Sujit K. Sahu

Introduction to Probability, Statistics & R

Foundations for Data-Based Sciences

Springer

Introduction to Probability, Statistics & R

Sujit K. Sahu

Introduction to Probability, Statistics & R

Foundations for Data-Based Sciences

Springer

Sujit K. Sahu
School of Mathematical Sciences
University of Southampton
Southampton, UK

ISBN 978-3-031-37864-5 ISBN 978-3-031-37865-2 (eBook)
https://doi.org/10.1007/978-3-031-37865-2

Mathematics Subject Classification: 62-01, 60-01

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

To my family, parents, teachers and
professors who taught me everything.

Preface

It is a daunting task to contemplate writing an introductory textbook in Statistics and Probability since there are so many excellent text books already available in the market. However there seems to be a lack of textbooks/literature detailing statistical software `R` which contain sufficient levels of mathematical rigour whilst also remaining unambigious and comprehensible to the reader. There are even fewer introductory books which can be adopted as entry level texts for degree programmes with a large mathematical content, such as mathematics with statistics, operations research, economics, actuarial science and data science.

The market is saturated with a vast number of books on introductory mathematical statistics. Among the most prominent books in this area are the ones written by Professor Robert Hogg and his co-authors, e.g. the book *Probability and Statistical Inference*, 8th Edition (2010) by Hogg and Tanis (Pearson). Laid out in 11 chapters, this book is one of the best out there for learning mathematical statistics. Competitive books include: (i) *Probability and Statistics* by M. H. Degroot and M. J. Schervish (Pearson), (iii) *Statistical Inference* by G. Casella and R. Berger (Duxbury), (iii) *A First Course in Probability* by S. A. Ross (Pearson), (iv) Mathematical statistics with applications by D. D. Wackerly, W. Mendenhall and R. L. Scheaffer (Duxbury). These books assume a higher level of preparedness in mathematical methods that the typical first year undergraduate students do not have. Also these texts typically do not provide a plethora of examples that can help put the target audience of the first year undergraduate students at ease. Such students, fresh out of secondary school, are used to seeing lots of examples in each topic they studied. Universities in USA mostly adopt such text books in their masters level statistics courses. Lastly, none of these books integrate `R` in their presentation of the topics.

Apart from the above list, the book most relevant to the current textbook is *Introductory Statistics with R* by Peter Dalgaard published by Springer in 2008. Dalgaard's book is more targeted for a biometric/medical science audience whereas the current textbook targets students in a wider field of data-based and mathematical sciences. For example, Dalgaard's book include multiple regression and survival analysis. Multiple regression is too advanced for first year and survival analysis is too advanced for even second year students, who still are in the process acquiring skills in statistical inference and modelling. Also, unlike the Dalgaard's book, the current textbook does not assume knowledge of basic statistics to start with and hence is appealing to a wider audience of mathematics students who have not learned probability and statistics in their previous studies. The book *Teaching Statistics: A Bag of Tricks* by Andrew Gelman and Deborah Nolan (2nd Edition), published by the Oxford University Press, discusses many excellent methods for teaching practical statistics. However, this book is concerned about teaching statistics to aspiring applied scientists as well as mathematicians.

The current book aims to fill this gap in the market by taking a more direct targeted approach in providing an authentic text for introducing both **mathematical and applied statistics** with R. The book aims to provide a gentle introduction by keeping in mind the knowledge gap created by previous, ether none or non-rigorous, studies of statistics without the proper and rigorous use of mathematical symbols and proofs. This self-contained introductory book is also designed to appeal to first time students who were not previously exposed to the ideas of probability and statistics but have some background in mathematics. Many worked examples in the book are likely to be attractive to them, and those examples will build a transition bridge from their previous studies to university level mathematics. Moreover, integration of R throughout is designed to make learning statistics easy to understand, fun and exciting.

The book is presented in five parts. Part I (Chaps. 1 and 2) introducing basic statistics and R does not assume knowledge and skills in higher level mathematics such as multivariate calculus and matrix algebra. Part II (Chaps. 3 to 8) introduces standard probability distributions and the central theorem. Part III (Chaps. 9 to 12) introduces basic ideas of statistical inference. As a result, and quite deliberately, this part presents statistical inference methods such as the t-tests and confidence intervals without first deriving the necessary t and χ^2 distributions. Such derivations are delayed until the later chapters in Part IV. In this part (Chaps. 13 to 16), we present materials for typical second year courses in statistical distribution theory discussing advanced concepts of moment generating functions, univariate and bivariate transformation, multivariate distributions and concepts of convergence. Both this and the final Part V (Chaps. 17, 18 and 19) assume familiarity of results in multivariate calculus and matrix algebra. Part V of the book is devoted to introducing ideas in statistical modelling, including simple and multiple linear regression and one way analysis of variance. Several data sets are used as running examples, and dedicated R code blocks are provided to illustrate many key concepts such as the Central Limit Theorem and the weak law of large numbers. The reader can access those by installing the accompanying R package `ipsRdbs` in their computer.

