta$  and  $\text{var}(Y) = n\theta(1 - \theta)$ , so the mean and the variance are related.
- (b) The application of a linear model can lead to absurd estimates, such as estimated proportions outside the range (0,1).

It is helpful to get an overall perspective on the modelling of proportions. The models usually start with the idea that we have binomial data — like the proportion of people testing positive for hepatitis B, or the proportion of musicians with performance-related medical disorder (PRMD) — and our aim is to *model*  $\theta$ , the probability of “success”, in terms of measured characteristics.

We are envisaging a situation in which we have sample proportions  $\hat{\theta}_i$ , where  $i$  indexes the distinct proportions. A given sample proportion can be expressed as  $\hat{\theta}_i = \frac{Y_i}{n_i}$ , where  $n_i$  is the total number of units (subjects etc.) and  $Y_i$  is the number of units among the total which have a characteristic of interest.

We model the numerators of the proportions,  $Y_i$ , as binomial random variables with parameters  $n_i$  and  $\theta_i$ , and we seek to relate  $\theta_i$  to measurable characteristics. We use  $Y$  for the numerator of the proportion. This is because we are about to use models analogous to linear models, so that  $x$  will be used for explanatory variables.

Some examples:

Context	$Y_i$	$n_i$	Characteristics
Biased coins	No. heads	No. tosses	Weight of coin, starting position of toss, variation in coin density, ...
Opinion poll	No. "yes"	Total no.	Demographics: Age, sex, education, occupation, social class, ...
Drug trial	No. survived	Total no.	Drug used, severity of diagnosis, age, type of tumour, ...
Case-control	No. cases	Cases + controls	Exposure category, level of confounding variable, sex, age, ...
Performance-related medical disorder (PRMD)	No. with PRMD	Total	Instrument played, age, hours of playing per day, ...
New car warranties	No. of claims	Total	Model of car, variant, types of components, ...

One simple situation involves a single, continuous explanatory variable,  $x$ , and an outcome which is a proportion,  $\frac{Y}{n}$  for each value of  $x$ . This is quite realistic: for example, a toxic substance is administered to several groups of mice at varying doses, and the proportions of mice developing a tumour are recorded, for each dose.

In such a situation the number developing the tumour,  $Y$ , is naturally modelled as a binomial random variable with mean  $n\theta$  and variance  $n\theta(1 - \theta)$ , where  $\theta$  represents the probability of a tumour developing. As mentioned above, the mean and the variance of  $Y$  are intrinsically related, so, in general, we know we will not have a constant population variance. Contrast this with the linear model, where one of the important assumptions is that the random error always has the same variance, regardless of the value of the mean.

Second, it is clear that — if the substance really is carcinogenic — the relationship between the dose and the proportion of tumours cannot be linear, because if we increase the dose beyond the minimum level that induces tumours in a whole group, we cannot increase the proportion further. Intuitively, this suggests that any plausible model will show “diminishing returns” as we approach the extremities of 0 and 1.

In other words, we seek an alternative to  $\theta = a + bx$  which has the following features:

- As  $x$  increases,  $\theta$  either increases or decreases;
- No matter what the value of  $x$ ,  $\theta$  lies between 0 and 1—so that we always get sensible estimates;
- As  $x$  gets very small or very large, the effect of a change in  $x$  diminishes.

## 14.2 The logistic regression model

There are a number of models which have the above properties. Here we are going to discuss the one most commonly used, the logistic model, whose basic form is:

$$\theta = \frac{\exp(x)}{1 + \exp(x)} = \frac{e^x}{1 + e^x}$$

The number  $e = 2.718\dots$ ; it is the base of the natural logarithms. Recall that  $e^0 = 1$ ,  $e^\infty = \infty$  and  $e^{-\infty} = 0$ . The basic logistic function is shown in Figure 93.

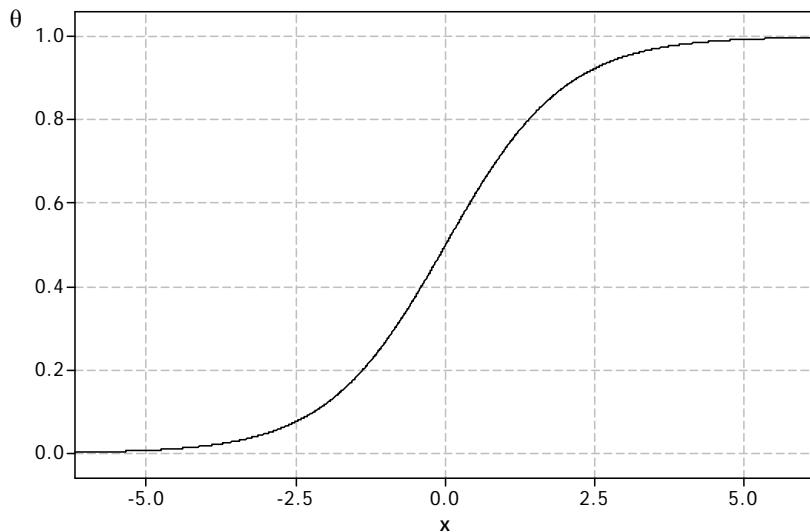


Figure 93: The logistic function; the probability or proportion is on the y-axis, and the explanatory variable, such as ‘dose’, is on the x-axis.

Now consider the modelling of actual proportions. A simple way to incorporate an unknown relationship between  $x$  and  $\theta$  is to introduce a linear function of  $x$  into the model. Now it looks like this:

$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

This gives us some flexibility:

- Different values of  $\beta_0$  shift the curve horizontally, allowing the increase in  $\theta$  to occur for the relevant values of  $x$ . Thus  $\beta_0$  performs the role of locating the model in an overall sense.
- Different values of  $\beta_1$  lead to different rates of increase (or decrease) in  $\theta$  with increasing  $x$ . If  $\beta_1 = 0$ , then there is no change in  $\theta$  when  $x$  changes. If  $\beta_1 > 0$ , then  $\theta$  increases as  $x$  increases, and the greater the size of  $\beta_1$  the “faster” the rate of increase. If  $\beta_1 < 0$ , then  $\theta$  decreases as

$x$  increases, and the greater the magnitude of  $\beta_1$  the “faster” the rate of decrease.

There is a very important piece of algebra which rearranges the model:

$$\begin{aligned}\theta(x) &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \\ \Rightarrow 1 - \theta(x) &= \frac{1}{1 + \exp(\beta_0 + \beta_1 x)} \\ \Rightarrow \frac{\theta(x)}{1 - \theta(x)} &= \exp(\beta_0 + \beta_1 x) \\ \Rightarrow \log\left(\frac{\theta(x)}{1 - \theta(x)}\right) &= \beta_0 + \beta_1 x\end{aligned}$$

$\frac{\theta(x)}{1 - \theta(x)}$  is referred as the “odds” of the event.

In words: the logarithm of the odds is a linear function of  $x$ . There is a name given to this particular function: the “logit” function. That is,  $\text{logit}(\theta)$  is equal to  $\log \frac{\theta}{1-\theta}$ .

What does this tell us about the odds ratio, comparing  $x+1$  to  $x$ ? The odds for  $x+1$  are  $\frac{\theta(x+1)}{1-\theta(x+1)}$  and the odds for  $x$  are  $\frac{\theta(x)}{1-\theta(x)}$ , so the odds ratio comparing  $x+1$  to  $x$  is

$$\frac{\theta(x+1)/(1-\theta(x+1))}{\theta(x)/(1-\theta(x))} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1x)} = \exp(\beta_1) = e^{\beta_1}.$$

There are two simple cases to consider:

- (a) When  $x$  is a continuous variable (e.g.  $x = \text{age in years}$ ),  $\beta_1$  represents the change in  $\log(\text{odds})$  corresponding to a change in one unit of  $x$ . It follows that in this case  $e^{\beta_1}$  is the odds ratio corresponding to an increase of one unit in  $x$ .
- (b) When  $x = 0$  or  $1$  indexes a particular group (e.g.  $x = 1$  for treated and  $0$  for placebo)  $\beta_1$  represents  $\log(OR)$  for exposure. For this case  $e^{\beta_1}$  is the odds ratio for treatment relative to placebo.

The estimates of the  $\beta$ s are obtained using the method of maximum likelihood, which means they will be approximately Normally distributed in large samples. This means that we can use the “estimate  $\pm 1.96$  standard error” approach to calculating a confidence interval for an individual  $\beta$ . These are on the logarithmic scale; they can then be transformed back to the odds ratio scale, and these are the final confidence intervals presented and interpreted.

### 14.2.1 One binary explanatory variable

▷ **EXAMPLE.** Data are collected on two brands of car over the period of the new car warranty (see `warranties.mwx`). Among 120 brand A owners, 31 warranty claims were made. Among 200 brand B owners, 25 claims were made. Brand A had 31/120 claims, while brand B had 25/200 claims.

Consider fitting a logistic regression. We can think of this in at least two (equivalent) ways.

- (a) We can have the data set as 320 rows (one per car owner), and columns for brand (A or B) and warranty claim (0: no claim or 1: claim). This is known as the ‘long’ form of the data. From a formal point of view, we are then thinking of 320 independent observations, each of which can take the value 0 or 1. The chance that one of these observations is equal to 1 (a claim) is modelled as

$$\theta = \begin{cases} \frac{e^{\beta_0}}{1+e^{\beta_0}} & \text{for brand A;} \\ \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} & \text{for brand B;} \end{cases}$$

Note that if we parameterise the model in this way,  $\beta_1$  is the log odds ratio for brand B, relative to brand A. However this is arbitrary really, since the inference is ultimately the same either way around.

Software will often assume that the lowest numbered level of a categorical variable, or the first alphabetically (if the software copes with string variables for categorical variables) is the “baseline” level against which other levels are compared. Often, software will allow you to set the baseline level to a different category.

- (b) The second way to look at this data is the way as two proportions, 31/120 and 25/200, indexed by the brand of car.

The MINITAB file `warranties.mwx` has the data in the long form.

The main way to fit a logistic regression in MINITAB is to use **Stat > Regression > Binary logistic regression**. Put the 0/1 response variable in the **Response** box. Enter the explanatory variable(s) in the **Continuous predictors** or **Categorical predictors** box as appropriate. Here the explanatory variable is categorical.

MINITAB 19 has a default setting in which each level of a categorical explanatory variable is compared to the *first* level, that is, the smallest when numerically coded or the first level set in the **Value order** when coded with text; the default order for text variables is alphabetical. However, the setting in MINITAB can be changed, by using the options under the **Coding** button and changing the **Reference level**.

In this analysis, we will use the MINITAB default setting and have the *first* level of brand (brand A) as the baseline.

### Regression Equation

$$\begin{aligned} P(1) &= \exp(Y')/(1 + \exp(Y')) \\ Y' &= -1.055 + 0.0 \text{ Brand\_A} - 0.891 \text{ Brand\_B} \end{aligned}$$

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-1.055	0.209	(-1.463, -0.646)	-5.06	0.000	
Brand B	-0.891	0.299	(-1.477, -0.306)	-2.98	0.003	1.00

### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Brand B	A	0.4101	(0.2284, 0.7365)

*Odds ratio for level A relative to level B*

In the MINITAB output you can see the estimate of the odds ratio and a 95% confidence interval. This gives the odds ratio comparison for brand B compared to brand A. The odds of a claim for brand B is estimated to be 0.41 times the odds for brand A, with an approximate 95% confidence interval of (0.23 to 0.74).

If we had reversed the baseline, the odds ratio we obtain is for brand A compared with brand B; it is the reciprocal of the odds ratio we obtained in the first analysis. The odds of a claim is estimated to be 2.44 times higher for brand A than for brand B, with an approximate 95% confidence interval of (1.36 to 4.38).

#### 14.2.2 Categorical explanatory variables with more than two levels

A categorical explanatory variable may have more than two levels. In that case, its effect can not be represented by a single parameter. If it has  $k$  levels, the usual practice is to present estimates of  $k - 1$  of the levels in relation to a reference level. It is customary, though not necessary, to make the reference level either the first or last category. The default in MINITAB is the first category, which can be altered, as explained above.

▷ **EXAMPLE.** Consider the data set bd3.mwx, which is from a case-control study of oesophageal cancer in the French department of Ille-et-Vilaine. We model case-control status as the outcome variable. There are two categorical explanatory variables, namely, age and level of alcohol consumption. They have the following levels:

alcohol	Count	age	Count
0	415	0	116
1	355	1	199
2	138	2	213
3	67	3	242
N=	975	4	161
		5	44
		N=	975

When we fit a “main effects” logistic regression in MINITAB the output includes the following:

#### Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
<b>alcohol</b>			
1	0	4.1967	(2.5975, 6.7807)
2	0	7.4418	(4.3189, 12.8228)
3	0	39.6469	(18.9614, 82.8986)
2	1	1.7732	(1.1197, 2.8083)
3	1	9.4471	(4.8243, 18.4995)
3	2	5.3276	(2.5983, 10.9237)
<b>age</b>			
1	0	5.1096	(0.6154, 42.4259)
2	0	30.7486	(4.0137, 235.5606)
3	0	51.5966	(6.7923, 391.9455)
4	0	78.0053	(10.1346, 600.4037)
5	0	83.4484	(9.8284, 708.5253)
2	1	6.0178	(2.7152, 13.3375)
3	1	10.0980	(4.6552, 21.9043)
4	1	15.2664	(6.8218, 34.1645)
5	1	16.3317	(5.8174, 45.8492)
3	2	1.6780	(1.0460, 2.6919)
4	2	2.5369	(1.5179, 4.2400)
5	2	2.7139	(1.1857, 6.2118)
4	3	1.5118	(0.9385, 2.4353)
5	3	1.6173	(0.7243, 3.6113)
5	4	1.0698	(0.4715, 2.4272)

*Odds ratio for level A relative to level B*

You will see that MINITAB provides odds ratios for all pairs of levels, with the ‘lower’ level as the baseline in each case.

For example, in the case of alcohol consumption, the lowest level is 0 – 39 g daily, on average. The fourth and highest category, coded 3, is 120+ g daily on average. The odds ratio comparing the highest consumption category to the lowest is estimated to be 40, with a 95% confidence interval of (19 to 83).

It is obvious that each of these categorical variables is statistically significant overall. The formal test of this in MINITAB is shown in the output for each explanatory variable. These are summarised below:

<b>Source</b>	<b>DF</b>	<b>Wald Test</b>	
		<b>Chi-Square</b>	<b>P-Value</b>
Regression	8	145.58	0.000
alcohol	3	108.55	0.000
age	5	63.81	0.000

### 14.2.3 One continuous explanatory variable

▷ **EXAMPLE.** Now we consider a simple example involving a continuous explanatory variable. The data set artery.mwx involves data collected on the health of arteries collected from patients undergoing coronary bypass surgery.

Let's consider fitting a logistic regression to estimate the effect of age on medial calcification in the radial artery; age is a common continuous explanatory variable, in many different contexts.

Note that we don't have grouped or aggregated data, as there are many different ages in the data. In fact, if a continuous variable is measured accurately enough, we may have a different  $x$  value for each subject.

The output we get from MINITAB is as follows.

#### Coefficients

<b>Term</b>	<b>Coef</b>	<b>SE Coef</b>	<b>95% CI</b>	<b>Z-Value</b>	<b>P-Value</b>	<b>VIF</b>
Constant	-7.34	2.93	(-13.08, -1.60)	-2.51	0.012	
AGE	0.0796	0.0417	(-0.0021, 0.1613)	1.91	0.056	1.00

#### Odds Ratios for Continuous Predictors

<b>Odds Ratio</b>	<b>95% CI</b>
AGE	1.0828 (0.9979, 1.1750)

Remember what this odds ratio of 1.08 means. It is the estimate of the odds ratio corresponding to an increase of one unit in the explanatory variable. In this case age was recorded in years, so we estimate that that odds of medial calcification in the radial artery is 1.08 times higher for each extra year of age.

Something that often causes confusion here is the comparison involved; people ask which two ages are being compared? In terms of the model and the result, the answer is: any two ages differing by one year. So 1.08 is the estimated odds ratio comparing 45 with 44, but also for comparing 55 with

54, and 80 with 79, etc.; in general, for age  $x + 1$  compared to age  $x$ .

This means that any  $\beta$  or odds ratio for a continuous explanatory variable coming from a logistic regression *must* be interpreted in terms of the units of  $x$ . This is also true of ordinary linear regression, of course.

The example is typical, in that the odds ratio for an increase of one year in age is quite small, but the lower bound of the 95% confidence interval is 1.00.

However, we should not and need not be constrained to use the unit of age inherent in the data. We can consider an alternative and perhaps more meaningful age interval, such as 10 years; the age range in the data is about 40 years.

Since  $e^{\hat{\beta}_1}$  is the odds ratio for an increase of 1 year in age,  $e^{10\hat{\beta}_1}$  is the odds ratio for an increase in 10 years. Both  $\hat{\beta}_1$  and  $se(\hat{\beta}_1)$  are multiplied by 10 to get the inferences for an increment of a decade.

The “lazy” way to do this in software is to create a new variable that is in the units we would really like to use. For example, if we convert age in years to age in decades by dividing by 10, and re-fit the logistic regression, here is what we get:

### Coefficients

Term	Coef	SE Coef	95% CI	Z-Value	P-Value	VIF
Constant	-7.34	2.93	(-13.08, -1.60)	-2.51	0.012	
Age (decades)	0.796	0.417	(-0.021, 1.613)	1.91	0.056	1.00

### Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Age (decades)	2.2161	(0.9790, 5.0164)

Note that the  $P$ -value for testing that the true odds ratio = 1 remains exactly the same, and the lower bound is still essentially at 1 (the lower limit of the odds ratio for the one year result was, in fact, just below 1, but rounded to 1.00). But the odds ratio is probably now more readily interpreted: the odds of medial calcification in the radial artery is estimated to be 2.2 times higher for each extra decade of age.

## 14.3 Estimated probabilities

Having estimated the parameters of the logistic regression, we can, in turn, estimate the probabilities of a response for a given set of explanatory variables. MINITAB produces these when you ask for them. This can be done via Logistic regression, Storage, then check the Fits (Event probabilities)

box. These are also called the “fitted values” (because they result from fitting the model) or the “predicted values.”

With grouped data, these estimated probabilities can be directly compared with the observed ones.

Whether the data are grouped or ungrouped, these estimated probabilities can be intrinsically useful: they give us an objective way to estimate the chance of the response occurring, given a particular configuration of explanatory variables.

In fact, this application of logistic regression was an important one historically in the early use of the model in medicine, and epidemiology in particular. In studies such as the Framingham heart study, the risk of, say, 5-year mortality, or some other adverse binary outcome, was estimated using logistic regression. If the data set is representative and large, and the model is a good fit, this enables us to estimate the risk of death for quite an extensive list of explanatory variables, such as age, gender, diabetes, hypertension, smoking status, previous acute myocardial infarction, peripheral vascular disease, and so on.

Logistic regression is also used in finance to predict the probability of a customer defaulting on a loan, and in many other applications. Since binary outcomes are very common, logistic regression is very widely used indeed.

## 14.4 Assessing goodness of fit

For “grouped” data, when there are multiple observations for each of the combined levels of the explanatory variables, there are a number of approaches to testing the goodness of fit of a logistic model.

- The overall deviance can be used as a measure of whether the model is a good fit; the deviance is distributed approximately as a  $\chi^2$  distribution with stated degrees of freedom. A small  $P$ -value indicates that the model is not a good fit.
- Pearson’s  $\chi^2$  statistic also can be used in this way and the two statistics are often similar.

Additional ways to assess the goodness of fit are:

- The Hosmer-Lemeshow test looks at the association between the response variable and the estimated distribution of the response probabilities from the model. It can be used in a similar way. It also appears in the MINITAB output.

## 14.5 Exercises

- 14.1 Open the data file `sholom.mwx`. These data are from a study of the association between maternal alcohol consumption and the risk of a low birth weight baby. The data shows the outcome for 900 births: being below the tenth percentile in birth weight (`lowbirthwt`) for the 18 categories defined by
- the mother's social class (`class`): three levels (see file);
  - drinking habits of the mother (`alcohol`): light, moderate or heavy;
  - whether or not the mother smoked (`smoke`): non-smoker or smoker.
- (a) Fit a logistic regression model predicting low birth weight from mother's smoking. Use **Stat > Regression > Binary logistic regression > Fit binary logistic model ...**. Use `lowbirthwt` as the **Response:** variable. Remember to include smoking as a **Categorical predictor**, and click on **Coding** to check the reference category. Make sure it is non-smokers.  
Report and explain the odds ratio and the associated confidence interval. What do you conclude?
- (b) Carry out a logistic regression predicting low birth weight from mother's alcohol consumption. First, examine the values of `alcohol` and decide on the appropriate baseline.  
Record the overall test for the effect of alcohol, and interpret the *P*-value.  
Record and explain the odds ratios and the associated confidence intervals. What do you conclude?
- (c) Fit a logistic regression predicting low birth weight from both mother's smoking and alcohol consumption. Examine the odds ratios and the associated confidence intervals. How do the results compare with the two separate models fitted above?
- 14.2 Open the data file `bloodfat.mwx`. Data were collected on the concentration of plasma cholesterol and plasma triglycerides (mg/dl) for 371 male patients evaluated for chest pain. For 51 patients there was no evidence of heart disease; for the remaining 320 there was evidence of narrowing of the arteries.  
In the data file, the outcome, Heart disease, is coded 0 = No heart disease, 1 = Narrowing of arteries.
- (a) Fit a logistic regression model predicting narrowing of the arteries from plasma triglycerides and plasma cholesterol. Report the relevant results.