I am highly indebted to all of my current and past mathematics and statistics colleagues in the Universities of Cardiff and Southampton, especially: Brian Bailey, Stefanie Biedermann, Dankmar Böhning, Russell Cheng, Jon Cockayne, Frank Dunstan, Jon Forster, Steven Gilmour, Terence Iles, Gerard Kennedy, Alan Kimber, Susan Lewis, Wei Liu, Zudi Lu, John W. McDonald, Robin Mitra, Barry Nix, Helen Ogden, Antony Overstall, Vesna Perisic, Philip Prescott, Dasha Semochkina, T. M. Fred Smith, Peter W. Smith, Alan Welsh, Dave Woods, Chieh-Hsi Wu, and Chao Zheng, whose lecture notes for various statistics courses inspired me to put together this manuscript. Often I have used excerpts from their lecture notes, included their data sets, examples, exercises and illustrations, without their full acknowledgement and explicit attribution. However, instead of them, I acknowledge responsibility for the full content of this book.

I also thank all my bachelor's and master's degree students who read and gave feedback on earlier versions of my lecture notes leading to drafting of this book. Specifically I thank three Southampton BSc students: Mr Minh Nguyen, Mr Ali Aziz and Mr Luke Brooke who read and corrected a preliminary draft of this book. I also thank PhD students Mr Indrajit Paul (University of Calcutta), who introduced me to use the latex package tikz for drawing several illustrations, and Ms Joanne Ellison (Southampton), who helped me typeset and proofread Parts I–III of the book. Lastly, I thank two anonymous reviewers whose suggestions I incorporated to improve various aspects including coverage and presentation.

Winchester, UK Sujit K. Sahu

Contents

Part II Introduction to Probability

Part III Introduction to Statistical Inference

Part V Introduction to Statistical Modelling

Part I

Introduction to Basic Statistics and R

Introduction to Basic Statistics

1

Abstract

Chapter 1: This chapter introduces basic statistics such as the mean, median and mode and standard deviation. It also provides introduction to many motivating data sets which are used as running examples throughout the book. An accessible discussion is also provided to debate issues like: “Lies, damned lies and statistics” and “Figures don’t lie but liars can figure.”

1.1 What Is Statistics?

1.1.1 Early and Modern Definitions

The word *statistics* has its roots in the Latin word *status* which means the state, and in the middle of the eighteenth century was intended to mean:

$$\textit{collection, processing and use of data by the state.}$$

With the rapid industrialisation of Europe in the first half of the nineteenth century, statistics became established as a discipline. This led to the formation of the Royal Statistical Society, the premier professional association of statisticians in the UK and also world-wide, in 1834. During this nineteenth century growth period, statistics acquired a new meaning as the interpretation of data or methods of extracting information from data for decision making. Thus statistics has its modern meaning as the methods for:

$$\textit{collection, analysis and interpretation of data.}$$

Indeed, the Oxford English Dictionary defines *statistics* as: “*The practice or science of collecting and analysing numerical data in large quantities, especially for the*

S. K. Sahu, *Introduction to Probability, Statistics & R*,
https://doi.org/10.1007/978-3-031-37865-2_1

purpose of inferring proportions in a whole from those in a representative sample." Note that the word 'state' has been dropped from its definition. Dropping of the word 'state' reflects the wide spread use of statistics in everyday life and in industry—not only the government. The ways to compile interesting statistics, more appropriately termed as statistical methods, are now essential for every decision maker wanting to answer questions and make predictions by observing data.

Example questions from everyday life may include: will it rain tomorrow? Will the stock market crash tomorrow? Does eating red meat make us live longer? Is smoking harmful during pregnancy? Is the new shampoo better than the old as claimed by its manufacturer? How do I invest my money to maximise the return? How long will I live for? A student joining university may want to ask: Given my background, what degree classification will I get at graduation? What prospects do I have in my future career?

1.1.2 Uncertainty: The Main Obstacle to Decision Making

The main obstacle to answering the types of questions above is *uncertainty*, which means **lack of one-to-one correspondence between cause and effect**. For example, having a diet of (hopefully well-cooked!) red meat for a period of time is not going to kill someone immediately. The effect of smoking during pregnancy is difficult to judge because of the presence of other factors, e.g. diet and lifestyle; such effects will not be known for a long time, e.g. at least until the birth. Thus, according to a famous quote:

> **"Uncertainty is the only certainty there is, ..."**

Yet another quote claims, "In statistics, there is uncertainty over the past, present and future." This again emphasises the importance of presence of uncertainty in the data for the purposes of drawing conclusions with certainty.

1.1.3 Statistics Tames Uncertainty

It[1] is clear that we may never be able to get to the bottom of every case to learn the full truth and so will have to make a decision under uncertainty; thus we conclude that mistakes cannot be avoided when making decisions based on uncertain data. If mistakes cannot be avoided, it is better to know how often we make mistakes (which provides knowledge of the amount of uncertainty) by following a particular rule of

[1] This section is based on Section 2.2 of the book *Statistics and Truth* by C. R. Rao cited as Rao [15].

decision making. Such knowledge could be put to use in finding a rule of decision making which does not betray us too often, or which minimises the frequency of wrong decisions, or which minimises the loss due to wrong decisions. Thus we have the following equation due to Rao [15]:

$$\boxed{\text{Uncertain knowledge}} + \boxed{\begin{array}{l}\text{Knowledge of the extent of}\\ \text{uncertainty in it}\end{array}} = \boxed{\text{Usable knowledge}}$$

In the above equation uncertain data are noted as uncertain knowledge and the decisions we make based on data are denoted by the phrase usable knowledge. The amount of uncertainty, as alluded to in the middle box, is evaluated by applying appropriate statistical methods. Without an explicit assessment of uncertainty, conclusions (or decisions) are often meaningless guesses with vast amounts of uncertainty. Although such conclusions may turn out to be correct just by sheer chance, or luck, in a given situation, the methods used to draw such conclusions cannot *always* be guaranteed to yield sound decisions. A carefully crafted statistical method, with its explicit assessment of uncertainty, will allow us to make better decisions on average, although it is to be understood that it is not possible to guess exactly always correctly in the presence of uncertainty.

How does statistics tame uncertainty? The short answer to this question is by evaluating it. Uncertainty, once evaluated, can be reduced by eliminating the causes and contributors of uncertainty as far as possible and then by hunting for better statistical methods which have lower levels of uncertainty. Explicit statistical model based methods may help in reducing uncertainties. This book will illustrate such uncertainty reduction in later chapters. *Uncertainty reduction* is often the most important task left to the statisticians as any experimenter is free to guess about any aspects of their experiments. Indeed, in many practical situations results (and conclusions) are quoted without any mention (and assessment) of uncertainty. Such cases are dangerous as those may give a false sense of security implied by the drawn conclusions. The associated, perhaps un-evaluated, levels of uncertainty may completely overwhelm the drawn conclusions.

1.1.4 Place of Statistics Among Other Disciplines

Studying statistics equips the learner with the basic skills in data analysis and doing science with data in any scientific discipline. Statistical methods are to be used wherever there is any presence of uncertainty in the drawn conclusions and decisions. Basic statistics, probability theory, and statistical modelling provide the solid foundation required to learn and use advanced methods in modern data science, machine learning and artificial intelligence. Students studying mathematics as their major subject may soon discover that learning of statistical theories gives them the opportunity to practice their deductive mathematical skills on real life problems.

In this way, they will be able to improve at mathematical methods while studying statistical methods. This book will illustrate these ideas repeatedly.

The following quote by Prof. C. R. Rao, see Rao [15], sums up the place of statistics among other disciplines.

> "**All *knowledge* is**, **in final analysis**, ***history***.
> **All *sciences* are**, **in the abstract**, ***mathematics***.
> **All *judgements* are**, **in their rationale**, ***statistics***."

Application of statistics and statistical methods require dealing with uncertainty which one can never be sure about. Hence the mention of the word 'judgements' in the above quote. Making judgements requires a lot of common sense. Hence common sense thinking together with applications of mathematical and inductive logic is very much required in any decision making using statistics.

1.1.5 Lies, Damned Lies and Statistics?

Statistics and statistical methods are often attacked by statements such as the famous quotation in the title of this section. Some people also say, "you can prove anything in statistics!" and many such jokes. Such remarks bear testimony to the fact that often statistics and statistical methods are miss-quoted without proper verification and robust justification. It is clear that some people may intentionally miss-use statistics to serve their own purposes while some other people may be incompetent in statistics to draw sound conclusions, and hence decisions in practical applications. Thus, admittedly and regretfully, statistics can be very much miss-used and miss-interpreted especially by dis-honest individuals.

However, we statisticians argue:

- "Figures won't lie, but liars can figure!"
- "Data does not speak for itself"
- "Every number is guilty unless proved innocent."

Hence, although people may miss-use the tools of statistics, it is our duty to learn, question and sharpen those tools to develop scientifically robust and strong arguments. As discussed before, statistical methods are the only viable tool whenever there is uncertainty in decision making. It will be wrong to feel certainty where no certainty exists in the presence of uncertainty. In scientific investigations, statistics is an inevitable instrument in search of truth when uncertainty cannot be totally removed from decision making. Of-course, a statistical method may not yield the best predictions in every practical situation, but a systematic and robust application of statistical methods will eventually win over pure guesses. For example, statistical methods are the only definitive proof that cigarette smoking is bad for human health.

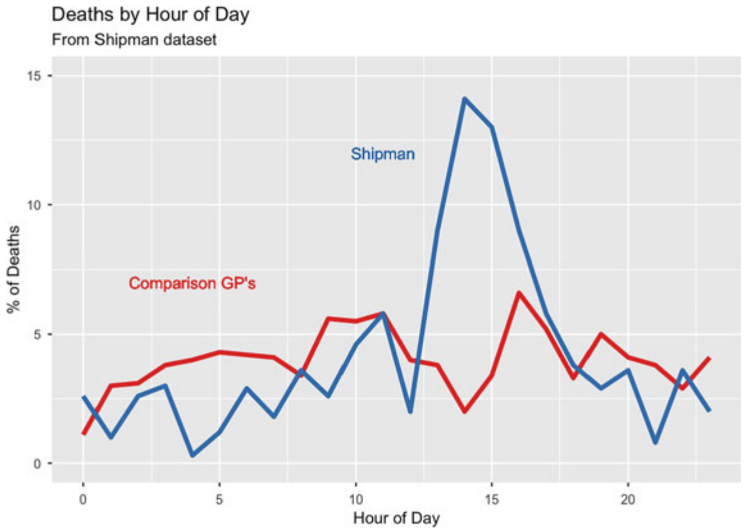

Fig. 1.1 The time at which Shipman's patients died, compared to the times at which patients of other local family doctors died. This is Figure 0.2 (reproduced here with permission) in the book "The Art of Statistics" by David Spiegelhalter

1.1.6 Example: Harold Shipman Murder Enquiry

To illustrate and motivate the study of statistics consider the Harold Shipman murder enquiry data example as discussed in the book, *The Art of Statistics* by Spiegelhalter [20]. Shipman was a British family doctor and serial killer who killed at least 215 of his most elderly patients by injecting opium between 1975 and 1998. His killing spree went undetected until an enquiry was launched during 1998–1999. He was finally convicted in January 2000. Figure 1.1 provides a graph of the percentages of patients dying in each of the 24 hours in a day. No sophisticated statistical analysis is required to detect the obvious pattern in the data, which shows that 2PM is the very unusual peak time of death for Shipman's patients. Further background and details are provided in Chapter 1 of the book by Prof David Spiegelhalter [20].