- (b) Explain the interpretation of the odds ratio for one of the explanatory variables.
- (c) Investigate the distribution of both the explanatory variables. Describe the range of each of the explanatory variables and decide on an appropriate unit for each variable.
- (d) Re-calculate the odds ratios and confidence intervals reported above using the new units for the explanatory variables. You will need to calculate new variables using **Calc > Calculator**.
- (e) Describe the effect of each explanatory variable measured in the new units.

## 14.6 Answers

- 14.1 (a) The odds ratio for low birth weight, using non-smoking mothers as the baseline, is 2.03. The odds of having a low birth weight baby for smoking mothers are twice the odds for non-smoking mothers. The 95% confidence interval is (1.31, 3.14).

**Odds Ratios for Categorical Predictors**

Level A	Level B	Odds Ratio	95% CI
smoke			
smoker	non-smoker	2.0300	(1.3136, 3.1372)

*Odds ratio for level A relative to level B*

**Analysis of Variance**

Wald Test			
Source	DF	Chi-Square	P-Value
Regression	1	10.17	0.001
smoke	1	10.17	0.001

- (b) The overall test for the effect of alcohol is  $\chi^2 = 9.88$ ,  $P = 0.007$ . The  $P$ -value suggests that the odds of having a low birth weight baby for different levels of alcohol consumption are not consistent with the null hypothesis of no association.

**Odds Ratios for Categorical Predictors**

Level A	Level B	Odds Ratio	95% CI
alcohol			
medium	heavy	0.5278	(0.2778, 1.0029)
light	heavy	0.4694	(0.2906, 0.7580)
light	medium	0.8893	(0.4917, 1.6085)

*Odds ratio for level A relative to level B*

**Analysis of Variance**

Wald Test			
Source	DF	Chi-Square	P-Value
Regression	2	9.88	0.007
alcohol	2	9.88	0.007

In this analysis, the pairwise odds ratios are provided. For example, the odds of having a low birth weight baby for moderate drinkers are about half (0.53) the odds for heavy drinkers. The 95% confidence interval is (0.28, 1.00). The odds of having a low birth weight baby for light drinkers are also about half (0.47) the odds for heavy drinkers but the 95% confidence interval is narrower: (0.29, 0.76).

- (c) The logistic regression predicting low birth weight from both mother's smoking and alcohol consumption:

**Odds Ratios for Categorical Predictors**

Level A	Level B	Odds Ratio	95% CI
alcohol			
medium heavy		0.5664	(0.2966, 1.0819)
light	heavy	0.5020	(0.3093, 0.8146)
light	medium	0.8862	(0.4887, 1.6069)
smoke			
smoker	non-smoker	1.9116	(1.2313, 2.9676)

*Odds ratio for level A relative to level B*

**Analysis of Variance**

Source	DF	Wald Test	
		Chi-Square	P-Value
Regression	3	18.00	0.000
alcohol	2	7.98	0.018
smoke	1	8.34	0.004

The odds ratio for low birth weight, using non-smoking mothers as the baseline, is slightly reduced compared to the single variable model; it is 1.91. The 95% confidence interval is (1.23, 2.97).

In this model, the odds of having a low birth weight baby for moderate drinkers are 0.57 times the odds for heavy drinkers. The 95% confidence interval is (0.30, 1.08). The odds of having a low birth weight baby for light drinkers is half the odds for heavy drinkers (odds ratio is 0.50). The 95% confidence interval is (0.31, 0.81). The odds have slightly increased compared with the single variable model; they are closer to one in this two variable model. The odds ratios in the two variable model are closer to one than the odds ratios in the separate single variable model; however, there still appear to be quite strong independent effects of smoking and alcohol consumption on low birth weight.

14.2 (a) Here is some output:

**Odds Ratios for Continuous Predictors**

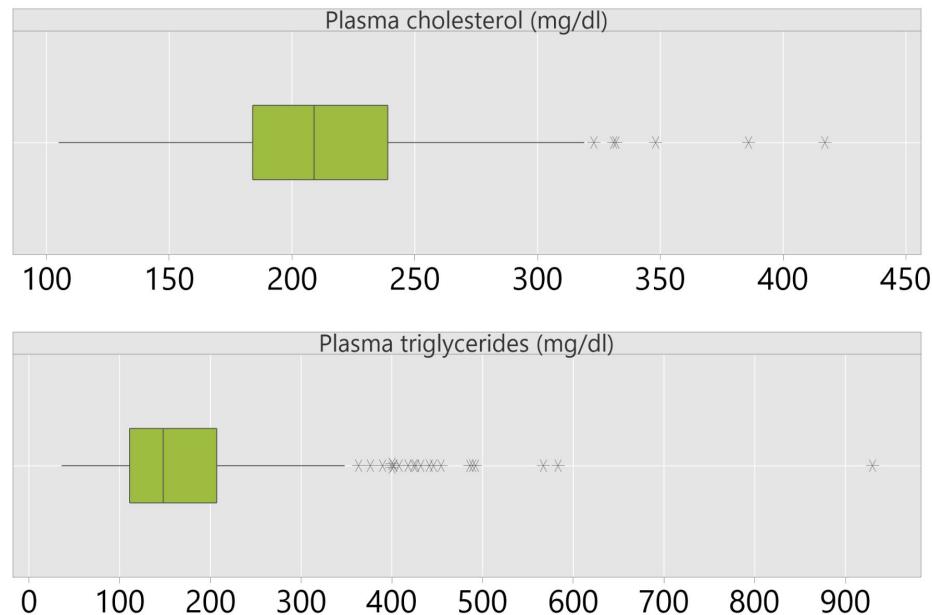
	Odds Ratio	95% CI
Plasma cholesterol (mg/dl)	1.0108	(1.0025, 1.0193)
Plasma triglycerides (mg/dl)	1.0045	(0.9999, 1.0090)

**Analysis of Variance**

Source	DF	Wald Test		
		Chi-Square	P-Value	
Regression	2	13.71	0.001	
Plasma cholesterol (mg/dl)	1	6.50	0.011	
Plasma triglycerides (mg/dl)	1	3.70	0.054	

(b) The odds ratio for plasma cholesterol is 1.01. This means that for each increase of one mg/dl of plasma cholesterol there is a 1.01 increase in the odds of narrowing of the arteries.

- (c) Plasma cholesterol has a range of about 300 and plasma triglycerides has a range of about 900. Both these variables could be re-scaled by dividing by 100.



- (d) The new variables are called Cholesterol/100 (mg/dl) and Triglycerides/100 (mg/dl). The output with the two re-scaled explanatory variables appears below:

**Odds Ratios for Continuous Predictors**

	Odds Ratio	95% CI
Cholesterol/100 (mg/dl)	2.9393	(1.2831, 6.7331)
Triglycerides/100 (mg/dl)	1.5617	(0.9915, 2.4598)

- (e) For every 100 point increase in plasma triglycerides, the odds of narrowing of the arteries increases 1.56 times. The 95% confidence interval for the odds ratio indicates that plausible values for the increase in odds range from 0.99 to 2.46.

For every 100 point increase in plasma cholesterol, the odds of narrowing of the arteries increases nearly three times (odds ratio = 2.94). The 95% confidence interval for the odds ratio indicates that plausible values for the increase in odds range from 1.28 to 6.73.

Note that the re-scaling has not changed the results of the significance tests or the  $P$ -values. However the re-scaling provides more informative odds ratios corresponding to meaningful changes in the explanatory variables.



## 15 Inference — categorical outcome, simple methods

We have already considered statistical inference in relation to categorical outcomes in the following ways:

- finding a confidence interval for a difference of two proportions (Chapter 5);
- fitting a logistic regression model for a binary outcome (Chapter 14).

Logistic regression is a general and useful framework to consider when modelling a binary categorical outcome, as it allows for consideration of both categorical and numerical explanatory variables. It provides hypothesis testing and confidence intervals. There are general models for ordinal outcomes, and categorical outcomes with more than two levels, but these are beyond the scope here.

In this chapter, we consider some methods of hypothesis testing in relation to categorical data. These methods have been popular historically and are not uncommon today. They are:

- Approximate test for comparing two proportions,
- Fisher's exact test, and
- Pearson's  $\chi^2$  test.

When two variables are genuinely categorical, we cannot use the correlation coefficient to measure the degree of association between them. We may want to test for a relationship of some sort. The natural null hypothesis is that there is no (true) association: the two variables are independent of each other. Pearson's  $\chi^2$  test can be used here.

### 15.1 Hypothesis test for comparing two population proportions

Suppose that we have two independent random samples of size  $n_1$  and  $n_2$  from two populations. We want to make inferences about the difference of proportions  $\theta_1 - \theta_2$ , where  $\theta_i$  is the proportion of population  $i$  that has a characteristic of interest.

This setting was considered in Section 5.3.4, where we looked at an approximate confidence interval for  $\theta_1 - \theta_2$ ; the result was derived for large samples, based on the Central Limit Theorem.

As a reminder, the result being used as follows. Let  $X_1 \stackrel{d}{=} \text{Bi}(n_1, \theta_1)$  and  $X_2 \stackrel{d}{=} \text{Bi}(n_2, \theta_2)$ . If  $n_1$  and  $n_2$  are large, then

$$\frac{X_1}{n_1} - \frac{X_2}{n_2} \stackrel{d}{\approx} N(\theta_1 - \theta_2, \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2})$$

and an approximate 95% confidence interval for  $(\theta_1 - \theta_2)$  is given by

$$\frac{x_1}{n_1} - \frac{x_2}{n_2} \pm 1.96 \sqrt{\frac{\frac{x_1}{n_1}(1 - \frac{x_1}{n_1})}{n_1} + \frac{\frac{x_2}{n_2}(1 - \frac{x_2}{n_2})}{n_2}}.$$

This can also be expressed as

$$\hat{\theta}_1 - \hat{\theta}_2 \pm 1.96 \sqrt{\frac{\hat{\theta}_1(1 - \hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1 - \hat{\theta}_2)}{n_2}}.$$

We may wish to carry out an **hypothesis test** of the null hypothesis  $H_0 : \theta_1 - \theta_2 = 0$ . We can use the Normal approximation above to test this hypothesis. In this case there is more than one way to estimate the standard error; we can use the same standard error as above, or we can estimate the standard error assuming that the null hypothesis is true, which gives a slightly different answer.

There is a good statistical argument for estimating the standard error assuming that  $H_0$  is true. However, this is not what MINITAB does by default. If we use the default MINITAB approach (more on this below) we find the ratio

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\frac{\hat{\theta}_1(1 - \hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1 - \hat{\theta}_2)}{n_2}}}$$

and compare this ratio with the standard normal distribution.

This test is one of the familiar types, in which an estimate divided by its standard error is compared to an appropriate reference distribution.

▷ **EXAMPLE.** A survey of musicians finds that among 73 pianists, the number that had playing-related musculo-skeletal disorders (PRMD) was 37. Among the 48 saxophone players in the survey, the number with PRMD was 14.

These data give  $\hat{\theta}_1 = 37/73 = 0.507$  and  $\hat{\theta}_2 = 14/48 = 0.292$ , so that  $\hat{\theta}_1 - \hat{\theta}_2 = 0.507 - 0.292 = 0.215$ . Using the above method, we find that the standard error of  $\hat{\theta}_1 - \hat{\theta}_2$  is 0.088, and hence an approximate 95% confidence interval for  $\hat{\theta}_1 - \hat{\theta}_2$  is  $0.215 \pm 1.96 \times 0.088$ , or (0.043 to 0.387).

We might quote this on the percentage scale, as an estimated difference in the prevalence of PRMD of 22%, with a 95% confidence interval of (4%, 39%).

Consider the comparison of the pianists and the saxophonists. We find that  $z = 0.215/0.088 = 2.45$ . When we compare this to the  $N(0, 1)$  distribution we find that  $P = 2 \times \Pr(Z \geq 2.45) = 0.014$ .

## Software

In MINITAB, Stat → Basic Statistics → 2-Proportions ... will provide an approximate 95% confidence interval and a test of the null hypothesis that there is no difference in the proportions. As described above, the result provided by MINITAB uses the normal approximation. You might note that under Options you can choose to “Use pooled estimate of  $p$  for test”. This assumes  $H_0$  is true and it is better to do this.

The default calculation does not use the pooled estimate for the test. Using the pooled estimate gives  $z = 2.35$  and  $P = 0.019$ , as shown in the following Minitab output.

### Test and CI for Two Proportions

#### Method

$p_1$ : proportion where Sample 1 = Event

$p_2$ : proportion where Sample 2 = Event

Difference:  $p_1 - p_2$

#### Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	73	37	0.506849
Sample 2	48	14	0.291667

#### Estimation for Difference

Difference	95% CI for
	Difference
0.215183	(0.042883, 0.387483)

CI based on normal approximation

#### Test

Null hypothesis  $H_0: p_1 - p_2 = 0$

Alternative hypothesis  $H_1: p_1 - p_2 \neq 0$

Method	Z-Value	P-Value
Normal approximation	2.35	0.019
Fisher's exact		0.024

The test based on the normal approximation uses the pooled estimate of the proportion (0.421488).

## 15.2 Fisher's exact test

It is important to keep in mind that the results considered above are “large sample approximations”. There is no straightforward way to deal with small sample sizes when finding a 95% confidence interval for  $\theta_1 - \theta_2$ .<sup>31</sup>

<sup>31</sup>A number of alternative approximations are considered in Newcombe, R.G. (1998) Interval estimation for the difference between independent proportions: Comparison of eleven methods, *Statistics in Medicine*, 17, 873-890.

Sometimes we do have small samples, and are still interested in testing the null hypothesis  $H_0 : \theta_1 - \theta_2 = 0$ .

Another way of thinking about the testing of  $H_0 : \theta_1 - \theta_2 = 0$  is to construct what is called a **contingency table** or a **cross-tabulation**.

We look at this approach in the musicians' data. We can make a  $2 \times 2$  table to compare the pianists and the saxophonists by classifying them according to their instrument, and, separately, according to whether or not they had PRMD. The table looks like this:

	Instrument		Total
	Piano	Saxophone	
PRMD	37	14	51
no PRMD	36	34	70
Total	73	48	121

Fisher's exact test uses the idea that the margins (the row and column totals) are non-informative about  $\theta_1 - \theta_2$ , or equivalently, about the association between instrument (piano/saxophone) and PRMD (yes/no).

This leads to the use of the hypergeometric distribution for the frequency in one of the 4 cells in the body of the table, if the null hypothesis is true. The hypergeometric distribution is a discrete distribution we have not considered previously; it arises in sampling from a finite population without replacement.

Define  $T = X_1 + X_2$ , where  $X_1$  and  $X_2$  are the binomial counts. Given the observed value of  $T$ , which we denote by  $t$ , the probability that  $X_1 = x_1$  is given by

$$\Pr(X_1 = x_1 | T = t) = \frac{\binom{t}{x_1} \binom{n_1+n_2-t}{n_1-x_1}}{\binom{n_1+n_2}{n_1}}$$

where the limits of  $x_1$  are determined by the margins of the table.

The  $P$ -value is then found by applying its general definition:  $P$  = probability of a result at least as extreme as that observed, given that the null hypothesis is true.

For Fisher's exact test, this entails calculating the probability of all possible tables, and summing the probabilities of tables whose probability is less than or equal to probability of the observed table.

For the musicians' data, Fisher's exact test is given in the output above, and is 0.024.

▷ **EXAMPLE.** Florey's first penicillin experiment

In May 1940, Florey and his team trialled penicillin in an experiment on eight mice. They first injected all mice with lethal doses of streptococci (bacteria). Four of the eight mice also got the treatment – penicillin; two received a single injection, and two received five injections. Florey is reported to have phoned his colleague, Margaret Jennings and said, "It looks like a miracle". Here are the miraculous results:

Group	<i>n</i>	Survival details
Controls	4	All died within "a matter of hours"
Penicillin	1 injection	2 days; 6 days
	5 injections	2 13 days; indefinitely

Considering the outcome of survival beyond 1 day only, the  $2 \times 2$  table is

Group	Survival		Total
	no	yes	
Controls	4	0	4
Penicillin	0	4	4
Total	4	4	8

For these data, the key Minitab output is shown below. Note the following:

- The point estimates of the proportions and the difference of proportions are correct; obviously in these data the difference is as extreme as it can be.
- The approximate confidence interval gives the result (\*, \*), which reflects the inadequacy of any 'large sample approximation' for these data.
- The approximate test is given with a caution about small sample sizes.
- The *P*-value for Fisher's exact test is  $P = 0.029$ .

A common question is: Why use the large sample approximation if Fisher's exact test is available? The answer is that there is no reason to do so: you should use Fisher's exact test if it is available.

## Test and CI for Two Proportions

### Method

$p_1$ : proportion where Sample 1 = Event  
 $p_2$ : proportion where Sample 2 = Event  
 Difference:  $p_1 - p_2$

### Descriptive Statistics

Sample	N	Event	Sample p
Sample 1	4	0	0.000000
Sample 2	4	4	1.000000

### Estimation for Difference

Difference	95% CI for Difference	
	-1	(*, *)

CI based on normal approximation

### Test

Null hypothesis	$H_0: p_1 - p_2 = 0$	
Alternative hypothesis	$H_1: p_1 - p_2 \neq 0$	
Method	Z-Value	P-Value
Normal approximation	-2.83	0.005
Fisher's exact		0.029

The test based on the normal approximation uses the pooled estimate of the proportion (0.5).  
 The normal approximation may be inaccurate for small samples.

## 15.3 $\chi^2$ test for $r \times c$ table

So far in this chapter the setting considered entails the simplest possible comparison of categorical data: two sample proportions. This can be represented as a  $2 \times 2$  table, as we saw.

However, we sometimes have categorical variables with more than two levels, and we may be interested in the association between them. This gives an  $r \times c$  table;  $r$  is used for the number of rows, and  $c$  for the number of columns. Such a table is called a **contingency table**. “Contingent” means “dependent”, so we consider the table to look for evidence of dependence between the two variables.

A very old statistical test used to test for the association between two categorical variables is **Pearson’s  $\chi^2$  test**.

Agresti, in his book *Categorical Data Analysis* quotes the author Doolittle (1887) who described the question raised by a general  $2 \times 2$  contingency table, as follows:

“Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so,

in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things."

Exactly.

▷ **EXAMPLE. Incidence of cerebral tumours** (cerebral.mwx)

Site	Type			Total
	Benign	Malignant	Other	
Frontal	23	9	6	38
Temporal	21	4	3	28
Other	34	24	17	75
Total	78	37	26	141

Figure 94 shows two different dot charts of the data. The null hypothesis  $H_0$  is that there is no association between the two variables type and site; the alternative hypothesis  $H_1$  is that there is an association of some sort.

The notation we use is:

- $r \times c$  table : one with  $r$  rows and  $c$  columns
- $O_{ij}$  : the observed frequency in the  $ij^{th}$  cell (row  $i$ , column  $j$ )
- $E_{ij}$  : the expected frequency in the  $ij^{th}$  cell, assuming  $H_0$  is true
- $N$  : the total number of observations

The entries in the cells are regarded as "observed" frequencies. The test is based on a comparison of these with the frequencies we would have "expected" if the null hypothesis is true. Consider the top left hand cell, which has an observed frequency of 23. How many would we "expect" there, on average, if there was really no association between Site and Type?

We reason as follows: In the example, there were 78 benign tumours. This was 55.3% ( $= 78/141$ ) of all the tumours. If  $H_0$  is true, on average, this 55.3% should apply within each row of the table. That is, the overall percentage should apply in each row. If the percentage of benign tumours depends on the site of the tumour, that is an association between site and tumour, and  $H_1$  is then true.

So we calculate "expected frequencies", assuming  $H_0$  is true, by applying the overall fractions or percentages within each row. For example, assuming independence, the expected frequency for the top left-hand cell is

$$21.02 = \frac{78}{141} \times 38.$$

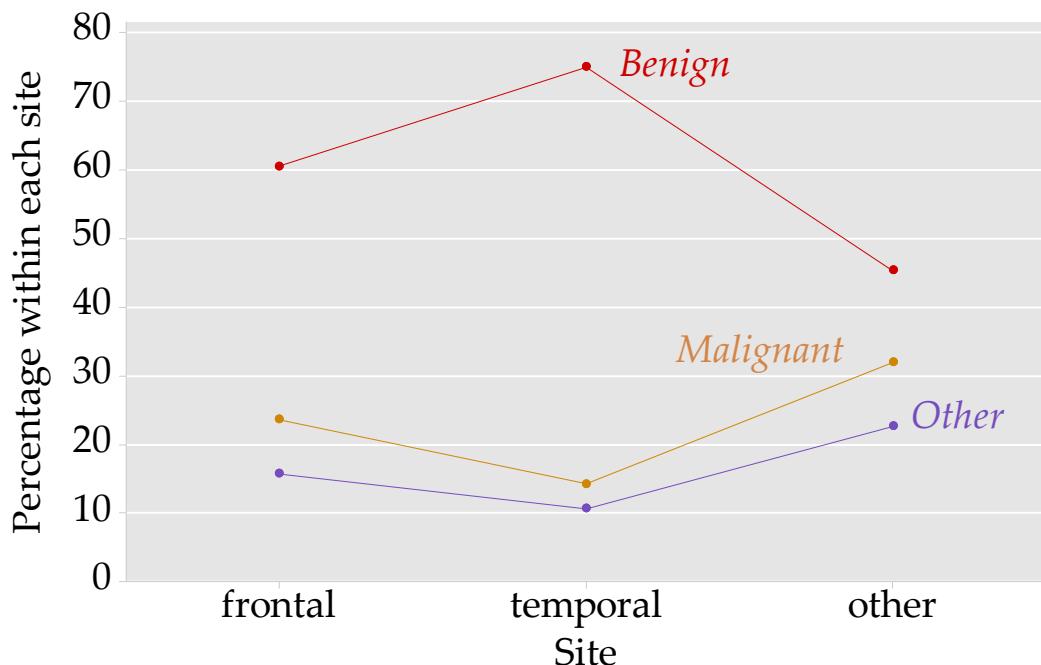
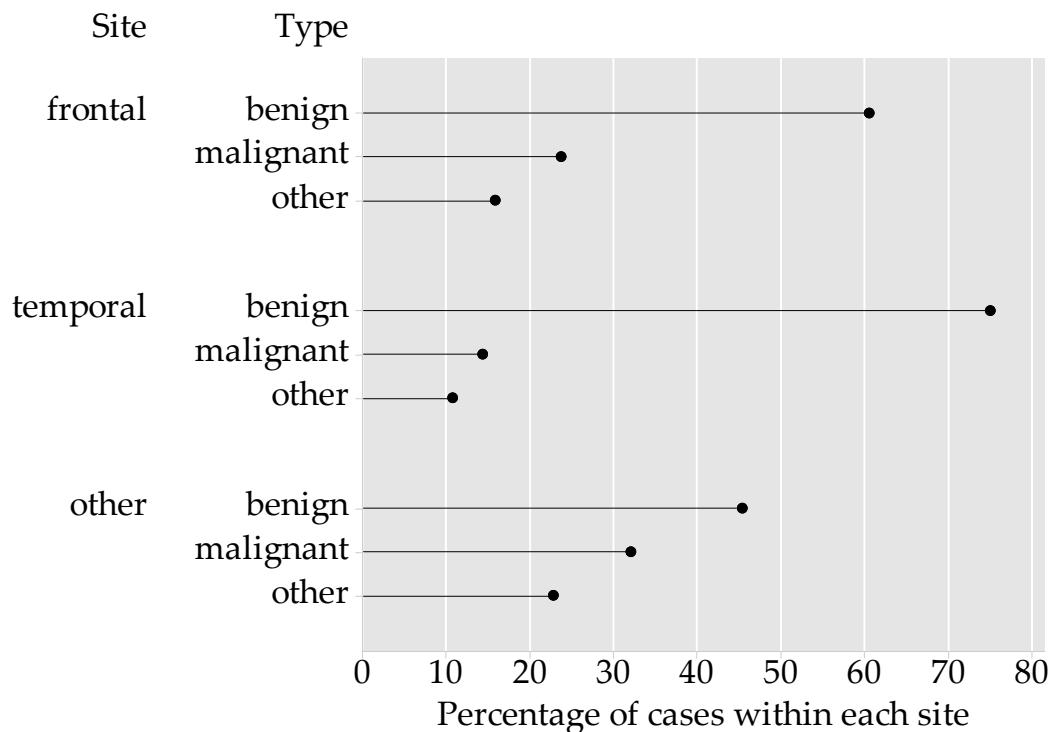


Figure 94: Two different dot charts showing the cerebral tumour data

We use the terms “expect” and “expected”. Even if  $H_0$  is true, we do not literally expect to see 21.02 frontal benign tumours; it has to be a whole number. We use the term “expected” to denote the long-term, average value, assuming that  $H_0$  is true, and imagining an infinite repetition of the experi-

ment with the same margins in the table.

Of course, we could have applied the argument in the other direction, but we would get exactly the same expected frequency. We could have argued that if  $H_0$  is true, the overall fraction of frontal sites should apply in each column, and hence the expected frequency in the top left-hand cell is (again)

$$21.02 = \frac{38}{141} \times 78.$$

Either way, the general expression we get is that:

$$\text{expected frequency in row } i \text{ and column } j = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{N},$$

where  $N$  is the total number of observations.

We calculate the following test statistic:

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Note that  $X^2$  will be large when the observed and expected frequencies differ a lot.

If the null hypothesis is true,  $X^2 \stackrel{d}{\approx} \chi^2_{(r-1)(c-1)}$ , and if  $H_1$  is true, the test statistic is inclined to be larger. It therefore has the properties than any test statistic needs: a known distribution (at least approximately) if the null hypothesis is true, and a signal, a shift in the distribution, if the alternative is true.

So we calculate the observed value  $x^2$  and compare it to the  $\chi^2$  distribution with  $(r-1)(c-1)$  degrees of freedom.

▷ **EXAMPLE.**  $x^2 = 7.84$ ; cf.  $\chi^2_4, P = 0.097$ .

This test is based on a “large sample” approximation. It involves normal approximations to discrete distributions. Usually, rules of thumb are given for what constitutes a large enough sample for the approximation to be adequate. Importantly, these rules are based on the “expected” frequencies and not on the observed ones.

- (a) No expected frequency should be less than 1.
- (b) The percentage of cells with expected frequencies less than 5 should be no greater than 20%.

It is important to be alert to these rules. This is because the way in which the  $\chi^2$  test fails, if the expected frequencies are too small but the test is used anyway, is that the  $P$ -values obtained are too small. This is called an “anti-conservative” bias, which we want to avoid.

## Software

There are two ways to obtain the  $\chi^2$  test in MINITAB.

If the data are in long form, use Stat → Tables → Cross-tabulation and Chi-square .... See cerebral.mwx for example.

If the data are stored as a table in the worksheet, you can also use: Stat → Tables → Cross-tabulation and Chi-square ... and then choose the option Summarised data in a two-way table.

The  $\chi^2$  test is a “global” test. In that sense, it does not go very far, and it does not lead easily to parameter estimates and confidence intervals.

A small  $P$ -value is evidence against  $H_0$ , but does not indicate where the lack of independence arises. We need to inspect the table of observed and expected frequencies, and the contributions of the individual cells to the overall test statistic. For this reason, MINITAB provides exactly this information. In the cerebral tumours case, the  $P$ -value was 0.10. It is on the small side. The MINITAB output shows:

	Rows: Site	Columns: Type			
		benign	malignant	other	All
frontal		23	9	6	38
		21.02	9.97	7.01	
		0.1863	0.0947	0.1447	
temporal		21	4	3	28
		15.49	7.35	5.16	
		1.9605	1.5251	0.9063	
other		34	24	17	75
		41.49	19.68	13.83	
		1.3519	0.9479	0.7267	
All		78	37	26	141
<i>Cell Contents</i>					
<i>Count</i>					
<i>Expected count</i>					
<i>Contribution to Chi-square</i>					

Within the first row, the expected and observed frequencies are very close. The slight departure from independence occurs in the second row; in particular, the observed number of benign temporal tumours (21) was rather more than expected, assuming  $H_0$ . The contribution to the overall test statistic from that cell was 1.96.

## 15.4 Other methods

In this chapter, we have demonstrated some simple hypothesis testing procedures for categorical outcome. There are statistical methods for categorical data arising from correlated samples, and, as mentioned for modelling ordinal outcomes or nominal outcomes. These are substantial areas of study in themselves.

## 15.5 Exercises

- 15.1 *Understanding that people have beliefs about the world allows us to anticipate what other people might think or do. Some psychologists have argued that autistic people fail to understand that people have beliefs and so have difficulty in social situations.*

A test for children involves two dolls. One doll, Sally, hides a marble. A second doll, Anne, moves the marble when Sally is out. Sally returns and the child is asked “Where will Sally look for her marble?” Correct answers refer to Sally’s original hiding place — they reflect an understanding that Sally believes the marble to be where she put it. Baron-Cohen, Leslie & Frith (1985) used this Sally-Anne test with a sample of 20 autistic children and a second sample of 27 non-autistic children. Their results were:

Response to question	Autistic	Not autistic
Correct	4	23
Incorrect	16	4
Total	20	27

The data are in mind.mwx.

- (a) Stat > Tables > Cross tabulation and Chi square to reproduce the table above. Additionally obtain the percentages answering correctly in each group.
  - (b) Carry out a test of the null hypothesis of no association between group and response to the question.
  - (c) Summarise the differences between the groups and the statistical inference you have made.
- 15.2 In a study on the effect of stress, 300 subjects were divided into four groups of 75, at random, and required to attempt a task under various degrees of stress. The subjects’ attempts were classified as successful, partially successful or unsuccessful.

The results of the study are summarised as follows:

Program	Stress level			
	None	Low	Medium	High
Successful	48	42	39	30
Partially Successful	17	18	18	16
Unsuccessful	10	15	18	29

- (a) Carry out a test of whether stress affects the subjects’ ability to succeed at the task.

[Enter the data into 4 columns in the Data Window. If you enter the data into

C1-C4, then C1 should contain the values 48, 17, and 10, C2 should contain the values 42, 18 and 15, etc.

The analysis can be carried out using Stat > Tables > Chi-square Test for Association, then choose Summarized data in two-way table; select C1-C4; click OK.]

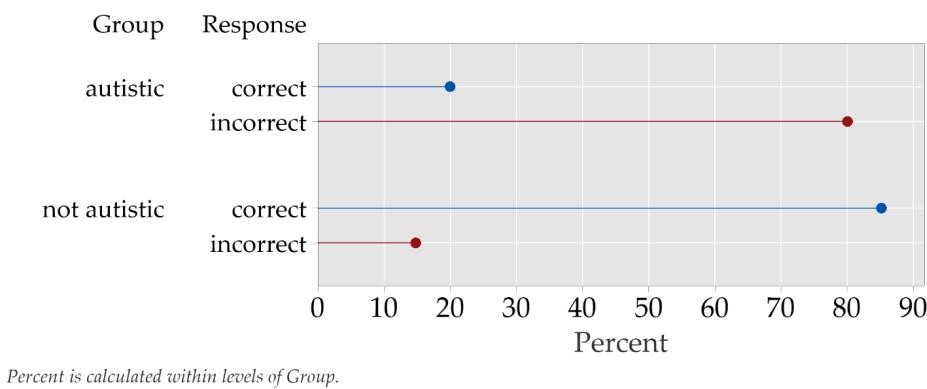
- (b) What inference can be drawn from the results?

## 15.6 Answers

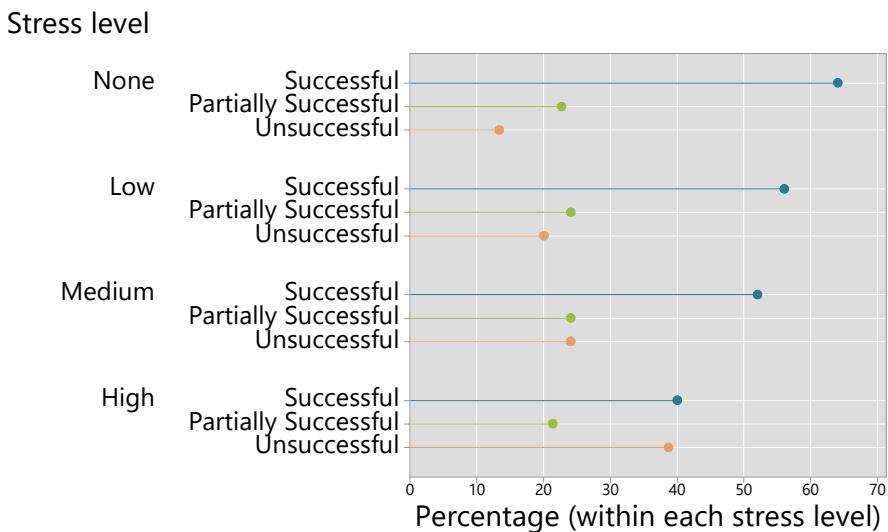
- 15.1 (a) The output from Stat > Tables > Crosstabulation and Chi-square:

	autistic	not autistic	All
correct	4 20.00	23 85.19	27 57.45
incorrect	16 80.00	4 14.81	20 42.55
All	20 100.00	27 100.00	47 100.00
<i>Cell Contents</i>			
<i>Count</i>			
<i>% of Column</i>			

Here is an example of a graph of the data:



- (b) The  $P$ -value for Fisher's exact test is very small,  $P < 0.001$ . The pattern of responses observed is not consistent with the null hypothesis of no association.
- (c) Only 20% of the autistic children answered correctly, while less than 20% of the non-autistic children answer incorrectly. The estimated difference in the percentage answering correctly is 65.2%. Note that it is also useful to provide a confidence interval for the the difference in percentage answering correct; the 95% confidence interval is (37.3%, 80.1%).
- 15.2 (a) The result for Pearson's  $\chi^2$  test of no association is  $\chi^2_6 = 15.18$ ,  $P = 0.019$ . Here is an example of a graph of the data:



Note that when there is an ordering of the groups (e.g. stress: none to high) you can carry out a Chi-square test for trend (a trend in the proportions across the ordered groups). There is a good discussion, with examples, in “Practical Statistics for Medical Research” by Douglas Altman (1991).]

- (b) The findings suggest a relationship between stress levels and success. As the stress level increases, the percentage that are successful declines and the percentage that are unsuccessful increases.



# 16 Study planning and design

Good design is the foundation for making valid inferences. An understanding of the principles of study design is a fundamental aspect of knowledge of a good analyst. No amount of sophisticated analytic work can undo the limitations arising from poor design.

“You can’t fix by analysis what you bungled by design.” Light, Singer and Willett (1990, page v)

The word ‘design’ can be used in a broad sense, defining the method or structure of the data collection. It has many aspects that we do not cover here, including defining the research question clearly, dealing with the management of the research, ethics, measurement protocols, ensuring integrity of the data collection, cleaning and storage, and data analysis plans. These are all important things. The design of surveys, the design of experiments, and the design of many other aspects of an empirical study are courses of study in themselves.

In this chapter we touch on two aspects of design — fundamental concepts in design of experiments and approaches to determining sample size in study planning.

## 16.1 Design of experiments

Any data collection that comes from a context in which an intervention has been used to change an outcome in some way, can be regarded as an experiment. An experiment usually and desirably involves a comparison between two or more interventions. There is a rich body of theoretical understanding of what makes a good experiment and in this section we briefly introduce some important principles in experimental design.

The main concepts in experimental design are:

- randomization
- blocking
- replication

### 16.1.1 Randomization

Subject to restrictions imposed by the type of design being used, treatments should be allocated to units (experimental units) at random. This guards against possible bias and helps to justify the assumptions that are usually made when the results are analysed.

Randomization is very strongly established as a desirable feature in practice, especially in some disciplines. In evaluating the efficacy of a new drug, for example, randomized trials get far more evidentiary weight than non-randomized trials, and regulatory bodies tend to dismiss the evidence from non-randomized trials altogether.

Random allocation of treatments to units in experiments allows valid conclusions about causation.

We have groups with different interventions, and we measure an outcome of interest. Suppose that the results in the two groups differ. Why did that happen? What caused the difference? If allocation to the groups was random, then the strongest explanation is the different interventions applied. This is an incredibly powerful concept and technique.

But what does “differ” mean here? Intuitively, we are imagining the results differing substantially. More precisely, it means that the difference between the groups is large, relative to natural variation. The way that distributions of variables behave when they are randomly split into groups is well understood. Hence this applies to outcomes in randomised studies. The reasoning behind analysing a randomised trial, after the intervention has had its opportunity, is as follows:

Do the groups look like the outcomes were distributed randomly? If so, data are consistent with an ineffective intervention. If not, the intervention must have caused this. Without randomisation, but an apparent effect, we are always left wondering:

Did the intervention cause the effect, or was it something else? Randomisation ensures that the groups are balanced, on average, on other possible causes of the outcome. Researchers sometimes try to do this in other ways, by matching on key variables (age, sex). Randomisation does this automatically; groups are balanced on known and unknown other causes of the outcome.

The random allocation of treatments means that the different treatments are the strongest explanation of markedly different group results. But this is not the only possible explanation. There could be some small amount of imbalance between the groups that is not completely eradicated by randomisation; this is known as “residual confounding”. For large sample sizes and big differences in the results, residual confounding becomes less and less plausible.

### 16.1.2 Blocking

A **block** is a collection of units which are considered likely to be more homogeneous than the entire collection of available units. For example, in an experiment to compare the effect of various diets on the weight gain of pigs,

animals from the same litter would constitute a reasonable block.

The purpose of blocking is to reduce the variance of the random error by accounting for some of the variation between units. This reduction can be quite substantial thereby greatly increasing the precision of the experiment. Blocks differ from factors in that it is generally assumed (hoped) that blocks and factors do not interact.

Much of experimental design is concerned with ways for dealing with different types of blocking.

We have already seen examples of blocking, particularly in Section 12.1, where we dealt with a randomised blocks design. Pairing is also an example of blocking.

### 16.1.3 Replication

Using the same treatment (factor combinations) a number of times not only increases the precision of estimators it also enables us to estimate the precision by estimating the error variance. Without such an estimate, formal inference (confidence intervals, hypothesis testing) would not be possible.

## 16.2 Some common designs

### 16.2.1 Completely randomized design

For all of the designs considered below it is possible for the ‘treatments’ to be combinations of two or more factors and for the treatment effects to be partitioned into main effects and interactions between the factors, that is, for the treatments to have a factorial structure, such as that considered in Section 12.2. However, for simplicity, such partitions will not be considered here.

In a completely randomized design, treatments are assigned to units at random, subject only to the number of times each treatment is to be used.

▷ **EXAMPLE.** If an intervention is applied to people, and there is no clear basis for discriminating between individuals in terms of their likely response, then a completely randomized design may be used.

Drug trials are quite often conducted in this way: if there is a pool of eligible subjects, the subjects may be allocated at random to one of the ‘arms’, or treatment groups, of the trial. There are a variety of ways that this is done in practice, using random number tables or computer programs, but they all entail the feature that each subject has the same probability of being allocated to each group.

**Analysis:** As a general linear model with a categorical explanatory variable

(one-way ANOVA).

### 16.2.2 Randomized block design

This design requires that each treatment be used the same number of times (usually once) in each block. Within each block, treatments are allocated to units at random with a separate randomization for each block. We have covered this in Section 12.2.

▷ **EXAMPLE.** In the early historical applications of this design in agriculture, the blocking factor was an area of land; hence the name “block”. This remains an important context; areas of land may vary in fertility, soil content and so on, so it is useful to exploit this in the design, to make a more efficient inference about the treatments.

In educational research on school children, say, a comparison of two strategies for teaching about safety, the blocking factor might be age, or school year.

**Analysis:** As a general linear model with two categorical explanatory variables (two-way ANOVA). It is usually assumed that treatments and blocks do not interact. If each treatment is used more than once in each block it is possible to test the validity of this assumption.

### 16.2.3 Latin square design

Allows for two-directional blocking (e.g. allows for both ‘row’ and ‘column’ blocking in agricultural experiments). The design requires that there be equal numbers of rows, columns and treatments, and that each treatment be used exactly once in each row and each column. Randomization is achieved by choosing a ‘square’ at random from all squares which satisfy the above requirements.

An example of a Latin square is shown in Figure 95.<sup>32</sup> Note the tendency for the trees to look healthy in the rows higher up on the hill (a block effect) and clear differences between the types of tree.

---

<sup>32</sup>The original source of this image is difficult to identify; it is found in many places on websites.