This example illustrates the importance of studying probability and statistics for data based sciences, although it did not require any sophisticated statistical methods about to be presented in the book. However, it is essential to learn a plethora of statistical methods to fully appreciate the strength of the evidence present in Fig. 1.1.

1.1.7 Summary

In summary, we note that statistical methods are, often, the only and essential tools to be applied whenever there is uncertainty and complete enumeration is not possible. Any analysis of empirically collected data must stand up to scientific (read as statistical) scrutiny and hence statistical knowledge is essential for conducting any scientific investigation. As a result, it is enormously advantageous to have good statistical data analysis skills to advance in most career paths in academia, industry and government.

This section has also discussed the main purpose of statistics—mainly to assess and to reduce uncertainty in the drawn conclusions. It also noted the place of statistics among different scientific disciplines and mathematics. Statistics and mathematics are best studied together as statistical applications provide rich training grounds for learning mathematical methods and mathematical theories and logic, on the other hand, help develop and justify complex statistical methods.

This section also tackles the often discussed misconceptions regarding the use of statistics in everyday life. It is often argued that statistics and statistical methods can be used to both prove or disprove a single assertion. This section has put forward the counter argument that data, being pure numbers, does not lie but users of statistics are liable to make mistakes either un-knowingly or knowingly through deceptions. A robust use of statistics is recommended so that the drawn conclusions can stand up to scientific scrutiny. Unfortunately, this task is to be taken by the producers of statistics so that only sound conclusions are reported in the first place.

For further reading, we note two accessible books: *Statistics and Truth* by Rao [15] and (ii) *The Art of Statistics* by David Spiegelhalter [20]. In addition, there are many online resources that discuss the joy of statistics. For example, we recommend the reader to watch the *YouTube* video **Joy of Statistics**.[2]

To acknowledge the references used for this book, we note the excellent textbooks written by Goon et al. [7], Casella and Berger [3], DeGroot and Schervish [4] and Ross [17]. We also acknowledge two books of worked examples in probability and statistics by Dunstan et al. [5, 6]. We also borrowed example exercises from the *Cambridge International AS and A-level Mathematics Statistics* (2012) book published by Hodder Education (ISBN-9781444146509) and written by Sophie Goldie and Roger Porkess.

1.2 Example Data Sets

Before introducing the example data sets it may be pertinent to ask the question, "How does one collect data in statistics?" Recall the definition of statistics in Sect. 1.1.1 where it states that statistics uses a representative sample to infer proportions in a whole, which we call the population. To collect a representative sample from a population the experimenter must select individuals randomly or haphazardly using a lottery for example. Otherwise we may introduce bias. For example, in order to gauge student opinion in a university, an investigator should not only survey the international students. However, there are cases when systematic sampling, e.g., selecting every third caller in a radio phone-in show for a prize, or

[2] https://www.youtube.com/watch?v=cdf0k545yDA.

sampling air pollution hourly or daily, may be preferable. This discussion regarding random sample collection and designed experiments, where the investigator controls the values of certain experimental variables and then measures a corresponding output or response variable, is deferred to Sects. 9.1 and 12.7. Until then we assume that we have data from n randomly selected sampling units.

For each of the n sampling units, we may collect information regarding a single or multiple characteristics. In the first case we conveniently denote the data by $x_1, x_2, \ldots, x_n$, so that these values are numeric, either discrete counts, e.g. number of road accidents, or continuous, e.g. heights of 18-year old girls, marks obtained in an examination. In case multiple measurements are taken, we often introduce more elaborate notations. For example, the variable of interest for the ith individual is denoted by y_i and the other variables, assuming p many, for individual i may be denoted by $x_{i1}, \ldots, x_{ip}$. The billionaires data set introduced below provides an example of this.

We now introduce several data sets that will be used as running examples throughout this book. Later chapters of this book may use the same data set to illustrate different statistical concepts and theories. All these data sets are downloadable from the online supplement of this book and also included in the `R` package `ipsRdbs` accompanying this book. Some of these data sets are taken from the data and story library.[3]

Example 1.1 (Fast Food Service Time)

The table (data obtained from the online Data and Story library in 2018) below provides the service times (in seconds) of customers at a fast-food restaurant (Fig. 1.2). The first row is for customers who were served from 9–10AM and the second row is for customers who were served from 2–3PM on the same day. The data set is `ffood` in the `R` package `ipsRdbs` with a help file obtained using the command `?ffood`.

AM	38,	100,	64,	43,	63,	59,	107,	52,	86,	77
PM	45,	62,	52,	72,	81,	88,	64,	75,	59,	70

Note that the service times are not paired, i.e. the times in the first column, 38 and 45, are not related. Those are time in seconds for two different customers possibly served by different workers in two different shifts. Issues that we would like to investigate include analyses of the AM and PM service times and comparison of differences between the times. ◀

[3] https://dasl.datadescription.com/.