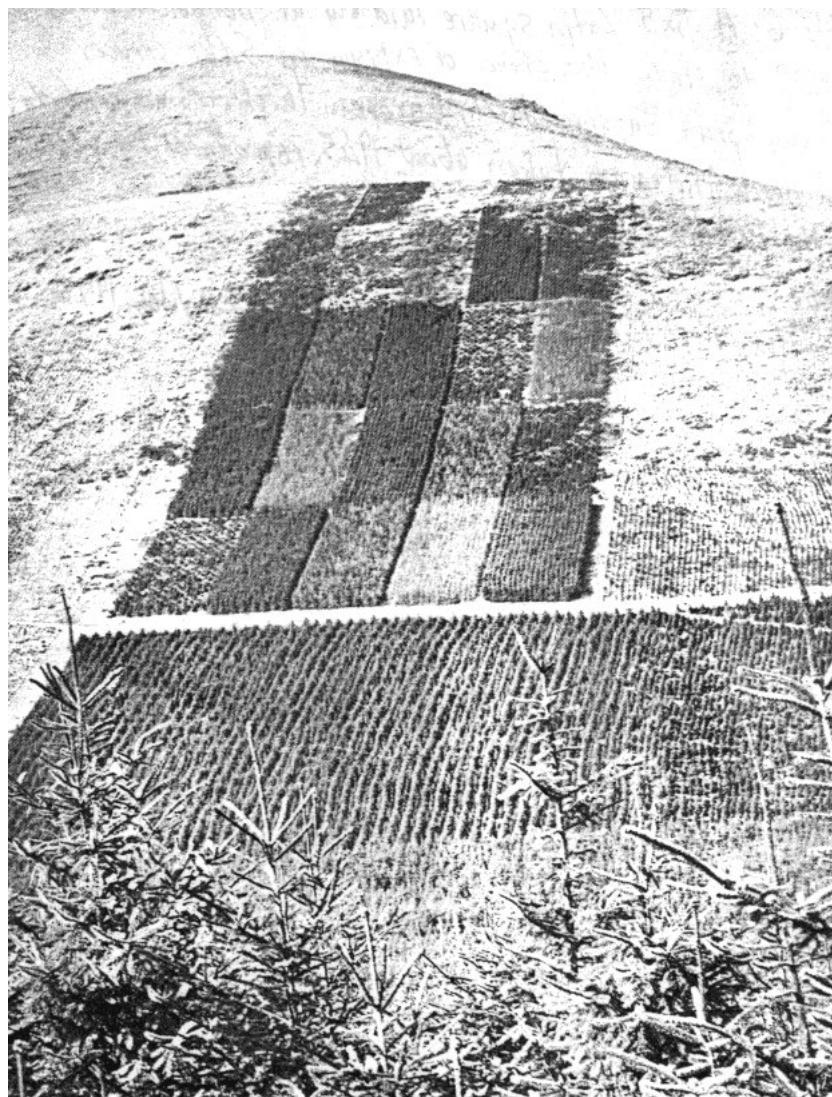


Figure 95: A  $5 \times 5$  forestry experiment in Beddgelert in Wales, to compare varieties of tree; designed by R.A. Fisher, laid out in 1929, and photographed in about 1945.

**Analysis:** A simple extension of a two-way ANOVA.

▷ **EXAMPLE. Oats yield (oats.mwx)**

Four varieties of oats were compared using a  $4 \times 4$  Latin square design to take account of possible fertility gradients in the soil. The yields (in kg) per plot were as follows, where the numbers 1, 2, 3 and 4 refer to the varieties.

47	40	50	57
3	4	2	1
49	53	37	29
2	1	3	4
28	34	46	37
4	3	1	2
48	44	25	30
1	2	4	3

The data are shown in Figure 96.

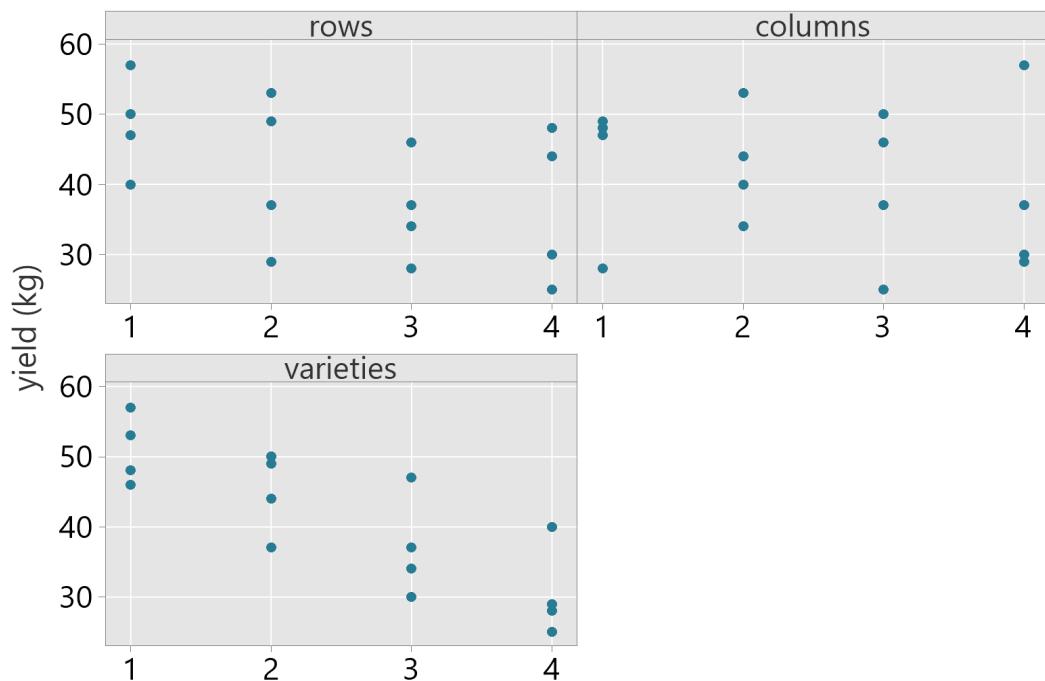


Figure 96: Scatter plots of the yields by row, by column and by variety, Latin Square example

Means			
	Rows	Columns	Varieties
1	48.5	43.0	51.0
2	42.0	42.8	45.0
3	36.3	39.5	37.0
4	40.9	38.3	30.5

## ANOVA

Source	df	SS	MS	F	P
Rows	3	391.25	130.42	92.18	< 0.001
Columns	3	67.25	22.42	15.8	0.003
Varieties	3	968.75	322.92	227.9	< 0.001
Residual	6	8.50	1.42		
Total	15	1435.75			

We conclude that there is a highly statistically significant difference between varieties ( $P < 0.001$ ) and that the Latin square was very effective in increasing the precision of the experiment as both the row and column effects are highly statistically significant.

It is useful to make inferences about the pairwise comparisons for varieties arising from this analysis; estimates and 95% confidence intervals are shown in Figure 97.

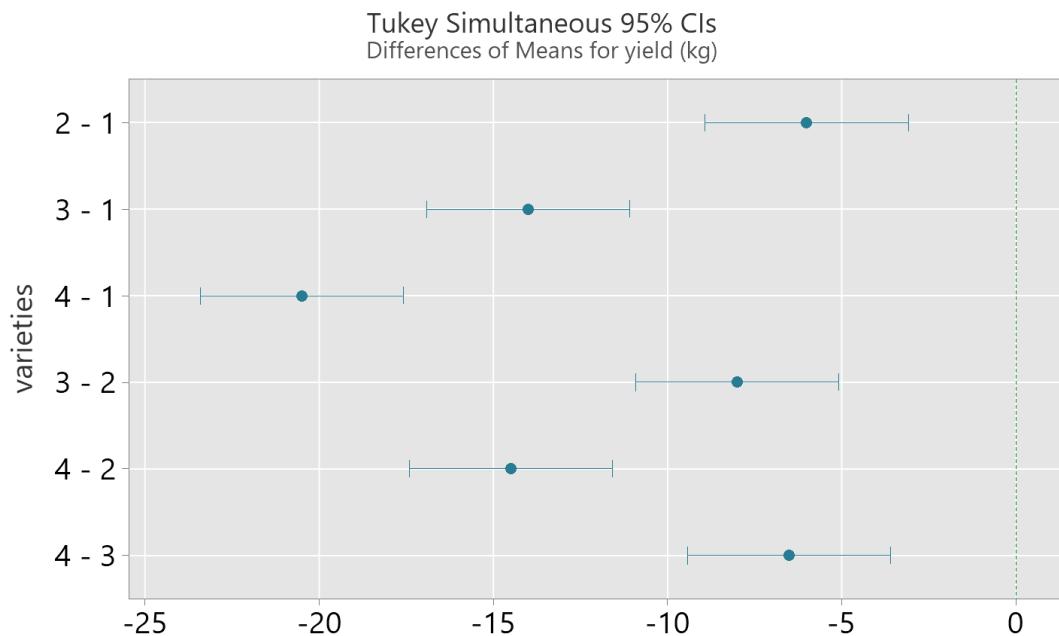


Figure 97: Estimates and 95% confidence intervals for the pairwise differences between varieties.

#### 16.2.4 Extensions

Formal experimental design is a major subject that has attracted a great deal of attention. Aspects of it intersect with the theory of combinatorics, in pure mathematics.

Here is a list of other types of designs, with brief information about them.

- Incomplete block design: similar to a randomized block design but with fewer units per block than the number of treatments. The design requires that there be the same number of units per block, that each treatment be used the same number of times and that each pair of treatments appear together in a block the same number of times.
- Latin rectangle, also known as a 'Youden square': similar to a Latin square but with different numbers of levels for the two blocking variables.
- Graeco-Latin square: three blocking variables, and a treatment variable, all with the same number of levels.
- Response surface design: used sometimes when the levels of a design factor are numeric, or at least ordered.
- Fractional factorial designs: used when there are a large number of experimental factors of interest, but the cost of an experimental unit is high, for example, settings for the factors of a whole production system.
- Designs with more than one level of variation, such as split plot design, cluster randomised trials, stepped wedge designs.

### 16.3 Study size

One of the decisions that needs to be made at the start of a study is the choice of sample size. In practice, it is often regarded as the most important design matter, but that is a misguided view.

In determining the sample size,

“On the first level the decision is a binary one—the choice between zero and non-zero as the size.” (Miettinen)

We need to work out a sample size:

(a) To get the precision right:

- Studies can be too big: precision gained is not worth the cost.  
(Soil management technique A is better than B by 1% (95% confidence interval: (0.4%, 1.6%)): study costs \$megabucks.)
- Studies can be too small: poor precision, uninformative (also not worth it!). We estimate that A is better than B by 50%, but the 95% confidence interval is (-20%, 120%).
- Studies need to be just right (in size).

(b) Because of ethical issues:

- Researchers have a responsibility not to waste the contribution that subjects make to their work, by carrying out a study that is too small or too large.
- The allocation of resources: when research is funded by someone else, they are entitled to ask: "What are we going to learn for this much expenditure of the research budget?"

(c) Because it concentrates the mind on important study objectives.

(d) Grant bodies require it.

In this chapter, we present two approaches to determining a sample size when planning a study. The first is based on consideration of the precision of a confidence interval; the second is in the framework of hypothesis testing.

## 16.4 Use of confidence intervals to determine sample size

The simplest way to choose a sample size is to consider the width of confidence intervals. This means that, at the design stage of the study, we think about how precisely we wish to estimate an unknown parameter, and we express this precision in terms of a desired width for the confidence interval.

This strategy is most appealing when estimation is the inferential goal, and hence is commonly considered in surveys. However, it can be used in experiments or situations where comparisons are made, since estimation should feature there too.

The basic idea is to set the size of the margin of error for the confidence interval at a desired value. Recall that a symmetric confidence interval can be expressed as

$$\text{estimate} \pm \text{margin of error.}$$

The margin of error depends on the standard error of the estimate, and that, in turn, depends on the sample size.

In application, the desired margin of error is a design choice; it is not obtained from a formula.

### Estimation of a population mean

For the case of estimating the mean of a (Normal) population, a 95% confidence interval is given by  $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$  so that equating  $1.96 \frac{\sigma}{\sqrt{n}}$  to some appropriate value  $m$  will give a suitable value for  $n$ . Note that this process is

setting a value for the desired margin or error,  $m$ . Solving the equation and making  $n$  the subject gives

$$n \geq \left(1.96 \times \frac{\sigma}{m}\right)^2.$$

For example, suppose we are interested in the mean fuel consumption (litres/100km) of a particular brand of car and suppose we know that  $\sigma$  is about 0.3. Then if a 95% confidence interval of width  $\pm 0.1$  seems reasonable, the required sample size is (about) 35 [  $= (1.96 \times \frac{0.3}{0.1})^2$ . ]

How do we arrive at the choice of a margin of error of  $m = 0.1$ , and hence a confidence interval that will be  $\bar{x} \pm 0.1$ ? Ultimately, this is not a statistical question. It is a question for the researcher to answer. In this case, however, the reasoning might be something like this: "I am going to report the average fuel economy of this car in a brochure, or a car magazine, to one decimal place (8.1, e.g., or 7.6 litres/100km). Given that, a 95% confidence interval of  $\pm 0.1$  (e.g.  $8.1 \pm 0.1$ ) is precise enough for my purposes."

Note that since the width of the confidence interval is proportional to  $\frac{1}{\sqrt{n}}$ , halving the width requires a four-fold increase in sample size. This means that the width of a confidence interval is not very sensitive to changes in sample size so that what is needed is a 'ball-park' figure rather than an exact value. If  $\sigma$  is unknown it will be necessary to obtain an estimate of it, either from prior knowledge (e.g. published papers) or by conducting a small pilot study.

### Estimation of a population proportion

Proportions are often thought of in terms of percentages instead: a percentage of 35% rather than a proportion of 0.35. It all depends on the scale used, 0 to 1 (proportions) or 0 to 100 (percentages). Importantly, all the formulae here assume that we are making an inference on a proportion.

For the case of estimating a population proportion,  $\theta$ , it is possible to use the properties of the variance of the estimator,  $\frac{X}{n}$ , to find an upper limit for the required sample size. The margin of error  $m$  of an approximate 95% confidence interval for  $\theta$  is given by

$$m = 1.96 \sqrt{\frac{\theta(1-\theta)}{n}}.$$

Solving this equation gives

$$n \geq \frac{3.84\theta(1-\theta)}{m^2}. \quad (3)$$

This looks very much as if it begs the question, since the unknown  $\theta$  is in the equation! When we were obtaining the confidence interval from data, we

could substitute the estimate  $\hat{\theta}$ . We cannot do that here as we do not have the data; we are at the design stage.

But note that  $\theta(1 - \theta)$  has a maximum value of  $\frac{1}{4}$  which occurs when  $\theta = 0.5$ . So it is possible to do a “worst case” calculation.

The following table gives the sample sizes required for estimating a proportion, based on specifying the width of the 95% confidence interval, and allowing for the possibility that the true value may be at or near  $\theta = 0.5$  (worst case).

$2m = \text{width of 95\% CI}$	0.01	0.02	0.03	0.04	0.05	0.10	0.2
i.e. $\pm m$ , where $m =$	0.005	0.01	0.015	0.02	0.025	0.05	0.1
$n$	38416	9604	4268	2401	1536	384	96

Note, further, that  $3.84 \times \frac{1}{4}$  is roughly 1, so to work out the figures in the above table approximately, one says that for a 95% confidence interval with a margin of error  $m$ , and hence width at most  $2m$ , a sample size of about  $\frac{1}{m^2}$  is required. For example, for  $m = 0.1$  a sample size of about 100 is required and for  $m = 0.01$  a sample size of about 10 000 is required.

The Excel workbook: Confidence intervals for sample size.xls also allows you to find sample sizes for single sample proportions using this worst case approach. You should use the sheet labelled Population proportion - worst, and enter the desired confidence interval width. Remember that this is  $2m$ .

▷ **QUESTION:** The publication of the results of opinion polls in the press is sometimes accompanied by claims about precision of the results. For example, in the January 26, 1998 issue of Time Magazine, a poll was reported with the footnote:

“From a telephone poll of 1 020 adult Americans ... Margin of error is  $\pm 3.1\%$ .”

Can you reconcile this with the above formula? Two hints: the polls which they use. When the sample size is around 1,000, as it is often is, they claim that for the estimation of percentages the maximum margin of error is “about 3 per cent”.

Hints:

1. this implies that the error in proportion terms is 0.031;
2. the percentages the poll is estimating are often near 50%.

It may be that we can be quite sure that the percentage we are estimating is not going to be near 50%. We may be sure that it's either quite high or very low. In these circumstances the formula (3) can be used, and the value of  $\theta$  that should be inserted is the numerical value closest to 50% (0.5, as a proportion). If you know that the relevant percentage has to be at least 70%,

use  $\theta = 0.7$ ; if you know it must be less than 20%, use  $\theta = 0.2$ .

There is a sheet in the Excel workbook Confidence intervals for sample size.xls called: Population proportion - other. You can use this sheet in situations when you are quite sure that the percentage you are estimating is not going to be near 50%.

## 16.5 The hypothesis testing decision making framework

In Chapter 6, we introduced the framework of hypothesis testing for carrying a statistical test of a null hypothesis using data from an empirical research study. We now consider an extension of this framework that can be used when we wish to determine a sample size for a study that we plan to carry out.

We will use the example of monitoring water quality, introduced in Chapter 6.

First, here is a reminder of the concepts covered in Chapter 6:

- The null hypothesis is a hypothesis about a population parameter that specifies the absence of a true effect.
- The null hypothesis specifies an exact value for the population parameter.
- The null hypothesis and a set of relevant assumptions allow the distribution of a test statistic to be defined, when the null hypothesis is true.
- Statistical significance specifies a criterion for small  $P$ -values; small  $P$ -values are typically taken to be  $P < 0.05$ .

In our water quality example:

- We defined  $X$  to be the pH of a water sample from a tap, and assumed that  $X$  was normally distributed.
- We assumed independence of  $n$  samples;  $n = 25$  in this case.
- The null hypothesis was the preferred mean value for the pH:  $\mu = 7.5$ .
- We further assumed that the population standard deviation of pH values was known to be 0.5, i.e.  $\sigma = 0.5$ .
- The test statistic was the sample mean.

With these assumptions we saw that the distribution of the sample mean,  $\bar{X}$ , is itself normally distributed, in fact

$$\bar{X} \stackrel{d}{=} N\left(7.5, \frac{0.5^2}{25}\right).$$

This means that the standard deviation of  $\bar{X} = 0.1$ , so that the distribution of  $\bar{X}$  is as shown in Figure 98.

Sampling distribution of the mean for samples of size 25

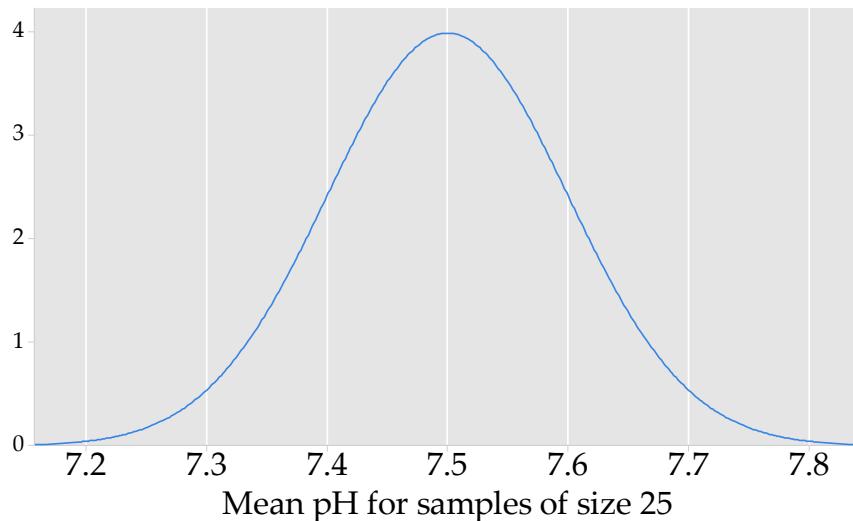


Figure 98: *Sampling distribution for the mean of a sample of 25 when the true mean is equal to 7.5 and the population standard deviation for an individual observation is 0.5.*

### 16.5.1 A competing alternative hypothesis ( $H_1$ )

In Chapter 6, we introduced an **alternative hypothesis** which was not precisely specified, in the sense that it does not specify a particular value for the parameter of interest. For the water quality example, the alternative hypothesis considered was  $\mu \neq 7.5$ .

In adopting a hypothesis testing framework for planning a study, a precise alternative hypothesis is considered. Quite often, several precise alternatives might be considered, in turn. For the moment, we consider one precise alternative in order to explain the important concepts.

In the water quality example, the water authority may be worried about the water being too alkaline. They want to know: how effective is our monitoring at detecting this? If the true mean pH is equal to 7.8, will the program pick this up?

Figure 99 illustrates the situation. If the water is actually alkaline, the sampling distribution of the mean will be shifted to the right, and it will be

centred around 7.8, if  $\mu = 7.8$ . This is an example of a specific alternative hypothesis. The null hypothesis (“situation OK”) is  $\mu = 7.5$ ; the alternative hypothesis (“alarm bells”) is  $\mu = 7.8$ .