Fig. 1.2 A fast food restaurant in Kyiv, Ukraine 2012. A photo by Sharon Hahn Darlin, source: Wikimedia Commons. https://www.flickr.com/photos/sharonhahndarlin/8088905486/. License: CC-BY-2.0

Table 1.1 Number of weekly computer failures over two years

4	0	0	0	3	2	0	0	6	7
6	2	1	11	6	1	2	1	1	2
0	2	2	1	0	12	8	4	5	0
5	4	1	0	8	2	5	2	1	12
8	9	10	17	2	3	4	8	1	2
5	1	2	2	3	1	2	0	2	1
6	3	3	6	11	10	4	3	0	2
4	2	1	5	3	3	2	5	3	4
1	3	6	4	4	5	2	10	4	1
5	6	9	7	3	1	3	0	2	2
1	4	2	13						

Example 1.2 (Computer Failures)

This data set (Table 1.1) contains weekly failures of a university computer system (See Fig. 1.3) over a period of two years. The source of the data set is the book ‘A Handbook of Small Data Sets’ by Hand et al. [8], thanks to Prof Jon Forster (author of Kendall et al. [9]) for sharing this. We will use this data set to illustrate commands in the `R` software package and also in statistical modelling. The data set is `cfail` in the `R` package `ipsRdbs` and there is a help file obtained by issuing the command `?cfail`. ◀

Example 1.3 (Number of Bomb Hits in London During World War II (See Fig. 1.4))

This data set is taken from the research article Shaw and Shaw [19] via the book by Hand et al. [8] and Prof Dankmar Böhning, (author of Böhning et al. [2]). The city of Greater London is divided into 576 small areas of one-quarter square

Fig. 1.3 PCs running Windows. Photo by Project Manhattan. License: CC BY-SA 3.0

Fig. 1.4 London Blitz. Photo by H. F. Davis

kilometre each. The number of bomb hits during World War II in each of the 576 areas was recorded. The table below provides the frequencies of the numbers of hits.

Number of hits	0	1	2	3	4	5	Total
Frequency	229	211	93	35	7	1	576

Thus, 229 areas were not hit at all, 211 areas had exactly one hit and so on. Like the previous computer failure data example, we will use this to illustrate and compare statistical modelling methods. The data set is `bombhits` in the `R` package `ipsRdbs` and there is a help file which is accessed by issuing the command `?bombhits`. ◀

Example 1.4 (Weight Gain of Students)

This data set was collected to investigate if students (see Fig. 1.5) tend to gain weight during their first year in college/university. In order to test this, David Levitsky, a Professor of Nutrition in the Cornell University (USA), recruited students from two large sections of an introductory course in health care, see the article Levitsky et al. [11]. Although they were volunteers, they appeared to match the rest of the freshman class in terms of demographic variables such as sex and ethnicity. Sixty-eight students were weighed during the first week of the semester, then again 12 weeks later. The table below provides the first and

Fig. 1.5 College students. Source: https://www.freestock.com/free-photos/college-university-students-smiling-classroom-23344615 Image used under license from Freestock.com

last three rows of the data set in kilograms, which was converted from imperial measurement in pounds and ounces.

Student number	Initial weight (kg)	Final weight (kg)
1	77.6	76.2
2	49.9	50.3
3	60.8	61.7
⋮	⋮	⋮
66	52.2	54.0
67	75.7	77.1
68	59.4	59.4

This data set will be used to illustrate simple exploratory charts in `R` and also to demonstrate what is known as statistical hypothesis testing. The data set is `wgain` in the `R` package `ipsRdbs`, with an associated help file `?wgain`. ◀

Example 1.5 (Body Fat Percentage)

Knowledge of the fat content of the human body is physiologically and medically important. The fat content may influence susceptibility to disease, the outcome of disease, the effectiveness of drugs (especially anaesthetics) and the ability to withstand adverse conditions including exposure to cold and starvation. In practice, fat content is difficult to measure directly—one way is by measuring body density which requires subjects to be weighed underwater! For this reason, it is useful to try to relate simpler measures such as skin-fold thicknesses (which

Fig. 1.6 Womens 5000 m start at the 2012 Olympics by Nick Webb. Source: Wikipedia. License: CC BY 2.0

are readily measured using calipers) to body fat content and then use these to estimate the body fat content.

Dr R. Telford, working for the Australian Institute of Sport (AIS), collected skin-fold (the sum of four skin-fold measurements) and percent body fat measurements on 102 elite athletes training at the AIS (see Fig. 1.6). Obtained from Prof Alan H. Welsh (author of Welsh [21]), the data set `bodyfat` is made available from the `R` package `ipsRdbs` and the `R` command `?bodyfat` provides further information and code for exploring and modelling the data, which has been perfprmed in Chap. 17. ◀

Athlete	Skin-fold	Body-fat (%)
1	44.5	8.47
2	41.8	7.68
3	33.7	6.16
⋮	⋮	⋮
100	47.6	8.51
101	60.4	11.50
102	34.9	6.26

Example 1.6 (Wealth of Billionaires)

Fortune magazine publishes a list of the world's billionaires each year. The 1992 list includes 225 individuals from five regions: **A**sia, **E**urope, **M**iddle East,

Fig. 1.7 Stack of 100 dollar bills. Source: Wikimedia Commons. License: CC BY-SA 3.0

United States, and **O**ther. For these 225 individuals we also have their wealth (in billions of dollars, see Fig. 1.7) and age (in years). The first and last two rows of the data set are given in the table below.