Sampling distributions of the mean for samples of size 25

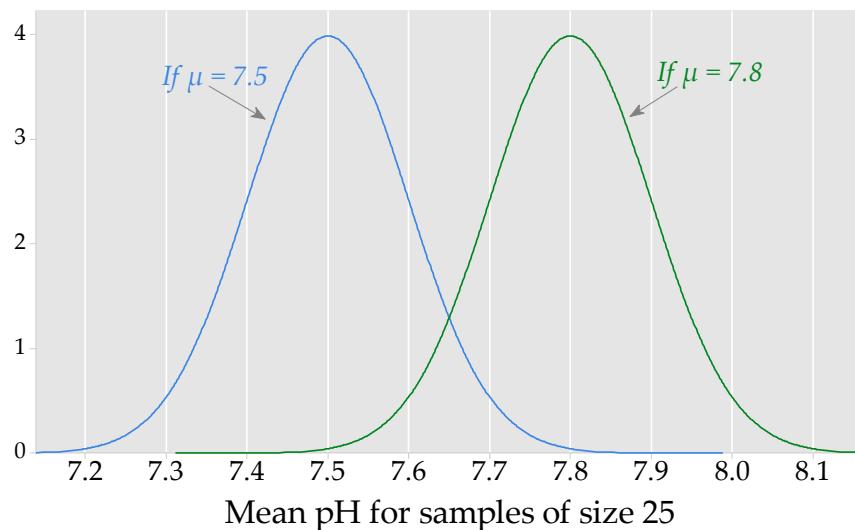


Figure 99: *Sampling distributions for the mean of a sample of 25 when the population standard deviation for an individual observation is 0.5, and the true mean is either  $\mu = 7.5$  (blue curve, null hypothesis) or  $\mu = 7.8$  (green curve, alternative hypothesis).*

### 16.5.2 Level of significance ( $\alpha$ ) and type I error

In Section 6.6, we introduced the term “statistically significant” to refer to a criteria for small  $P$ -values; small  $P$ -values are typically taken to be  $P < 0.05$ . In the context of study planning, we may wish to consider the implications if the decision is to doubt the null hypothesis and prefer the alternative when the  $P$ -value is small. Figure 100 shows, in blue, the area in the distribution under the null hypothesis that corresponds to small  $P$ -values.

In that case, the probability we define to be “small enough” is referred to as the “level of significance”; the symbol  $\alpha$  (‘alpha’) is often used. If the null hypothesis is true, the level of significance specifies improbable but not impossible values of the test statistic. Hence  $\alpha$  is also the probability of ‘rejecting’ (doubting)  $H_0$  when  $H_0$  is true. The traditional name for  $\alpha$  was the “size” of the test.

Corresponding to the level of significance are values of the test statistic that are regarded as implausible under the null hypothesis. These values are referred to as the “critical region” for the test statistic. A possible way to construct the test is to say that if the value of the test statistic belongs to the “critical region” we doubt  $H_0$ .

For the water quality example, this refers to values of mean pH corresponding to small  $P$ -values, in the tail of the blue distribution in Figure 100. Whenever the mean pH is in the tails, the  $P$ -value must be  $< 0.05$ .

If we doubt  $H_0$  when the test statistic falls in the critical region, sometimes we will do this even though  $H_0$  is true. This is because when  $H_0$  is true, it is possible for the test statistic to fall into the critical region, just improbable. If we ‘reject’  $H_0$  when it is true, this is known as a **type I error**. This means that  $\alpha = \Pr(\text{Type I error})$ . In planning a study, an a priori decision is made about the value of  $\alpha$  — about the long run risk of a type I error.

Sampling distributions of the mean for samples of size 25

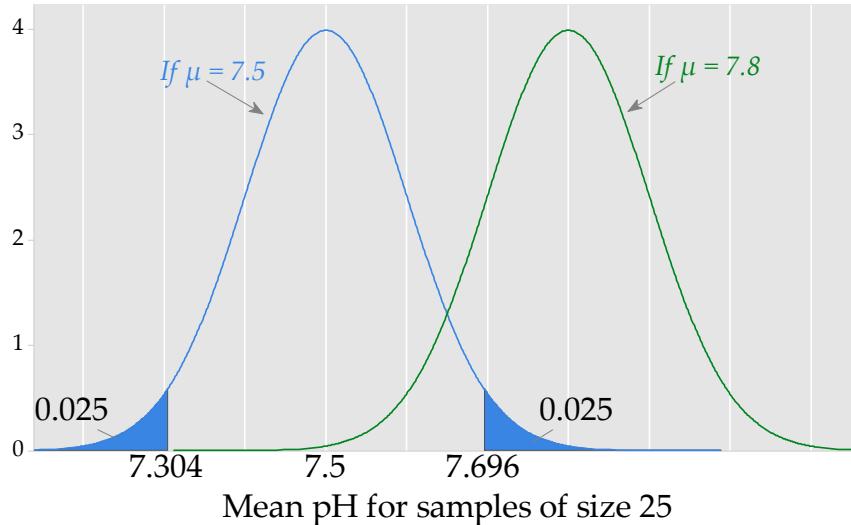


Figure 100: *Sampling distributions for the mean of a sample of 25 when the population standard deviation for an individual observation is 0.5, and the true mean is either  $\mu = 7.5$  (blue curve, null hypothesis) or  $\mu = 7.8$  (green curve, alternative hypothesis). The area corresponding to small  $P$ -values is shown in blue.*

### 16.5.3 Power

Ideally, if the null hypothesis is not true, our study will be appropriately set up to detect this. We hope that there is a good chance of finding evidence against  $H_0$  when, in fact, it is not true. The power of the test is defined to be the probability of rejecting  $H_0$  when  $H_0$  is false.

The power of the test is something we may assess in the design and planning of a study — before we have the data. So again we consider the properties of the theoretical distributions set up under the null and alternative hypotheses.

As we have just seen, we can set  $\alpha$ , the probability of a type I error, to our desired value:  $\alpha = 0.05$ , say. This determines the critical region for the test statistic, corresponding to outcomes when the decision would be to doubt

the null hypothesis. This also means we can work out the power for the given values of the parameter of interest.

In this framework, when  $H_0$  is taken to be false, the alternative hypothesis is taken to be true. The critical region determines the values that give a statistically significant result for  $\alpha = 0.05$ ; these values correspond to the blue shaded regions in Figure 100. The *power*, in this case, is the chance of a result in the critical region *when the alternative hypothesis is true*, that is, when  $\mu = 7.8$ . This is illustrated in Figure 101; the power is found to be 0.85, or 85%. So there is a good chance that a sample of 25 will give a statistically significant result (for  $\alpha = 0.05$ ) when  $\mu = 7.8$ , since 0.85 is quite a high probability.

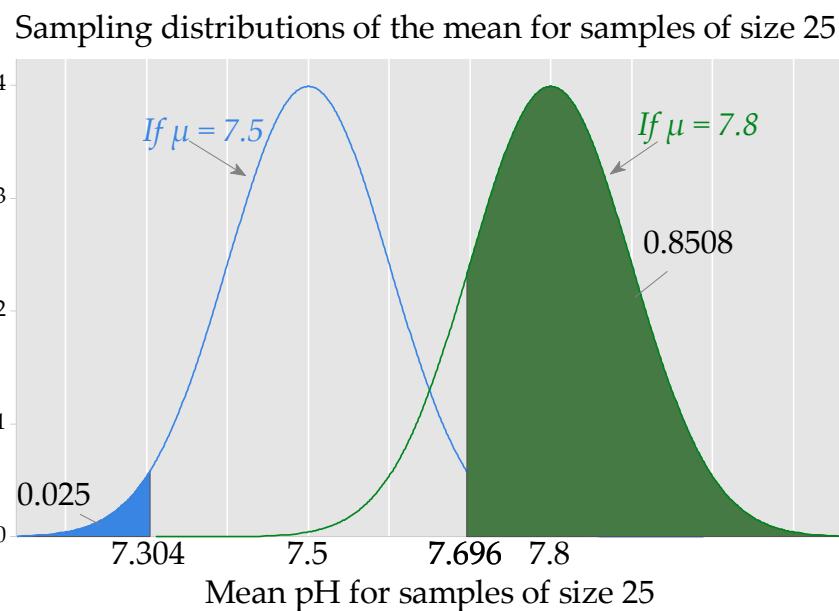


Figure 101: Statistical power for the water example; the power for the alternative hypothesis  $\mu = 7.8$  is equal to 0.85.

In practice in planning a study, a researcher might consider a number of different alternative hypotheses in order to understand what the power would be under different scenarios. Power can also be plotted continuously against continuous values corresponding to continuous values of an alternative hypothesis.

In the water quality example, the power for  $\mu = 7.6$  is 17%; this would mean that such studies have a chance of 17% of finding a statistically significant result at  $\alpha = 0.05$ , if the true difference in mean pH is 0.1. This is not a very high power, but 0.1 is a relatively small mean difference from the null hypothesis, so unless we have a very large study, it will be hard to detect such an effect.

As the true mean pH increases, the power correspondingly increases. The

power for  $\mu = 8.0$  is over 99%; this would mean that, if the true pH is 8.0, the study would be virtually certain to detect this. In other words, large true effects will make themselves known, while small true effects are hard to detect.

This illustrates how power can depend on a number of things. Of these, the one that can be most readily planned for in the design of a study is the sample size. All other things being equal, the power of the study increases as the sample size increases.

This relationship can be exploited to set sample sizes in studies, as we shall see. In the hypothesis testing approach to setting a sample size, we specify the power for a given value under the alternative hypothesis, and find what sample size is required to achieve this specification.

The notion of “post-hoc power”, calculated after the results are available, is found in the literature: it is a rather specious construct, since the best way to indicate the statistical effectiveness of a study after we have the data, is to quote an estimate and 95% confidence interval.

#### 16.5.4 Type II error

The power can be expressed in terms of the **type II error**.

Recall that the type I error is rejecting the null hypothesis when it is true; the type II error is not rejecting the null hypothesis when it is false.

Hence, the power can be defined thus:

$$\text{power} = 1 - \Pr(\text{type II error})$$

Of course, no one wants to make errors. But we can only avoid type I errors altogether if we never reject  $H_0$  under any circumstances (in which case we might as well not collect the data!). Similarly, we can only avoid type II errors altogether if we always reject  $H_0$ . In practice, we have to compromise, and we therefore have to make these errors sometimes.

Importantly it is not possible to know if we have made such errors in practice. These errors are defined within a planning framework and we can assess their probability in the long run. If however, when conducting a hypothesis test we find a small  $P$ -value and decide to doubt the null hypothesis, we can be sure we have not made a type II error. Why? Type II errors are not relevant when doubting the null hypothesis. A type I error is possible, but we cannot know if it occurred in our results. Why not? Because we don't know if the null hypothesis is true or not. And of course, it is also true that if we conduct a hypothesis test and choose not to doubt the null hypothesis, it is possible that we have made a type II error, but we still cannot know if this is what happened in this instance.

We emphasize that the decision making framework for hypothesis testing

where a ‘choice’ is made between hypotheses is most relevant to study planning. In applied practice, the results of a hypothesis test are best interpreted alongside a confidence interval which provides information about a set of plausible values for the parameter of interest — plausible hypotheses — consistent with the data.

### 16.5.5 A useful analogy

There is a helpful analogy between legal processes, at least in Westminster-style legal systems, and hypothesis testing. The parallels are shown in the following table.

Hypothesis testing	The law
null hypothesis $H_0$	accused is innocent
alternative hypothesis $H_1$	accused is guilty
don’t reject $H_0$ without strong evidence	innocent until proven guilty beyond a reasonable doubt
type I error	convict an innocent person
type II error	acquit a guilty person
$\alpha = \Pr(\text{type I error})$	beyond reasonable doubt
$\text{power} = 1 - \Pr(\text{type II error})$	effectiveness of system in convicting a guilty person

## 16.6 Use of power to determine sample size

We first looked at using confidence intervals to determine sample size. Here we consider the more traditional approach, based on hypothesis testing.

In the confidence interval approach, we saw that determining a sample size required the specification of a desired width in the confidence interval. In the hypothesis testing approach in practice, there are analogous settings that we need to specify.

The great majority of sample sizes worked out using hypothesis testing involve one of these two cases:

- a difference between two means
- a difference between two proportions

The approach can be used for other settings, such as the slope of a regression line, but this is less common.

Once we have collected the data, it is possible that we will do something more sophisticated than a simple two group comparison. Therefore, we

may feel that we should frame the sample size calculation in terms of the proposed analysis. This is a good principle, but difficult to implement, because we are setting the sample size at the design stage, and therefore don't have all the detail of the data, by the nature of the case. Hence we tolerate the simplifications involved in using a simple two group comparison.

For the same reason, sample sizes using hypothesis testing need to apply the traditional construct, of a threshold of significance, rather than the  $P$ -value. (Why can't we use the  $P$ -value?)

In this section, we deal with the two cases mentioned above, and then conclude with some general remarks about sample size.

### 16.6.1 Sample size for a comparison of means

We suppose that our study is going to compare two means,  $\mu_1$  and  $\mu_2$ , and that we will carry out a test of the null hypothesis  $H_0: \mu_1 = \mu_2$ , versus a two-sided alternative. We will carry out a test with level of significance  $\alpha = 0.05$ , say.

To set a sample size, we need an inferential goal. In hypothesis testing, this is done by asking: "How large a sample size do I need to detect a particular true difference,  $\mu_1 - \mu_2$ , with high probability?" In this framework, we "detect" a difference when we have a statistically significant result. Hence the question we have asked is essentially a question about statistical power.

Recall that the power of a test is the probability of rejecting the null hypothesis when it is false; that is, when the alternative hypothesis is true. When we discussed statistical power earlier, we noted that it is a function of the actual value of the parameter under the alternative hypothesis. This becomes important here: the smaller the true difference we hope to detect, the larger the required sample size. Conversely, if we specify a very large true difference to detect, a small sample size will do: if there really is a large effect, it will probably stand out in a small study.

We have to choose the power: it is usually set at 80% at least, and sometimes calculations are made for a variety of values.

To make the application of the ideas more concrete, consider the design of a study of the use of screen time in children, including TV, smart phones, tablets etc. A number of studies have indicated a weak but consistent association between hours using a screen and childhood obesity, not to mention other health concerns. A researcher in health promotion proposes a randomized trial in children aged 5 to 13. Half the children will get no intervention, and half will get an intervention designed to reduce their time using a screen, by promoting appealing alternatives after school, in particular.

Some studies have found that the average weekly screen time in this age group is approximately 40 hours. Suppose that the distribution among these children has a standard deviation of about 10 hours, again, based on past studies.

The researchers consider that even a reduction of only 5 hours per week on average would be worthwhile, and likely to impact on obesity, if the link is really causal.

Now these study plans and the information from earlier studies can be used to answer the sample size question. We put the study in a statistical framework.

- We assume that the number of weekly hours watched by the children has a standard deviation of 10 hours, and that this standard deviation will apply in each of the two groups.
- We assume that the distribution of this variable is at least approximately Normally distributed, although the Central Limit Theorem means that this assumption is not very important, unless the sample size is small.
- We will test the null hypothesis  $H_0 : \mu_1 = \mu_2$ , where  $\mu_1$  is the true mean in the control group and  $\mu_2$  is the true mean in the intervention group.
- The test will be two-sided, with a level of significance of 5% ( $\alpha = 0.05$ ) and a power of 80%.
- We plan to have the same number of children in both of the groups; while this is not necessary, it is statistically efficient.
- We want a power of 80% for the specific alternative  $H_1: \mu_1 - \mu_2 = 5$ , which corresponds to a mean in the intervention group that is 5 hours less than in the control group.

The way the test will actually be carried out is by calculating

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

using standard notation; this test statistic will be compared to the  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom (see Section 10.2).

The chance that this test statistic is beyond the threshold of statistical significance, when  $\mu_1 - \mu_2 = 5$ , depends on the sample size.

The answer is available in MINITAB using Stat > Power and Sample Size ► 2-sample t ...

For a “Difference” of 2 units, a “Power value” of 0.8, with a “Standard deviation” of 4 units, we obtain the answer that 64 children are required in each group, and hence a total sample size of 128 is needed.

The sample size was the output of this calculation. As is apparent from the MINITAB dialogue box, which asks us to “Specify values for any two of the following”, we can:

- Specify the difference  $\mu_1 - \mu_2$  of interest and the required power, and obtain the sample size needed to achieve this;
- Specify the difference  $\mu_1 - \mu_2$  of interest and the sample size in each group, and obtain the power that this will achieve;
- Specify the sample size and power, and obtain the smallest difference  $\mu_1 - \mu_2$  that will satisfy these settings.

For example, suppose that the researcher only has the resources for a study of 30 children per group. For the same difference of interest,  $\mu_1 - \mu_2 = 5$ , what power is achievable? The answer given is 0.478, or 48%; there is only a chance of about 1 in 2 that the study will detect the desired effect with a sample size of 30.

What if the researchers believe that they can reduce children’s screen use by 15 hours, on average? What sample size is required for a power of 80%, for this case? We find that  $n$ , the number required in each group, is 9, and hence a total sample size of 18 is required. Is this credible? This is indeed a rather small sample size: but the hypothesized effect, a shift of 15 hours per week, is huge: it amounts to reducing the number of hours by more than a third, on average. It equals 1.5 standard deviations of the distribution of weekly screen use. If such a change really occurred, it would indeed be very noticeable, in even a small study.

The formal calculation in MINITAB is rather complicated.<sup>33</sup> However, simple approximations can be given.

For a power of 80%, an assumed standard deviation of  $\sigma$ , a true difference of  $\mu_1 - \mu_2$  and a level of significance of 5%, the sample size,  $n$ , required in *each* of two equal sized groups is approximately

$$n = \frac{15.7\sigma^2}{(\mu_1 - \mu_2)^2}.$$

For the example, this gives  $n = 62.8$ , very close to the value we obtained using MINITAB, which was 64.

For a power of 90%, an assumed standard deviation of  $\sigma$ , a true difference of  $\mu_1 - \mu_2$  and a level of significance of 5%, the sample size,  $n$ , required in *each* of two equal sized groups is approximately

$$n = \frac{21\sigma^2}{(\mu_1 - \mu_2)^2}.$$

---

<sup>33</sup>It uses the “non-central  $t$  distribution”, which is beyond the scope of this course.

These approximations are adequate for practical purposes. Also, the form of the equation in the approximation tells us how the sample size depends on  $\sigma$  and  $\mu_1 - \mu_2$ . The larger the value of  $\sigma$ , the larger the required sample size. On the other hand, the larger the value of  $\mu_1 - \mu_2$ , the smaller the required sample size.

Setting a sample size, even in this simplified way, is not easy. The two most common obstacles are:

- Finding a reasonable estimate of  $\sigma$ , the standard deviation in each group;
- Setting the minimum difference of interest,  $\mu_1 - \mu_2$ .

The first of these needs some information about the distribution of the relevant measure. That might come from past studies, or a pilot study. We may be told “I have no idea what the variance (or standard deviation) of the measurements will be”. Even if we have very limited information, we may know something about the range of the variable; if so, we may consider that the range is about  $6\sigma$ , since almost all of most distributions will be within 3 standard deviations of the mean.

Since the value of  $\sigma$  is often not known precisely, it is usually appropriate to try values in a plausible range.

Setting the minimum difference of interest is a matter of deciding the smallest difference that could matter, for a practical purpose. In medical contexts it is sometimes called the “minimum clinical difference”, that is, the smallest difference that would matter, from a clinical point of view. In this context, it is vital to keep in mind that we are talking about a difference between the true *means* of the distributions under the two treatments, and not a shift of this amount for a single individual. If there is a true difference in the means, this has a “population” impact, and this may be important, even for a difference that may look small on an individual basis.

In the example, we planned to have equal numbers in each group. This is a good idea, if it can be achieved. It is not required for a valid analysis; rather, it is a matter of statistical efficiency. If the total sample size available is  $N$ , then in standard situations we get the best statistical efficiency (narrowest confidence intervals, most powerful tests) if half are allocated to each group.

Sometimes, reasons of cost or availability prohibit this.

If the sample sizes are required to be in a different ratio from 1:1, then a sample size calculation can still be done, but it is not available directly in MINITAB.

### 16.6.2 Sample size for a comparison of proportions

The second common application is the comparison of proportions.

The structure is essentially the same as for the comparison of means; we have to think about the difference that is important to detect, and we have to make choices about statistical power and so on.

Suppose that we are carrying out a study to compare two proportions,  $\theta_1$  and  $\theta_2$ , and that we will carry out a test of the null hypothesis  $H_0: \theta_1 = \theta_2$ , versus a two-sided alternative. We will carry out a test with level of significance  $\alpha = 0.05$ , say.