Wealth	Age	Region
37.0	50	M
24.0	88	U
⋮	⋮	⋮
1	9	M
1	59	E

This example will investigate differences in wealth of billionaires due to age and region using many exploratory graphical tools and statistical methods. The data set is `bill` in the `R` package `ipsRdbs` and a help file is obtained by issuing the command `?bill`. ◀

1.3 Basic Statistics

Having motivated to study statistics and introduced the data sets, our mission in this section is to learn some basic summary statistics and graphical tools through the use of the `R` software package. This section also aims to explore the summary statistics using basic mathematical tools such as the summation symbol $\sum$ and minimisation methods, which will be used repeatedly in the later chapters.

Suitable summaries of data are used to describe the data sets. The needs and motivation behind data collection often dictate what particular statistical summaries to report in the data description. Usually the non-numeric categorical variables are summarised by frequency tables. For example, we may report the number of billionaires in each of the five regions. For numeric variables we would like to know the centre of the data, i.e., measures of location or central tendency, and the spread or variability. The two sections below introduce these measures.

1.3.1 Measures of Location

This section defines three most common measures of location: mean, median and mode. It also justifies appropriateness of their use as a representative value for the data through theoretical arguments. The section ends with a discussion to decide the most suitable measure in practical applications.

1.3.1.1 Mean, Median and Mode

Suppose that the we have the data $x_1, x_2, \ldots, x_n$ for which we are seeking a representative value, which will be a function of the data. The sample mean denoted by $\bar{x}$ and defined by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

is a candidate for that representative value. Two other popular measures are the sample median and sample mode which we define below.

The sample median is the middle value in the ordered list of values $x_1, x_2, \ldots, x_n$. Consider the AM service time data in Example 1.1 where the values are: 38, 100, 64, 43, 63, 59, 107, 52, 86, 77. Obviously, we first write these values in order:

$$38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107.$$

There does not exist a unique middle value. But it is clear that the middle value must lie between 63 and 64. Hence, median is defined to be any value between 63 and 64. For the sake of definiteness, we may chose the mid-point 63.5 as the median.

In general, how do we find the middle value of the numbers $x_1, x_2, \ldots, x_n$? Mimicking the above example, we first write these values in order:

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)},$$

where $x_{(1)}$ denotes the minimum and $x_{(n)}$ denotes the maximum of the n data values $x_1, x_2, \ldots, x_n$. Note that we simply do not write $x_1 \le x_2 \le \cdots \le x_n$ since that would be wrong when data are not arranged in increasing order. Hence the new notation $x_{(i)}, i = 1, \ldots, n$ has been introduced here. For example, if 10 randomly observed values are: 9, 1, 5, 6, 8, 2, 10, 3, 4, and 7, then $x_1 = 9$ but $x_{(1)} = 1$.

If n is odd then $x_{(\frac{n+1}{2})}$ is the median value. For example, if $n = 11$ then the 6th value in the ordering of 11 sample values is the sample median. If n is even then sample median is defined to be any value in the interval $\left(x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}\right)$. For convenience, we often take the mid-point of the interval as the sample median. Thus, if $n = 10$ then the sample median is defined to be the mean of the 5th and 6th ordered values out of these $n = 10$ sample values. Thus, for the 10 observed values 9, 1, 5, 6, 8, 2, 10, 3, 4, and 7, the sample median is 5.5.

To recap, the sample median is defined as the observation ranked $\frac{1}{2}(n + 1)$ in the ordered list if n is odd. If n is even, the median is any value between $\frac{n}{2}$th and $(\frac{n}{2} + 1)$th in the ordered list. For example, for the AM service times, $n = 10$ and $38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107$. So the median is any value between 63 and 64. For convenience, we often take the mean of these. So the median is 63.5 seconds. Note that we use the unit of the observations when reporting any measure of location.

The third measure of location is the sample mode which is the most frequent value in the sample. In the London bomb hits Example 1.3, 0 is the mode of the number of bomb hits. If all sample values are distinct, as in the AM service time data example, then there is no unique mode. In such cases, especially when n is large, sample data may be presented in groups and the modal class may be found leading to an approximation for the mode. We, however, do not discuss frequency data in any further detail.

1.3.1.2 Which of the Three Measures to Use?

Which one of the three candidate representative values, sample mean, median, and mode shall we choose in a given situation? This question can be answered by considering a possibly fictitious imaginary idea of loss incurred in choosing a particular value, i.e., either the sample mean or median, as the representative value for all the observations. (Here we are thinking that we are guessing all the sample values by the representative value and there will be a loss, i.e. a penalty to be paid for incorrect guessing.)

Suppose a particular number a, e.g. the sample mean, is chosen to be the value representing the numbers $x_1, x_2, \ldots, x_n$. The loss we may incur in choosing a to represent any x_i could be a function of the error $x_i - a$. Hence the total loss is $\sum_{i=1}^{n}(x_i - a)$. But note that the total loss is not a going to be a good measure since some of the individual losses will be negative and some will be positive resulting in a small value or even a negative value of total loss. Hence, we often assume the squared-error loss, $(x_i - a)^2$ or the absolute error loss $|x_i - a|$ for representing the observation x_i so that the errors in ether direction (positive or negative) attracts similar amount of losses. Then we may choose the a that minimises the total error given by $\sum_{i=1}^{n}(x_i - a)^2$ in the case of squared-error loss. In case we assume absolute error loss we will have to find the a that minimises the sum of the absolute losses, $\sum_{i=1}^{n} |x_i - a|$.