Again, we need an inferential goal. Here, we ask: "How large a sample size is needed to detect a particular true difference,  $\theta_1 - \theta_2$ , with high probability?" If we set the power to be 0.90, and plan for equal sample sizes in both groups, then the sample size required in *each* group is approximately given by

$$n = \left( \frac{1.2816 \sqrt{\theta_1(1-\theta_1) + \theta_2(1-\theta_2)} + 1.96 \sqrt{2\bar{\theta}(1-\bar{\theta})}}{\theta_1 - \theta_2} \right)^2, \text{ where } \bar{\theta} = \frac{1}{2}(\theta_1 + \theta_2).$$

The numbers 1.2816 and 1.96 in this formula come from the standard Normal distribution; 1.2816 comes the choice of 0.90 as the power, and 1.96 comes from choosing  $\alpha = 0.05$ .

MINITAB does the work for us in Stat > Power and Sample Size ► 2 Proportions

...

For example, suppose we are interested in testing whether the usage of a particular managerial technique is similar in Australia and New Zealand, among companies listed on the stock exchange. We have reason to believe that the percentage of companies using the technique is about 30%. We are interested in detecting a difference between the two countries of 20%, that is,  $\theta_1 - \theta_2 = 0.2$ .

Then, regarding the Australian companies as group 1 and the New Zealand companies as group 2, we might take  $\theta_1 = 0.4$  and  $\theta_2 = 0.2$  as our specific alternative hypothesis, and we find that applying the formula gives a sample size required (in each sample) of about 108.2, which we round *up* to 109, and hence a total sample size of 218.

In fact, this approximation is a bit rough, and tends to give values that are too low. There are a number of improvements: one of the simple ones is:<sup>34</sup>

$$n = \left( \frac{1.2816 \sqrt{\theta_1(1-\theta_1) + \theta_2(1-\theta_2)} + 1.96 \sqrt{2\bar{\theta}(1-\bar{\theta})}}{\theta_1 - \theta_2} \right)^2 + \frac{2}{|\theta_1 - \theta_2|}.$$

Comparison with the previous formula shows that this improved approxi-

---

<sup>34</sup>This formula may be found in Fleiss JL, Tytun A and Ury HK (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 36:343–346.

mation is just the previous value, plus  $\frac{2}{|\theta_1 - \theta_2|}$ .

For the example, this gives  $n = 119$ , and a total sample size required of 238.

A “quick and dirty” two-proportions formula for equal-sized groups is:

$$n = \frac{5.25}{(\theta_1 - \theta_2)^2}.$$

This uses  $\alpha = 0.05$ , power = 0.90, and the fact that for  $\theta$  in the interval (0.3, 0.7),  $\sqrt{\theta(1-\theta)}$  doesn’t vary too much from its maximum of 0.5. But it doesn’t work well if the proportions we are comparing are outside the interval (0.3, 0.7).

Suppose we’re comparing two proportions that are close to 0.5, and we want to be able to detect a true difference of  $\theta_1 - \theta_2 = 0.1$ , i.e. 10% on the percentage scale.

- The quick and dirty formula says 525;
- If it is reasonable to say that, under the alternative hypothesis,  $\theta_1 = 0.55$  and  $\theta_2 = 0.45$ , the approximation used in MINITAB gives an answer of 524 per group;
- The improved approximation gives 544 per group.

Note that the sample size depends somewhat on the actual values  $\theta_1$  and  $\theta_2$  under the specific alternative hypothesis, although not a lot.

In practice, the formula used in MINITAB is good enough to give a reasonable idea of the required sample size. However, for something where the requirements are a little more rigorous, such as a grant application, the improved approximation should be used.

### 16.6.3 General issues in sample size determination

The ideas applied in the previous two sub-sections can be extended to other cases.

Conceptually, you need:

- A parameter  $\lambda$  that you want to make an inference about;
- A null hypothesis about  $\lambda$  you wish to test;
- A statistic to test it, based on an estimator of  $\lambda$ ;
- A prescribed level of significance,  $\alpha$ ;
- The value of the parameter corresponding to the smallest departure from  $H_0$  that is important to detect (this means practical significance or importance, not statistical significance);

- A prescribed power, that is, the probability of declaring the result statistically significant when the true value of the parameter is at least as far away from the null hypothesis value as you specified;
- (Something like) the inherent variation (continuous measure) or the background proportion.

This sounds like a lot ... and it is. When determining sample sizes statistically, it is necessary to have quite a large amount of information, at least approximately.

In practice:

- The choice of  $\lambda$  is not as obvious as it seems. This is where a “concentration of mind” on the part of the researcher may need to occur. For example, a question like:

“In my study of housing styles, how many houses should be in my sample to get an idea of the typical house?”

needs some work before it is translated into a question amenable to a formal sample size calculation.

- The null hypothesis will generally “choose itself”, as will the statistic or estimator.
- As in analysis, it is usual to use a two-sided test.
- $\alpha = 0.05$  is almost universal (Fisher), like 95% confidence intervals;
- The smallest departure from  $H_0$  that is important to detect needs to be argued for in terms of the relevant discipline.

It could be important to detect smaller effects for outcomes with more serious consequences.

Information from a pilot or another study may help, but this will only give you an estimate of the *likely* value of the parameter, not the minimum you wish to detect.

- Power should be at least 0.80, and often calculations are shown for 0.80, 0.90 and 0.95. Remember that setting the power at a high value does not guarantee a statistically significant result; it just makes it likely, if the parameter value is as specified.
- For many cases we have, at least approximately, a test statistic that is Normally distributed:

$$U \stackrel{d}{\approx} N(\lambda, \sigma_U^2).$$

Note that  $\sigma_U^2$  may depend on  $\lambda$ .

This means that approximate sample sizes can be based on Normal theory.

- There may be more than one parameter of interest; then pick a few key ones.
- Keep in mind the sample size requirements of sub-analyses. If they're important, look at the sample size issue for these too.
- If estimation is the “natural” inferential goal—for example, you have a survey—then work out  $n$  based on confidence interval theory.
- Often the maximum sample size that can be used in a study depends on financial and/or time restrictions and under these circumstances the width of confidence intervals and/or the power of tests can be used to help decide whether or not it is worthwhile carrying out the study.

Other sample size and power calculations are available in the MINITAB module, notably:

- for one-way ANOVA (discussed in Chapter 10);
- for 2-level factorial designs (Section 12.2).

Sample size calculations should be regarded in general terms rather than as specific “answers”.

## 16.7 Exercises

- 16.1 A research study is proposed for examining the attitudes of cat owners to the proposal that cats should be confined to the property of the owner. Assume that a random sample of cat owners can be found. (..)
- What sample size is needed so that the 95% confidence interval for the true percentage in favour of the proposal should be no wider than 20%? That is, we want the confidence interval to be of the form  $\hat{\theta} \pm 10\%$ .
  - Use Minitab to check your answer in (a) by finding 95% confidence intervals for the value of  $n$  you obtained, for various sample percentages (i.e. various numerators), and examining their widths.
  - Now suppose that the researcher requires a more demanding precision, that the 95% confidence interval for the true percentage in favour of the proposal should be no wider than 8%. What sample size is required for that?
  - The researcher discovers that the sample size you have worked out in (c) allows for an estimated percentage anywhere in the possible range of 0% to 100%, and that percentages near 50% are “worst case” scenarios. She says that based on anecdotal observations, the percentage in favour of the proposal must be 25% or less. Recalculate the required sample size in the light of this extra information, still requiring that the 95% confidence interval for the true percentage in favour of the proposal should be no wider than 8%.
  - What percentage reduction in sample size was achieved, using this additional information?
- 16.2 The Harm Avoidance scale is part of a general instrument for measuring temperament and personality characteristics. In subjects with no mental illness, it is found to have a mean of 13 and a standard deviation of about 7. A study is proposed to test whether patients with bipolar disorder have a higher mean for Harm Avoidance, by comparing a number of the controls with an equal number of patients with bipolar disorder.
- How would you determine the minimum difference in means which is clinically important?

- (b) Determine the sample size required in each group if it is desired to carry out a test with a power of 0.90 to detect a true difference in means of 6 units, using a 2-sided test and a level of significance of 5%. What assumption(s) have you made?

[Stat > Power and Sample Size > 2-Sample t; enter appropriate values for Differences, Power values and Standard deviation (the population standard deviation). The alternative hypothesis and level of significance can be changed in Options.]

- (c) Repeat the calculation for a true difference in means of 1 unit; 3 units; 10 units.

## 16.8 Answers

- 16.1 (a) If the width of the 95% CI for the percentage is 20%, that means a CI of the form estimate  $\pm 10\%$ , hence, on the proportion scale, estimate  $\pm 0.1$ . The rough formula gives  $n \geq 1/(0.1^2) = 100$ ; using the more precise formula gives  $n \geq 96$ . In Minitab, if we fix  $n = 96$  and find 95% CIs for a varying number of successes  $x$ , we get (e.g.):

$x = 30$ : point estimate,  $30/96 = 0.313$ , 95% CI: (0.222 to 0.415), width 0.19;

$x = 48$ : point estimate,  $0.500$ , 95% CI: (0.396 to 0.604), width 0.21;

$x = 80$ : point estimate,  $0.833$ , 95% CI: (0.744 to 0.902), width 0.16.

- (b) Rough formula:  $n \geq 1/(0.04^2) = 625$ ;  
more precise formula:  $n \geq (3.84 \times \frac{1}{4})/(0.04^2) = 600$ . (*It's actually 3.8416, which gives 600.25.*)
- (c) Now the worst case is the value of  $\theta$  closest to 0.5, which is 0.25, and we get  $n \geq \frac{3.84 \times 0.25 \times 0.75}{0.04^2} = 450$ .
- (d) The sample size required has decreased by 150, that is by 25%.

- 16.2 (a) This is usually very difficult! It is not really a statistical question. In this case, for example, you might try to choose a value which is going to have a distribution with 10% of subjects having Harm Avoidance scores which lead to institutionalisation, or some other undesirable outcome.
- (b) Using the formula in the notes, the number in each equal-sized group needs to be 29. The assumptions are that you can obtain random samples from the populations about which you wish to make inferences; that the sample size is large enough for the CLT approximation to be adequate, and crucially, that the standard deviation in the manic depressive group will be about the same as in the healthy controls.
- (c) Using the formula in the notes, for 1 unit:  $n = 1029$ ; for 3 units:  $n = 115$ ; and for 10 units:  $n = 11$ . If there really is a big difference, a small study will show it.



# 17 Extensions and modern methods

Much of the topics on analysis in this subject are about the general linear model: modelling a continuous outcome with a linear function of some parameters and explanatory variables, plus a random error. This family of models is very broad, includes simple methods that are pervasive in research, has stood the test of time and is used commonly in data analysis, throughout academic research, business analytics and every other possible field of application.

## 17.1 Extensions of linear models

We have focussed on the linear models with either one or more categorical explanatory variables, or one or more continuous explanatory variables.

Here we briefly list some of the further extensions that we have not considered here.

- It is possible to use a linear model when there is both categorical and continuous explanatory variables and such models are commonly fitted. They involve no new concepts analytically, although the way they are applied needs thought, depending on the design of the study from which the data comes. Interactions of categorical and continuous explanatory variables can be included in such models. The principles of model fitting and the interpretation of the estimated parameters follow the methods illustrated in this subject.
- The linear model can be extended to situations in which there are “random factors” as well as “fixed factors”. We have restricted attention to models with fixed factors here; models with random factors entail more than one level of random variation, a feature that needs special attention. Models with both fixed and random factors are referred to as “linear mixed models”; the word ‘mixed’ refers to the mixture of types of explanatory variables, fixed and random.
- Another very broad and varied class of model that extends the linear model is the class of “generalised linear models”, in which a linear term features in the explanatory component of the model, but through a non-linear function of some sort. We have seen one (very important) example of this, namely, logistic regression. In logistic regression, the outcome, a probability, is not modelled as a directly linear function, but there is a linear function embedded in the model. There are more special cases in common use, such as Poisson regression and negative binomial regression, both used for counts.

- An important model that is not a generalised linear model, but has a linear component, is Cox's proportional hazards model, used for survival data when there is censoring present.

Our focus has been limited by the scope of an introductory subject.

## 17.2 Some modern methods

The linear model and logistic regression are both analytic techniques that are mentioned in the topic of "machine learning". They were both in common use long before the term 'machine learning' was coined.

There are other approaches to the problem of prediction, in particular, that are alternatives in the contexts that linear models are used. We mention some examples here to illustrate the growing diversity of modern analytics and to discuss important statistical questions that arise about these methods.

### 17.2.1 Regression trees

An important alternative that emerged when computing power overcame computational obstacles is an approach known as regression trees. You may see this referred to in the more comprehensive term classification and regression trees, or CART. The central idea here, which captures the essence of the approach, is a tree. We will see why this term is appropriate.

Classification trees deal with an outcome that is categorical, which may be binary but does not need to be. This often occurs and is an important area of application; many of the principles of regression trees also apply to classification trees.

Regression trees are for continuous outcomes. Hence, they are used for the type of situation when a general linear model may be appropriate. The term regression is really only used to suggest this connection; there is usually not a regression model of any kind involved.

The purpose of a regression tree is to make predictions of a continuous response variable, in an optimal way. The method relies on successively splitting the data; hence groups are formed by branching out further and further, until an end is reached. So, there is an initial fork (a split into two) and the two branches down that fork may split again (another fork) and so on, until the end of a branch is reached. Terms such as fork, branch and split suggest why the word tree is appropriate. When you see the method applied, and graphical representations of it, this tree metaphor is very evident.

The standard regression tree algorithm uses a time-honoured and traditional criterion of the total of the within group sum of squares. This is used

repeatedly throughout the tree.

The goal with fitting a regression tree, like any analytics for prediction, is to find a suitably parsimonious model that makes predictions efficiently and accurately. In making predictions with such a model we certainly may gain insights into the relationships between the explanatory variables and the outcome, and researchers often find them to be a useful exploratory tool.

However, precisely because of their intrinsic complexity, and natural tendency to allow for interactions between variables, it can be harder, with a regression tree, to provide the kind of summarising inference that science prefers. This is seen as a limitation of regression trees and may be why their use is not prolific in disciplines such as medicine and epidemiology, where measures of causal effect are sought.

On the other hand, clinical decisions are often made using a tree-like approach, so classification trees are more prevalent in medical fields. A related issue is that regression trees do not entail, usually, many of the elements of the standard framework of statistical inference, such as hypothesis testing and confidence intervals.

### 17.2.2 Random forests

A single regression tree is a function of the training data used to develop it, and, in general, may overfit to these data; this means that it may have aspects that are more sensitive to the dataset than is desirable. We now briefly discuss random forests, which are an approach to dealing with this issue. One justification of random forests is empirical; they typically give better predictive accuracy than just one tree.

A random forest is an ensemble of regression trees. Sampling with replacement is the basis of this method. Samples of the data are taken and for each, a regression tree is obtained. Sampling occurs in two ways, typically: samples of observations are used, and a sample of the predictor variables is also used at each node, so that a particular tree uses both a sampled data set and samples of predictors. This means that each tree is very likely to be unique.

The number of such trees may be hundreds or thousands, and they make up the ‘forest’, which is ‘random’ because of the random sampling involved.

Each tree will have its own set of predictive rules. For a particular set of values for the explanatory variables (e.g. person is male, aged 43, works part-time, lives in rental accommodation, is divorced, has two children both as school ...) each of these trees will make a prediction (e.g. annual expenditure on technology). These predictions will vary across the trees.

The prediction made by the forest is then the average of the predictions of all of the trees. The strength of this approach is the averaging; the final

prediction is not dependent on the idiosyncrasies of one set of data.

Random forests have been quite successful in tests of prediction. Key outputs from a random forest include the predictions themselves, and relative importance measure for the explanatory variables; this can be measured from the extent of an explanatory variables involvement in the trees making up the forest.

For example, age may be found to be used in all of the trees, whereas number of children at school might be in only 3% of the trees. A drawback of random forests is that there is no easy way to represent them in a diagram. However, this is also true of many other models of some complexity, such as linear models with several predictor variables.

### 17.2.3 Training and testing

Linear models and regression trees can be considered to be pattern recognition techniques: aiming to optimise the modelling or prediction of an outcome on the basis of explanatory variable. They do this in different ways, and each has its strengths and weaknesses.

An important issue – whichever approach is used – is the realistic assessment of the performance of a model or tree. How do we know how well it does? This can be addressed, for example, using a measure used for the closeness of predictions to the actual values for regression trees, and the same question arises in linear models.

You may have been left with a vague unease about one aspect of this. We ask: how well did the method perform? If we answer that question with the same set of data used to derive the model or tree, it seems that there is some circularity.

This concern is right. When any methods predictive performance is assessed on the same data used to derive the model or tree for prediction, the performance will typically be over-optimistic. Beyond the main structures and patterns that the approach may successfully capture, the particular data set used will have minor quirks that will tweak the model slightly, to do well for that dataset. In formal statistical terms, the predictions and observed values are not statistically independent.

A better, more realistic assessment of any predictive technique involves testing the model on a different set of data than that used for the development of the fitted model.

In practice, this is done for a number of different contexts and models. For example, in time series analysis, if a method's predictive performance is being assessed, a so-called hold-out sample of consecutive outcomes is defined at the end of the series, and not used in the fitting of the model. The

models parameters are then estimated on the rest of the data (only). Finally, predictions are made, using the fitted model, on the hold-out sample, and compared to the observed values.

This principle applies widely and is relevant to regression trees. One method of legitimately estimating a tree's performance in a realistic way is **cross-validation**. If the dataset consists of  $n$  (multivariate) observations, the cross-validation algorithm proceeds as follows.

- Omit the first observation from the dataset. Estimate the tree on the remaining data. Obtain a prediction for the first observation, from the fitted tree. Store the prediction with the actual first observation.
- Repeat for all observations separately.
- Finally, compare the predictions with the actual values.

This entails fitting  $n$  trees; if the data set has 5,000 observations, 5,000 trees are obtained, and if random forests are used, 5,000 random forests. That sounds like a lot of computation, and it is. But the fitting of trees and random forests is computationally intensive anyway, and this is not a greatly demanding additional step. In other contexts, such as prediction competitions, simpler approaches are sometimes used: a team may be provided with a training set of data to develop its model. The fitted model is then submitted and evaluated on an (unseen) test set.



## Glossary

Throughout the glossary the notation [sample] and [population] has been used to distinguish sample and population versions of a particular term: thus, for example, under Quantiles there are entries for [population] Quantiles and [sample] Quantiles — these entries do not appear under Population or Sample.

- Additive model for two way classification —  $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$ , i.e. observation = overall mean + row effect + column effect + random error. Thus it is assumed that (apart from random error) the effects of rows and columns simply add. This means that the row-effects are the same in each column (and that the column-effects are the same in each row).  
[Note:  $y_{ijk}$  denotes the  $k$ th observation made with row level  $i$  and column level  $j$ .]
- Adjusted  $R^2$  — makes an adjustment to the coefficient of determination,  $R^2$ , for the number of predictors in the model, which better indicates the “value” of the model.

$$R_{\text{adj}}^2 = 1 - \frac{\text{error SS}/(n - p - 1)}{\text{total SS}/(n - 1)}$$

where  $p$  is the number of explanatory variables.

- Alternative hypothesis, experimental hypothesis  $H_1$  — describes the effect under investigation. For example, the new additive is better than the old one:  $\mu_{\text{new}} > \mu_{\text{old}}$ .
- Analysis of variance (ANOVA) — a general procedure for testing hypotheses about the mean by studying the variation of the data. This may seem odd at first, but the variance of a set of data is increased by variation of the mean (with differing groups for example). If this increase in variance is sufficiently large then we can infer a variation in the underlying mean value.

This general procedure goes by the name of analysis of variance, and it does what its name suggests: analyses the variance. It achieves this by subdividing the variance into components attributable to variation around the mean value and variation of the mean. In some cases the variation of the mean can be further sub-divided enabling more precise attribution of the sources of variation.

- Balanced incomplete block design — like a randomised block design but with fewer units per block than the number of treatments.
- Binomial distribution —  $\text{Bi}(n, p)$ . This is the distribution of the number of successes obtained in a sequence of  $n$  independent trials each having probability of success  $p$ . This is a discrete distribution with probability mass function

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (x = 0, 1, 2, \dots, n)$$

If  $X \stackrel{d}{=} \text{Bi}(n, p)$ , then  
 $E(X) = np$  and  $\text{var}(X) = np(1 - p)$ .