It turns out that the sample mean is the a that minimises $\sum_{i=1}^{n}(x_i - a)^2$ and sample median is the a that minimises $\sum_{i=1}^{n} |x_i - a|$. The sample mode minimises a third type of loss obtained be considering the 0–1 loss function. In 0–1 loss, the loss is defined to be zero if $x_i = a$ and 1 if $x_i \neq a$ for $i = 1, \ldots, n$. That is, the loss is zero if a is the correct guess for x_i and 1 if a is an incorrect guess. It is now intuitively clear that the sample mode will minimise the total of the 0–1 loss function since if a =sample mode then the loss is going to be 0 for most of the observations, $x_1, \ldots, x_n$, resulting in the smallest value for total of the 0–1 loss.

Here now prove that *the sample mean minimises the sum of squares of the errors*, denoted by:

$$\text{SSE} = \sum_{i=1}^{n}(x_i - a)^2.$$

To establish this we can use the derivative method in Calculus, see the exercises. Here is an important alternative derivative free proof which will be used in other similar circumstances in later chapters.

Proof Here the trick is to subtract and then add the sample mean $\bar{x}$ inside the square $(x_i - a)^2$. Then the task is to simplify after expanding the square as follows:

$$\begin{aligned}
\textstyle\sum_{i=1}^{n}(x_i - a)^2 &= \textstyle\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \quad \text{[subtract and add } \overline{x}\text{]} \\
&= \textstyle\sum_{i=1}^{n}\left\{(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2\right\} \\
&= \textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - a)\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2,
\end{aligned}$$

since $\sum_{i=1}^{n}(x_i - \bar{x}) = n\overline{x} - n\overline{x} = 0$. Now note that the first term is free of a; the second term is non-negative for any value of a. Hence the minimum occurs when the second term is zero, i.e. when $a = \bar{x}$. This completes the proof. □

The trick of adding and subtracting the mean, expanding the square and then showing that the cross-product term is zero will be used several times in this book in the later chapters. Hence it is important to learn this proof. The result is a very important in statistics, and this will be used several times in this book. Note that, SSE= $\sum_{i=1}^{n}(x_1 - a)^2$ is the sum of squares of the deviations of $x_1, x_2, \ldots, x_n$ from any number a. The established identity states that:

The sum of (or mean) squares of the deviations of $x_1, x_2, \ldots, x_n$ from any number a is minimised when a is the sample mean of $x_1, x_2, \ldots, x_n$.

In the proof we also noted that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. This is stated as:

The sum of the deviations of a set of numbers from their mean is zero. (1.1)

The sum of the deviations of a set of numbers from their mean is zero.

We now prove that *the sample median minimises the sum of the absolute deviations*, defined by:

$$\mathrm{SAD} = \sum_{i=1}^{n} |x_i - a|.$$

Here the derivative approach, stated in the exercises to prove the previous result for the sample mean, does not work since the derivative does not exist for the absolute function. Instead we use the following argument.

Proof First, order the observations (see Fig. 1.8):

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

Now note that:

$$\begin{aligned}\mathrm{SAD} &= \textstyle\sum_{i=1}^{n} |x_i - a| \\ &= \textstyle\sum_{i=1}^{n} |x_{(i)} - a| \\ &= |x_{(1)} - a| + |x_{(n)} - a| + |x_{(2)} - a| + |x_{(n-1)} - a| + \cdots\end{aligned}$$

Now see Fig. 1.9 for a visualisation of the following argument. From the top line in the figure it is clear that $S_1(a) = |x_{(1)} - a| + |x_{(n)} - a|$ is minimised when a is such that $x_{(1)} \leq a \leq x_{(n)}$, in which case $S_1(a)$ takes the value $|x_{(1)} - x_{(n)}|$. Otherwise, suppose a lies outside of the interval $\left(x_{(1)}, x_{(n)}\right)$. For example, if $a < x_{(1)}$ then

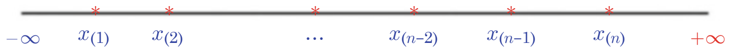

Fig. 1.8 Illustration of ordered observations

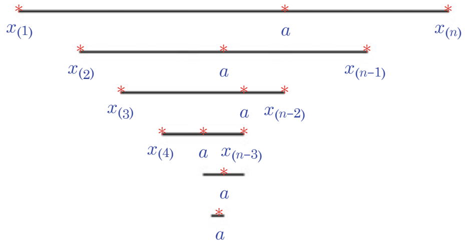

Fig. 1.9 Visualisation of the proof that the sample median minimises the sum of the absolute deviations

$S_1(a)$ will take the value $|x_{(1)} - x_{(n)}|$ plus $2 \times |x_{(1)} - a|$. Thus to minimise $S_1(a)$ we must put a somewhere in the interval $\left(x_{(1)}, x_{(n)}\right)$.