- Bivariate data displays — see Line plot (mainly for time series) and Scatter plot.
- Block — a collection of a homogeneous group of plots (experimental units); these should be such that the variation of plots within a block is less than the variation between blocks. This makes the error variance smaller, leading to increased precision.
- Blocking — the allocation of experimental units to blocks, i.e., groups of units which are considered likely to be more homogeneous than the entire collection of available units. The purpose of blocking is to reduce the variance of the random error by accounting for some of the variation between units.
- Boxplot — A boxplot is a visual summary of a data set based on the median, quartiles, and extreme values. Boxplots are especially useful for comparing (two or more) data sets. The rectangle with sides at  $Q_1$  and  $Q_3$  is referred to as the 'box' and the lines, which extend from the sides of the box as the 'whiskers'. Thus the length of the box is the interquartile range (IQR). The whiskers are drawn from each end of the box to the furthest data value not greater than 1.5 IQR from the end of the box. This means that the whiskers extend to the maximum and the minimum points, if these points are within 1.5 IQR of the box. Observations which are between 1.5 IQR and 3 IQR from the box are designated by '\*', and are called *possible outliers*; observations which are more than 3 IQR from the box are designated by 'o' and are called *probable outliers*.
- Central limit theorem — essentially says that if you add up lots of little things then the result is approximately normally distributed. And it doesn't matter how the little things are distributed, just so long as they are relatively little (compared to the sum). Note that the average is just a multiple of the sum, so if the sum is normal then the average is too.

The simplest mathematical version of the central limit theorem is as follows: if  $X_1, X_2, \dots, X_n$  are independent identically distributed random variables with mean  $\mu$  and variance  $\sigma^2$ , then  $\bar{X} \xrightarrow{d} N(\mu, \sigma^2/n)$  as  $n \rightarrow \infty$ .

- (Chi-squared)  $\chi^2$  test of independence — a test for the independence of the categorical factors in a contingency table. Expected frequencies based on the assumption of independence are calculated and compared with the observed frequencies using the statistic  $\chi^2 = \sum(o - e)^2/e$ . If the value of  $\chi^2$  is too large (compared to the upper tail of a  $\chi^2$  distribution with  $(r - 1)(c - 1)$  degrees of freedom) then the independence hypothesis is rejected.
- (Chi-squared)  $\chi^2$ -distribution — the distribution of a sum of squares of standard normal random variables. Formally: if  $Z_1, Z_2, \dots, Z_p$  are independent  $N(0, 1)$  random variables, then  $U = \sum Z_i^2$  has a  $\chi^2$  (pronounced "kie-squared") distribution with  $p$  degrees of freedom; we write  $U \xrightarrow{d} \chi_p^2$ . The mean of a  $\chi_p^2$  distribution is  $p$ .

For a random sample from a normal population, the sample variance has a distribution related to  $\chi^2$ :  $(n - 1)S^2 \xrightarrow{d} \chi_{n-1}^2$ .

In more complicated situations when  $S^2$  is obtained as a residual mean square,  $pS^2 \xrightarrow{d} \chi_p^2$ , where  $p$  is the number of degrees of freedom of the residual mean square in the ANOVA table.

- Coefficient of determination — It is useful to consider how well the regression model performs, in terms of explaining the variation in  $Y$  as a function of the dependent variables (the  $xs$ ). A measure that is used for this purpose is

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \frac{\text{error SS}}{\text{total SS}}.$$

Roughly speaking,  $R^2$  is the proportion of the variation in  $y$  explained by the  $xs$ . See also Adjusted  $R^2$ .

Note: if there is only one  $x$ , then  $R^2 = r^2$ , where  $r$  denotes the correlation coefficient (between  $x$  and  $y$ ).

- [sample] Coefficient of variation;  $s/\bar{x}$ . The coefficient of variation is useful when we want to compare the spread of different sets of values with different scales.
- Confidence coefficient — the level of confidence associated with a confidence interval: thus if the confidence coefficient is  $100\gamma\%$ , then the  $100\gamma\%$  confidence interval for  $\theta$  is an interval within which contains the true value of  $\theta$  with probability  $\gamma$ .
- Confidence interval — a 95% confidence interval for  $\theta$  is an interval which contains the true value of  $\theta$  with probability 0.95. The level of confidence is typically chosen to be 95%. But it may take other values: in which case, the 95 and 0.95 would be replaced in the definition. See Confidence coefficient.

An approximate 95% confidence interval is often given by estimate  $\pm 2$  standard error.

- Confounding — effects are confounded when we cannot tell whether the observations are attributable to one effect or the other.
- Contingency tables (with two factors) — count data categorised by two categorical variables: this results in a two-way table with rows representing the levels of one categorical variable and columns the levels of the other. The entries in the table are counts or observed frequencies.
- Continuous random variables — can take any value on a continuous scale within the range of possible values. The distribution of a continuous random variable is defined by specifying a curve which relates the height of the curve at any particular value to the chance of that value occurring. This curve is called the *probability density function* (pdf), and it is denoted by  $f(x)$ :

$$f(x)h \approx \Pr(x - 0.5h < X < x + 0.5h)$$

(for small  $h$ ). Thus, if  $X$  is measured to the second decimal place, then

$$\Pr(X \text{ observed as } 2.34) \tag{4}$$

$$= \Pr(2.335 < X < 2.345) \approx 0.01 f(2.34). \tag{5}$$

The probability that a continuous random variable takes a value in an interval between two points  $a$  and  $b$  is the area under the curve between  $a$  and  $b$ .

$$\Pr(a < X < b) = \int_a^b f(x)dx$$

- [population] Correlation,  $\rho$  is a measure of the strength of the linear association between random variables. It is defined by

$$\rho = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and is such that  $-1 \leq \rho \leq 1$  with a value of  $\rho$  close to  $\pm 1$  indicating a strong (linear) relationship and values close to 0 indicating a weak (linear) relationship.  $\rho$  does not depend on the units used to measure  $X$  and  $Y$ . If  $\rho = \pm 1$  then  $Y = a + bX$ . If  $\rho = 0$  then  $X$  and  $Y$  are uncorrelated. If  $X$  and  $Y$  are independent then  $\rho = 0$ , but  $\rho = 0$  does not necessarily mean that  $X$  and  $Y$  are independent.

- [sample] Correlation, the correlation coefficient,  $r$ , the Pearson's product moment correlation;  $r$  is a measure of the strength of the *linear relationship* between two variables. It is defined by

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

and is such that  $-1 \leq r \leq 1$  with a value of  $r$  close to  $\pm 1$  indicating a strong (linear) relationship and values close to 0 indicating a weak (linear) relationship. The association is positive (i.e. one variable tends to increase as the other increases) if  $r > 0$ , and negative (i.e. one variable tends to decrease as the other increases) if  $r < 0$ . See also Rank coorelation.

- [population] Covariance — the covariance of random variables  $X$  and  $Y$ , is denoted by  $\sigma_{X,Y}$  or  $\text{Cov}(X, Y)$ . It is the weighted average of the products of the deviation of  $X$  from its mean and the deviation of  $Y$  from its mean, where the weights are provided by the probability distribution.
- [sample] Covariance,  $s_{xy}$ ; roughly, the average of the products of the deviations of the  $x$  and  $y$  data values from their sample means:

$$s_{xy} = \frac{1}{n-1} \{(x_1 - \bar{x})(y_1 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})\}$$

$$= \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

- Covariates — a covariate is a continuous variable associated with a plot which may affect its response but which is (generally) not part of the main objective of the study.
- Critical region (rejection region) — the values of the test statistic for which the null hypothesis is rejected.
- Cumulative distribution function (cdf) — usually denoted by  $F(x)$ , this gives the probability that the random variable takes a value less than or equal to  $x$ :

$$F(x) = \Pr(X \leq x)$$

For example,  $F(2) = \Pr(X \leq 2)$ .

- Degrees of freedom — the degrees of freedom of a set of elements, is the number of the elements which are free to vary. For a random sample of size  $n$ , the degrees of freedom is simple  $n$ . But for the set of deviations from the sample mean:  $(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$ , the degrees of freedom is not  $n$  but  $n - 1$ . This is because the sum of these deviations is zero; so that if we know  $n - 1$  of the deviations the the remaining one is not “free”, but is determined.
- Differences —  $E(X - Y) = E(X) - E(Y)$  always; and  $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$  if  $X$  and  $Y$  are independent. These results can be derived using  $X - Y = X + (-Y)$  and  $E(-Y) = -E(Y)$ ,  $\text{var}(-Y) = \text{var}(Y)$ .
- Discrete random variables — can only take some values (usually integer values); almost always, they are based on counts of some sort. The distribution of a discrete random variable is specified by the probabilities corresponding to each possible value that the random variable may take. [This is called the *probability mass function*]. For example:

$$p(x) = \Pr(X = x), \quad (x = 0, 1, 2, \dots)$$

- Distribution of the sample mean —  $\bar{X}$ :  $E(\bar{X}) = \mu$ ,  $\text{var}(\bar{X}) = \sigma^2/n$ .  
And further, for large  $n$ ,  $\bar{X} \xrightarrow{d} N(\mu, \sigma^2/n)$  (by the central limit theorem).
- Dunnett’s procedure for multiple comparisons — a method for adjusting for multiple comparisons of a set of group means against a control group mean.
- Error
  - random error — a random error is simply a name for a random element indicating a deviation from the mean. It is not a fault or a mistake.
  - type I error — the error of rejecting the null hypothesis when it is true.
  - type II error — the error of accepting the null hypothesis when it is false.
- Estimate — a sample statistic; a value computed from the sample which is used to indicate the value of the unknown population characteristic; it is often the corresponding sample characteristic. Thus the sample mean,  $\bar{x}$  estimates the population mean  $\mu$ . An estimate is often indicated by a “hat”: for example, the estimate of the parameter  $p$  is denoted by  $\hat{p}$ .
- Estimator — a random variable, the realisation of which is the estimate. Thus the estimator, being a random variable, has a probability distribution; and it has a mean and a variance and so on. The estimate is a number.  
For example: the random variable  $\bar{X}$  is an estimator (of  $\mu$ ) — it has a distribution, a mean and a variance; the number  $\bar{x}$  is an estimate (of  $\mu$ ) — it does not have a distribution, it is just a number.
- Experimental design — the planning of a statistical experiment with the intention that (i) the conclusions reached are valid; and (ii) the results are as precise as possible.
- Experimental unit — the object to which the treatment is applied in a statistical experiment, and in which the variable under investigation is measured. Experimental units are often referred to as plots — because much of the early work on the design of statistical experiments was done by R. A. Fisher on agricultural experiments.

- F-distribution — the distribution of a ratio of independent sample variances (with sampling from a normal population); alternatively the  $F_{m,n}$  distribution is obtained from the scaled ratio of  $\chi^2$ -distributions:  $F_{m,n} = (\chi_m^2/m)/(\chi_n^2/n)$ . The centre of an F-distribution is around 1.
- F-test — the basic test of the analysis of variance: since if the null hypothesis is true (i.e. if there are no effects — due to groups, or rows, or whatever) then the ratio of mean squares,  $F$ , is like a ratio of sample variances: and it has an F distribution. If the null hypothesis is not true the ratio tends to be larger. Thus large values of  $F$  indicate departure from the null hypothesis (i.e. there is a significant difference between the effects).
- Factorial experiment — an experiment to investigate the effects of a number of treatments (factors) by applying each factor at each possible level. For example, a  $2 \times 3$  experiment would have factor  $A$  applied at each of 2 levels and factor  $B$  applied at each of three levels. Thus there are  $2 \times 3 = 6$  factor combinations. This would lead to a two way classification with 2 rows (corresponding to the levels of  $A$ ) and 3 columns (corresponding to the levels of  $B$ ).
- Family error rate — the probability that at least one of a group or “family” of tests falsely rejects the null hypothesis.
- (data) Frequency representations — histograms, dotplots, frequency polygons, bar graphs (for nominal and discrete data), pie charts (for percentages, usually with nominal data), stem-and-leaf display
- Friedman test (two way ANOVA for ranked data) — rank-based test for comparison of  $k (> 2)$  treatment (row) effects, in the presence of block (column) effects.
- Histogram — a pictorial display which gives the frequencies (or relative frequencies) of various categories or intervals.
- [statistical] Hypothesis — a statement about the population distribution. For example  $\mu = 10$ ,  $\mu > 12$ ,  $\sigma < 0.1$ ,  $m \neq 20$ , etc.
- Independence — the random variables  $X$  and  $Y$  are independent if the distribution of  $X$  is unaffected by the value of  $Y$  (and the distribution of  $Y$  is unaffected by the value of  $X$ ); or, in other words, the distribution of  $X$ , given  $Y$ , is the same as the distribution of  $X$  without knowledge of  $Y$  (and vice versa).
- Independent samples t-test (independent samples test for normal data). [Note: the test remains valid for data which are close to normal — practically, this means that the data are not obviously non-normal.] This is a test for comparison of location using independent normal samples. It is based on the means and standard deviations of the two samples.

There are two versions (depending on whether or not it is reasonable to assume the variances are equal). [Note rank based tests essentially assume equal variances — and, as in the case of the rank-based test, wrongly assuming equal variances does not invalidate the test, it just makes it less powerful.]

equal variances: combine the sample variances to obtain a pooled estimate of  $\sigma^2$ :  $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}$  (a weighted average of the sample variances, with weights equal to their degrees of freedom). Then, the test statistic is given by

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Under  $H_0 (\mu_1 = \mu_2)$ ,  $T \stackrel{d}{=} t_{n_1+n_2-2}$ . (Our estimate of  $\sigma^2$  is now based on  $n_1 + n_2 - 2$  degrees of freedom.)

unequal variances: in this case the sample variances are not pooled, and the test statistic is given by

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Under  $H_0 (\mu_1 = \mu_2)$ ,  $T \stackrel{d}{\approx} t_\nu$ , where  $\nu$  is given by a complicated formula;  $\nu$  will always be between the smaller of  $n_1 - 1$  and  $n_2 - 1$ , and  $n_1 + n_2 - 2$ .

[This is seen by some as the safer alternative — it is the default on MINITAB — as it is close to the equal variances test when  $s_1^2 \approx s_2^2$ , but covers the case when they are clearly different.]

- Individual error rate — the probability that one particular individual test (of a family of tests) falsely rejects the null hypothesis.
- Integer distribution —  $I(k, l)$ . This is sometimes called the discrete uniform distribution. The random variable is equally likely to take any integer value from  $k$  to  $l$  inclusive. This is a discrete distribution with probability mass function

$$\Pr(X = x) = \frac{1}{k-l+1} \quad (x = k, k+1, \dots, l)$$

For example, if  $X \stackrel{d}{=} I(1, 6)$ , then

$$\Pr(X = x) = \frac{1}{6} \quad (x = 1, 2, \dots, 6)$$

This is the probability distribution of the number obtained when a fair die is rolled.

If  $X \stackrel{d}{=} I(k, l)$ , then

$$\text{E}(X) = \frac{1}{2}(k+l) \text{ and } \text{var}(X) = \frac{1}{12}(l-k)(l-k+2).$$

- Interaction effects — the interaction effects in a factorial experiment indicate the deviation from additivity of the effects of  $A$  and  $B$ : an interaction would indicate that factor  $A$  has a different effect at different levels of  $B$ . (The interaction effects correspond to  $\gamma$  in the two-way classification model below.)
- Interactive (non-additive) model for two way classification —  $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ , i.e. observation = overall mean + row effect + column effect + interaction effect + random error. The interaction is simply a deviation from the additive model — an indication that the row-effects are not the same in each column (or equivalently, that the column-effects are not the same in each row).

- [sample] Interquartile range (IQR) — the difference between the upper and lower quartiles:  $IQR = Q_3 - Q_1$ ; the interquartile range is therefore the range of the middle 50% of the data.
- Interval — a continuous set of values between two limits: the set of numbers on the continuous scale between  $a$  and  $b$ , i.e.,  $a < x < b$ , is an interval.
- Kruskal-Wallis test (one way ANOVA for ranked data) — rank-based test for comparison of location of  $k (> 2)$  populations, based on independent samples.
- Kurtosis — an indicator of the bell-shape of a distribution: a distribution with positive kurtosis has longer tails than a normal distribution and is more peaked in the middle; a distribution with negative kurtosis has shorter tails and is flatter in the middle; a normal distribution has zero kurtosis. Kurtosis is unaffected by location or spread.

With an alternative definition, the normal distribution has kurtosis 3 — in which case the shape is indicated by whether the kurtosis is greater or less than 3.

- Latin square design — allows for two-directional blocking. The design requires that there be equal numbers of rows, columns and treatments, and that each treatment be used exactly once in each row and each column.
- Line plot (for bivariate data) — the  $(x, y)$  data values are plotted and points with adjacent  $x$ -values are joined by straight line segments. For a line plot to be appropriate, the  $x$ -variable has to have a natural ordering (usually time) and there can be only one  $y$ -value for each distinct  $x$ -value.
- Linear transformations:  $Y = c + kX \Rightarrow E(Y) = c + kE(X)$  and  $\text{var}(Y) = k^2\text{var}(X)$ , so that  $\text{sd}(Y) = |k|\text{sd}(X)$ .

Note:  $(k = 0) \quad E(c) = c, \text{var}(c) = 0;$

$(c = 0) \quad E(kX) = kE(X), \text{var}(kX) = k^2\text{var}(X).$

$[E(-X) = -E(X), \text{var}(-X) = \text{var}(X).]$

- Linear regression — It is often assumed, as we do, that the regression is a linear function: so that, given  $X = x$ , the mean of  $Y$  is  $\alpha + \beta x$ . This may also be expressed by writing:  $y = \alpha + \beta x + e$ , where  $e$  is random error (with zero mean).
- Main effects — the effects of the factors  $A, B, \dots$  in a factorial experiment (which correspond to  $\alpha$  and  $\beta$  in the two-way classification model): see Additive model, Interactive (non-additive) model.
- [population] Mean — the mean of a population or of a random variable  $X$ , specifying the population, is denoted by  $\mu$  or  $\mu_X$  or  $E(X)$ . It is the weighted average of values that  $X$  can take, where the weights are provided by the distribution of  $X$ . It is at the “centre of mass” of the distribution. Sometimes the term the “expectation of  $X$ ” is used, which is where the notation  $E(X)$  originates.
- [sample] Mean (or average), or the mean of a data set — the mean of a set of observations on the variable  $x$  is usually denoted by  $\bar{x}$ ; it is just the average of the values, that is  $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum x_i$ .

- Mean square — a sum of squares divided by its number of degrees of freedom.
- Measures of location are also referred to as measures of central tendency: see Mean, Median, Trimmed mean.
- Measures of spread are also referred to as measures of scale: see Variance, Standard deviation, Interquartile range, Range.
- [population] Median — the population 0.5-quantile: about half the population is above the median (and half below). See [population] Quantiles.
- [sample] Median, or the median of a data set — if the values are arranged in ascending order then the median is given by the ‘middle’ observation: if  $n$  is odd the median is the middle value, and if  $n$  is even the median is the average of the two middle values. It is often denoted by  $M$ . Note: about half the sample is above the sample median and half below.
- Method of least squares — a method of estimation by which the values of estimates are chosen so that the sum of squares of the residuals is minimised (residual = observed – fitted).
- Multiple comparisons — when performing a large number of tests, each with type I error rate  $\alpha$ , the probability that at least one of these tests falsely rejects the null hypothesis can be quite large; thus — according to the conventional multiple comparisons argument — some allowance needs to be made for this when multiple tests are carried out.
- Multiple correlation — a measure of association between one variable and two or more other variables. It can be expressed as the correlation between two, appropriately defined, variables.
- Multiple regression — In the real world, there are often several explanatory variables,  $x_1, x_2, \dots, x_p$ , rather than only one. So we may want to consider how they jointly affect the outcome,  $Y$ . Sometimes  $Y$  is called the “dependent” variable, and the  $xs$  are called the “independent” variables. In this course, we always assume the multiple regression is linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e.$$

Again the  $\beta$ s are estimated by the method of least squares

- Normal distribution —  $N(\mu, \sigma^2)$ . This is a commonly occurring continuous distribution with probability density function taking the form of the classical “bell-shaped curve”. The probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

If  $X \stackrel{d}{=} N(\mu, \sigma^2)$ , then  $E(X) = \mu$  and  $\text{var}(X) = \sigma^2$ .

- Null hypothesis  $H_0$  — the hypothesis of no (= null) effect.  
We do not reject  $H_0$  unless there is strong evidence against it — in favour of the alternative hypothesis  $H_1$ .
- One-sided alternative hypothesis — an alternative hypothesis of the form  $\theta > \theta_0$  or of the form  $\theta < \theta_0$ .

- One-way ANOVA — an analysis of variance for data classified in one way (usually designated as groups). The ANOVA produces an F-test for the null hypothesis that there is no group effect (or that the groups all have the same mean).
- Outliers — observations or data values that apparently do not belong with the rest of the sample or data set. They may be mistakes (mis-reading the scale, decimal point in the wrong place, incorrect number transcribed, data entry error) or they may be important. They should be identified and investigated.
- *P*-value — the probability of obtaining a result at least as extreme as that actually obtained, if the experiment were to be repeated with the null hypothesis true.

$P < 0.05$  = “significant” = “significant at the 5% level”,

$P < 0.01$  = “very significant” = “significant at the 1% level”,

$P < 0.001$  = “extremely significant” = “significant at the 0.1% level”.

- Paired t-test (paired samples test for normal data) — test for comparison of location using paired samples. It is based on the actual differences of the paired observations. It is essentially a *t*-test that the population from which the pairs are drawn has mean zero.

The test statistic  $T = \frac{\bar{D}}{S_D/\sqrt{n}}$ , where  $\bar{D}$  and  $S_D$  denote the mean and standard deviation of the sample of differences, and  $n$  the number of pairs. Under  $H_0$  (that  $\mu_D = 0$ ),  $T \stackrel{d}{=} t_{n-1}$ .

- Parameter — a population characteristic, which is usually unknown and the subject of our statistical enquiries; it is often denoted by a Greek letter, like the population mean  $\mu$ .
- Partial correlation — the correlation between two variables when both have been adjusted for one or more other variables. It can be expressed as the correlation between two, appropriately defined, variables.
- Polynomial regression — a special case of multiple regression where the explanatory variables are powers of one basic explanatory variable:  $x_k = x^k$ . Thus:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + e.$$

- Population — a population is the entire collection of people or things in which we are interested. Populations may be finite or infinite, they may be real or hypothetical and they may be easy or difficult to define.
- Power of a test — the probability of making a correct decision when  $H_1$  is true,  
i.e.,  $\Pr(\text{reject } H_0 \text{ given that } H_1 \text{ is true})$ . Thus, power =  $1 - \Pr(\text{type II error})$ . Note that *P*-value and power are very different. They both start with P, and they both relate to statistical hypothesis testing, but they have little else in common.

- Power transformations — a transformation of the (positive-valued) variable  $x$  using a power:  $y = x^p$ , with  $y = \log x$  for  $p = 0$ . Usually the convention is used that for negative  $p$ , we use  $y = -x^p$  so that the order of the  $ys$  is the same as the order of the  $xs$ .
- Practical significance — is not the same as statistical significance. The result that the mean increase is significantly different from zero does not imply that this increase has any practical significance: for example, the mean increase may be 0.1 whereas it may need to be at least 1.0 to be of practical significance.
- Prediction interval — a 95% prediction interval for  $Y$  is an interval within which a future value of  $Y$  will lie with probability 0.95.
- Probability density, Cumulative probability and Inverse cumulative probability on MINITAB

(MINITAB) probability density = pdf or pmf (according as to whether the distribution is continuous or discrete).

(MINITAB) cumulative probability = cdf

for a random variable  $X$  it is obtained as follows. For any value  $x$ , the cumulative probability is the value  $q_x$  that  $X$  is less than or equal to  $x$ ; i.e.,  $q_x = \Pr(X \leq x)$ .

(MINITAB) inverse cumulative probability = inverse cdf = quantiles

for a random variable  $X$  is obtained as follows. For any probability  $q$ , the inverse cumulative probability is the value  $x_q$  such that the random variable has a probability  $q$  of being less than or equal to  $x_q$ , i.e.,  $\Pr(X \leq x_q) = q$ .  $x_q$  is also called the  $q$ -quantile of  $X$  and is also denoted by  $c_q(X)$ .

- Probability distribution — is the specification of the probabilities attached to the possible values of a random variable. This specification may be achieved by a probability mass function (for discrete random variables), a probability density function (for continuous random variables) or a cumulative distribution function (for any type of random variable).

- [population] quantiles (inverse cdf) — the  $q$ -quantile of a random variable  $X$ , usually denoted by  $c_q$ , is a number such that  $\Pr(X \leq c_q) = q$ . In the case of a discrete random variable, such a number may not exist; in which case we define  $c_q$  as the smallest number such that  $\Pr(X \leq c_q) \geq q$ .

The 0.5-quantile,  $c_{0.5}$ , is the median; the 0.25-quantile,  $c_{0.25}$ , and the 0.75-quantile,  $c_{0.75}$ , are the lower and upper quartiles respectively.

- [sample] quantiles — the  $q$ -quantile of a sample is (roughly) a number such that a proportion  $q$  of the sample is less than or equal to it. It thus reflects the definition of a population quantile, and indeed the sample quantile is an estimate of the corresponding population quantile. The sample  $q$ -quantile is usually denoted by  $\hat{c}_q$ .

Our rule is to define  $\hat{c}_q$  as the value in the  $k$ -th position when the  $n$  sample values are placed in ascending order, where  $k = (n+1)q$ . [See the definitions of the sample median ( $\hat{c}_{0.5}$ ) and the sample quartiles ( $\hat{c}_{0.25}$  and  $\hat{c}_{0.75}$ ). ]

- [sample] Quartiles — if the  $n$  data values are arranged in ascending order then the lower quartile  $Q_1$  is at position  $(n + 1)/4$  and the upper quartile  $Q_3$  is at position  $3(n + 1)/4$ . If the position is not an integer, interpolation is used. [ $Q_2$  is the median: it is at position  $(n + 1)/2$ .]
- Random error — a random error is simply a name for a random element indicating a deviation from the mean. It is not a fault or a mistake. See Additive model, Interactive (non-additive) model
- Random sample — A random sample of size  $n$  is such that any selection of  $n$  items from the population is equally likely to be chosen.

Mathematically, a random sample from an infinite population can be represented by  $X_1, X_2, \dots, X_n$  where the random variables are independent and identically distributed (iidrvs) each having the same distribution  $\mathcal{D}$  — the population distribution. This is sometimes described as a random sample on  $X$  (i.e., a random sample from an infinite population with a distribution specified by the distribution of  $X$ ).

For example if the population distribution is  $N(50, 10^2)$ , then a random sample on this population is represented by  $X_1, X_2, \dots, X_n$  where each  $X_i$  is distributed as  $N(50, 10^2)$ .

- Random variable — a random variable is a variable whose value depends on the outcome of a random phenomenon. In this usage, “random” does not mean haphazard or chaotic, but simply uncertain: before we make an observation, we do not know what its value will be. For example, a random variable might be a count, a measure on a continuous scale, a binary (zero-one) variable, a proportion, or an average.

The realisation (or observed value) of a random variable is the value actually observed. Thus when we roll a fair die, the number observed is equally likely to be  $1, 2, \dots, 6$ . When it has been rolled the number actually obtained: 4, say, is the realisation of this random variable. We use capital letters ( $X, Y, \dots$ ) to denote random variables, and lower case letters ( $x, y \dots$ ) to denote the *observed values or realisations* of the random variable: see Realisation.

- Randomisation — treatments should be allocated to the experimental units at random. This guards against possible bias and provides justification for the assumptions that are usually made when the results are analysed.
- Randomised block design — this design requires that each treatment be used the same number of times (usually once) in each block. Within each block, treatments are allocated to units at random with a separate randomisation for each block.
- (completely) Randomised design — treatments are assigned to the experimental units at random, subject only to the number of times each treatment is to be used.
- [sample] Range — the difference between the largest and smallest values; the range is the simplest measure of spread, but it is generally used only for small samples as it is too easily affected by ‘outliers’.
- Rank correlation — a sample measure of association, which is little affected by outliers and unaffected by non-linearity. The  $x$ -values and  $y$ -values are

ranked separately (from smallest to largest); the values are replaced by the ranks and the correlation coefficient evaluated for the ranks. Thus the rank correlation is a correlation of the ranks. It is sometimes called Spearman's rank correlation.

In the absence of ties the ranks take the values  $1, 2, \dots, n$  (in some order or other). As a consequence the formula for the rank correlation can be simplified to

$$r' = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where  $d_i$  denotes the difference in the  $x$ -rank and the  $y$ -rank for the  $i$ -th point. In this form it is often referred to as *Spearman's rho*. This formula becomes approximate if there are ties, and may be substantially in error if there are many ties.

- Rank sum test, the Mann-Whitney test (independent samples test for ranked data) — test for comparison of location using independent samples. It is based on the relative magnitudes (ranks) of the observations in the two samples. It requires no assumptions other than the independence of the observations. See also Ranking.

Test statistic  $W_1$  = sum of ranks in sample 1 (having ranked all the observations in both samples from 1 to  $n_1 + n_2$ ). Under  $H_0$  (that the populations are identical),  $W_1 \stackrel{d}{\approx} N(\frac{1}{2}n_1(n_1 + n_2 + 1), \frac{1}{12}n_1n_2(n_1 + n_2 + 1))$ . This approximation is quite good even for small  $n_1$  and  $n_2$ ; MINITAB always uses this approximation. Critical values are tabulated: in tables it is usually assumed that  $n_1 \leq n_2$ , so it may be necessary to reverse the samples.

[This test is analogous to a  $t$ -test for independent samples, but with the observations replaced by their ranks.]

- Ranking — assigning ranks to a set of observations. The observations are ordered from smallest to largest, the smallest value is given rank 1, the next smallest rank 2, and so on up to the largest observation which is given rank  $n$ .

It is often the case that some of the observations are equal, or "tied". In that event, each of the tied observations is assigned the average of the ranks the observations would have got were they not tied. For example

obs	2.3	3.4	3.5	3.6	4.1	4.6	4.7	5.9
rank	1	2	3	4	5	6	7	8
obs	2.3	3.5	3.5	3.5	4.1	4.7	4.7	5.9
rank	1	3	3	3	5	6.5	6.5	8

Similarly, if (in a larger sample) six tied observations would have got ranks 11, 12, 13, 14, 15 and 16 had they not been tied, then they are each assigned rank 13.5. (The effect of this is that the mean rank is unaltered by ties, but the variance is slightly altered.)

- Realisation (or observed value) of a random variable — is the value actually observed. Thus when we roll a fair die, the number observed is equally likely to be  $1, 2, \dots, 6$ . When it has been rolled the number actually obtained: 4 say, is the realisation of this random variable.

Note: we use capital letters ( $X, Y, \dots$ ) to denote random variables, and lower case letters ( $x, y \dots$ ) to denote the *observed values* or *realisations* of the random variable.

- Regression — the ‘regression’ of  $Y$  on  $X$  is the conditional mean of  $Y$  given the value of  $X$ . If  $Y$  tends to be large when  $X$  is large and small when  $X$  is small then the regression is an increasing function.

Regression is used for prediction: it is the best predictor of  $Y$  based on  $X$ .

‘Regression’ is a silly name for what is just a mean value (or a set of mean values — one for each value of  $X$ ). The name comes from the fact that the conditional mean tends to be closer to the overall mean (giving a safer and better prediction): early researchers thought this meant a regression to the mean and some sort of convergence. In fact, it just indicates that you are likely to make a smaller error if you use a predictor which is closer to the overall mean.

- Replication — a repetition of the complete set of factor combinations. Using the same factor combinations a number of times not only increases the precision of estimators but it also enables us to estimate the precision by estimating the error variance.
- Residual analysis — examination of the residuals to determine whether their behaviour approximates the behaviour that random error terms should have: if the model is correct then the residuals should approximate random error.
- Residuals — the difference between the observed value  $y_i$  and its fitted mean  $\hat{\mu}$  estimated with a given model. For example, with straight line regression,  $\hat{\mu} = \hat{\alpha} + \hat{\beta}x_i$  and the  $i$ th residual  $\hat{e}_i = y_i - \hat{\alpha} + \hat{\beta}x_i$ .

The residual sum of squares is the sum of the squared residuals:  $\sum \hat{e}_i^2$ ; and the residual mean square is the residual sum of squares divided by the number of degrees of freedom (which is usually the number of observations minus the number of parameters estimated in the fitted model). The residual mean square is used to estimate the error variance and is usually denoted by  $s^2$ .

- Sample — a sample is a subset of a population. One type of sample which is of special interest in statistical investigations is the so-called *random sample*: see Random sample
- Sample mean — a random variable  $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$ . It is a random variable and therefore has a probability distribution [see Distribution of the sample mean]. The realisation or observed value of the sample mean is denoted by  $\bar{x}$ ; this is simply the average of the observed values in the sample.
- Sample size —  $n$ , the number of observations in the sample. One of the most common question in statistics is “How big a sample should I take?”. To determine an appropriate sample size, you first need to specify what you require: for example width of confidence interval or power against a specified alternative. Or indeed whether it’s worthwhile to sample at all! (Miettinen).
- Scatter plot (for bivariate data) — the  $(x, y)$  data values are plotted as points relative to the coordinate axes ( $x$  on the horizontal axis and  $y$  on the vertical axis). In general this will give a (more or less) scattered cloud of points which

is indicative of the relationship between the variables. It is not appropriate to join up the points as in a line plot, though it might be appropriate to plot a line, or curve, 'of best fit'.

- Scatter plot matrix (for multivariate data) — consists of an array of scatter plots, with each variable plotted against each other variable.
- Sign test — test for comparison of location using paired samples. It is based only on the number of positive and negative differences and therefore makes minimal assumptions. This also means it is less powerful when more assumptions are appropriate. The test statistic  $Z = \text{number of positive differences}$ ; under  $H_0$  (that the populations are identically distributed),  $Z \stackrel{d}{=} \text{Bi}(n^*, \frac{1}{2})$  (where  $n^*$  denotes the number of non-zero differences).
- Signed rank sum test, the Wilcoxon matched-pairs signed rank test (paired samples test for ranked data) — test for comparison of location using paired samples. It is based not only on the sign of the differences but also their relative magnitudes, by way of their ranks. See also Ranking.  
The test statistic  $T = \text{sum of ranks of magnitudes of positive differences}$  (having ranked the magnitudes without regard to sign). Critical values are tabulated. Under  $H_0$  (the populations are identically distributed),  $T \stackrel{d}{\approx} N(\frac{1}{4}n(n+1), \frac{1}{24}n(n+1)(2n+1))$ .
- Significance level of a test (level of significance,  $\alpha$ ; size) — the probability of making a wrong decision when  $H_0$  is true,  $\alpha = \Pr(\text{reject } H_0 \text{ given that } H_0 \text{ is true})$ , i.e., the probability of a type I error. This is sometimes referred to as the size of the test (which is different from the sample size,  $n = \text{number of observations}$ ).
- Skewness — a measure of asymmetry of a distribution: a positively skew distribution has a longer tail to the right (positive end of the line); a negatively skew distribution has a longer tail to the left; zero skewness indicates symmetry. Skewness is unaffected by location or spread.
- Split-plot experiments — if it is necessary to confound effects, it is usual to confound higher order interactions rather than main effects, but there are circumstances under which it is necessary, due to practical limitations, to confound one or more main effects with blocks. Experiments of this type are referred to as split-plot experiments.
- [population] standard deviation — the standard deviation of a population or of a random variable  $X$ , specifying the population, is the square root of the variance. It is denoted by  $\sigma$  or  $\sigma_X$  or  $\text{sd}(X)$ . Thus  $\text{sd}(X) = \sqrt{\text{var}(X)}$ .
- [sample] Standard deviation,  $s$ ; the square-root of the [sample] variance. The standard deviation (and not the variance) is in the same units as the original values. For many samples, approximately 95% of the sample will be within 2 standard deviations of the mean. Calculators:  $\text{SD}_{n-1}$ ; MINITAB: **STDEV**.
- Standard error — estimate of the standard deviation of an estimator. For example,  $\text{sd}(\bar{X}) = \sigma/\sqrt{n}$ , and  $\sigma$  is estimated by  $s$ , so the standard error of  $\bar{X}$  is given by  $\text{se}(\bar{X}) = s/\sqrt{n}$ . This is sometimes called the *standard error of the mean*.

An approximate 95% confidence interval is often given by estimate  $\pm 2$  standard error.

- standard normal distribution —  $N(0, 1)$ . This is the normal distribution with mean 0 and variance 1.
- Standardised range distribution  $Q_{k,r}$  — the distribution of the range of a sample of  $k$  normal random variables divided by an estimate of their standard deviation (which has  $r$  degrees of freedom).
- Standardised residuals — residuals adjusted so that they have unit variance; they already have zero mean. This adjustment entails division of each residual by a factor which estimates its standard deviation.
- standardised variable —  $z = \frac{x - \mu}{\sigma}$  and  $\hat{z} = \frac{x - \bar{x}}{s}$ .

A standardised variable has mean zero and variance one. Any variable can be standardised by subtracting off its mean and then dividing by its standard deviation. There are two versions: a population version ( $z$ ) which makes the population mean zero and population variance one; and a sample version ( $\hat{z}$ ) which makes the sample mean zero and the sample variance one.

- Statistical hypothesis — a statement about the population distribution. For example  $\mu = 10$ ,  $\mu > 12$ ,  $\sigma < 0.1$ ,  $m \neq 20$ , etc.
- Statistical significance — a result is said to be statistically significant if the null hypothesis is rejected (and  $P < 0.05$  etc.).
- Sums — the mean of a sum is the sum of the means:  
 $E(X+Y) = E(X)+E(Y)$ ;  $E(X_1+X_2+\dots+X_n) = E(X_1)+E(X_2)+\dots+E(X_n)$ .  
 and if the variables are independent, the variance of a sum is the sum of the variances.  
 $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ ;  $\text{var}(X_1 + X_2 + \dots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \dots + \text{var}(X_n)$ .

- Sums of squares (SS) — a sum of squares is what it says: a sum of squared terms like  $z_1^2 + z_2^2 + \dots + z_k^2$ ; for example, Total SS.
- t-distribution — If  $W \stackrel{d}{=} N(\mu, c\sigma^2)$ , then  $\frac{W - \mu}{\sigma\sqrt{c}} \stackrel{d}{=} N(0, 1)$ . If however,  $\sigma$  is replaced by an estimate with  $k$  degrees of freedom then  $\frac{W - \mu}{\hat{\sigma}\sqrt{c}} \stackrel{d}{=} t_k$ , where  $t_k$  denotes the t-distribution with  $k$  degrees of freedom.

In most applications,  $W$  denotes an estimator based on normal data. The simplest standard application is  $\frac{\bar{X} - \mu}{s/\sqrt{n}} \stackrel{d}{=} t_{n-1}$  for a sample from a normal distribution — in which  $W = \bar{X}$  and  $c = 1/n$ .

- Test statistic — a statistic on which the decision to accept or reject  $H_0$  is based. If a test statistic is to be any good, it must be sensitive to departures from  $H_0$ . In general, it is usually better to use a test statistic whose validity does not require stringent assumptions — unless you are very confident of their correctness.
- Treatment — the aim of a statistical experiment is to compare the effects of a number of treatments on the yield.

- [sample] ( $\alpha\%$ ) Trimmed mean — the mean of the values that remain after the largest and the smallest  $\alpha\%$  of the observations are omitted. MINITAB calculates a 5% trimmed mean as part of its basic descriptive statistics output: TRMEAN. The sample median can be thought of as a 50% trimmed mean, the sample mean as a 0% trimmed mean.
- Tukey's method for multiple comparisons — a method for multiple comparisons between group means based on the standardised range distribution: using the fact that any difference between any two of  $k$  variables is less than the range of the variables.
- Two-sided alternative hypothesis — an alternative hypothesis of the form  $\theta \neq \theta_0$ .  
The  $P$ -value for a two-sided alternative is taken to be twice the  $P$ -value for the one-sided hypothesis with the direction of the observed value of the test statistic.
- Two-way ANOVA — an analysis of variance for data classified in two ways (usually designated as rows and columns, producing a data table with row classifications and column classifications). The ANOVA produces F-tests for the null hypothesis that there is no row effect, and for the null hypothesis that there is no column effect.
- Type I error — rejecting  $H_0$  when  $H_0$  is true.
- Type II error — accepting  $H_0$  when  $H_1$  is true.
- Variable types
  - categorical (= nominal)
  - ordinal
  - numerical — discrete or continuous

It should be noted that these variable types are hierarchical:

$$\begin{aligned} \text{categorical} &\ll \text{category} \\ \text{ordinal} &\ll \text{category + order} \\ \text{numerical} &\ll \text{category + order + scale} \end{aligned}$$

Thus an ordinal variable can be treated as a categorical variable (ignoring the ordering); and a numerical variable can be treated as an ordinal variable (ignoring the scaling) or as a categorical variable (ignoring the ordering and the scaling).

Numerical variables can be further subdivided into discrete or continuous (according to whether or not there is a rounding error in observing the variable).

- [population] Variance — the variance of a population or of a random variable  $X$ , specifying the population, is denoted by  $\sigma^2$  or  $\sigma_X^2$  or  $\text{var}(X)$ . It is the weighted average of squared deviations from the mean of  $X$ , where the weights are provided by the distribution of  $X$ .

- [sample] Variance,  $s^2$ ; roughly, the average of the squares of the deviations of the data values from the sample mean,  $\bar{x}$ :

$$s^2 = \frac{1}{n-1} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\}$$

$$= \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

- $\bar{x}, \bar{X}$  (x-bar, X-bar) — see Sample mean.