Continuing this argument we conclude that, $|x_{(2)} - a| + |x_{(n-1)} - a|$ is minimised when a is such that $x_{(2)} \leq a \leq x_{(n-1)}$. Finally, when n is odd, the last term $|x_{(\frac{n+1}{2})} - a|$ is minimised when $a = x_{(\frac{n+1}{2})}$ or the middle value in the ordered list. In this case a has been defined as the sample median above. If, however, n is even, the last pair of terms will be $|x_{(\frac{n}{2})} - a| + |x_{(\frac{n}{2}+1)} - a|$. This will be minimised when a is any value between $x_{(\frac{n}{2})}$ and $x_{(\frac{n}{2}+1)}$, which has been defined as the sample median in case n is even. Hence this completes the proof. □

This establishes the fact that:

> the sum (or mean) of the absolute deviations of $x_1, \ldots, x_n$ from any number a is minimised when a is the sample median of $x_1, \ldots, x_n$.

There is the concept of third type of loss, called a 0-1 loss, when we are searching for a measure of central tendency. In this case, it is intuitive that the best guess a will be the mode of the data, which is the most frequent value.

Which of the Three (Mean, Median and Mode) Should We Prefer? Obviously, the answer will depend on the type of loss we may assume for the particular problem. The decision may also be guided by the fact the sample mean gets more affected by extreme observations while the sample median does not. For example for the AM service times, suppose the next observation is 190. The median will be 64 instead of 63.5 but the mean will shoot up to 79.9.

1.3.2 Measures of Spread

The measures of central tendency defined in the previous section does not convey anything regarding the variability or spread of the data. Often, it is of interest to know how tightly packed the data are around the chosen centre, one of sample mean, median or mode. This section discusses three measures of spread or variability.

A quick measure of the spread is the *range*, which is defined as the difference between the maximum and minimum observations. For example, for the AM service times in the Fast Food Example 1.1 the range is 69 $(= 107 - 38)$ seconds. The range, however, is not a very useful measure of spread, as it is extremely sensitive to the values of the two extreme observations. Furthermore, it gives little information about the distribution of the observations between the two extremes.

A better measure of spread is given by the sample standard deviation, denoted by s, which the square-root of the *sample variance*, s^2, defined by

$$\text{Var}(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

The sample variance is defined with the divisor $n - 1$ since there are some advantages which will be discussed in Sect. 9.4.1. The divisor $n-1$ is the default in R. for the command **`var`**.

The population variance is defined with the divisor n instead of $n-1$ in the above.

Although s^2 above is defined as the mean of sum of squares of the deviations from the mean, we do not normally calculate it using that formula. Instead, we use the following fundamental identity:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\\ &= \sum_{i=1}^{n}x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2\\ &= \sum_{i=1}^{n}x_i^2 - n\bar{x}^2.\end{aligned}$$

Hence we prefer to calculate variance by the formula:

$$\text{Var}(x) = s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)$$

and the *standard deviation* is taken as the square-root of the variance. For example, the standard deviation of the AM service times is 23.2 seconds. Note that standard deviation has the same unit as the observations.

A third measure of spread is what is known as the inter-quartile range (IQR). The IQR is the difference between the third, Q_3 and first, Q_1 quartiles, which are respectively the observations ranked $\frac{1}{4}(3n + 1)$ and $\frac{1}{4}(n + 3)$ in the ordered list, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Note that the sample median is the second quartile, Q_2. When n is even, definitions of Q_3 and Q_1 are similar to that of the median, Q_2. The lower and upper quartiles, together with the median, divide the observations up into four sets of equal size. For the AM service times

$$38 < 43 < 52 < 59 < 63 < 64 < 77 < 86 < 100 < 107$$

Q_1 lies between 52 and 59, while Q_3 lies between 77 and 86. Some linear interpolation methods are used to find approximate values in R. We, however, do not discuss this any further.

Usually the three measures: range, sd and IQR are not used interchangeably. The range is often used in data description, the most popular measure, standard

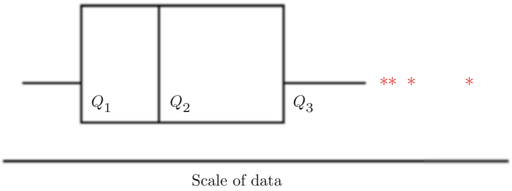

Fig. 1.10 A sketch of a boxplot diagram

deviation, is used as a measure of variability or concentration around the sample mean and the IQR is most often used in graphical summaries of the data such as the boxplot which is described in the next section.

1.3.3 Boxplot

A boxplot of sample data, e.g. computer failure data, plots the three quartiles and also provides valuable information regarding the shape and concentration of the data. From a boxplot, we can immediately gain information concerning the centre, spread, and extremes of the distribution of the observations (Fig. 1.10).

Constructing a boxplot involves the following steps:

1. Draw a vertical (or horizontal) axis representing the interval scale on which the observations are made.
2. Calculate the median, and upper and lower quartiles (Q_1, Q_3) as described above. Calculate the inter-quartile range (or 'midspread') $H = Q_3 - Q_1$.
3. Draw a rectangular box alongside the axis, the ends of which are positioned at Q_1 and Q_3. Hence, the box covers the 'middle half' of the observations). Q_1 and Q_3 are referred to as the 'hinges'.
4. Divide the box into two by drawing a line across it at the median.
5. The whiskers are lines which extend from the hinges as far as the most extreme observation which lies within a distance $1.5 \times H$, of the hinges.
6. Any observations beyond the ends of the whiskers (further than $1.5 \times H$ from the hinges) are suspected outliers and are each marked on the plot as individual points at the appropriate values. (Sometimes a different style of marking is used for any outliers which are at a distance greater than H from the end of the whiskers).

Figure 1.11 shows a boxplot of the number of weekly computer failure data introduced in Example 1.2. The two quartiles Q_1 and Q_3 are drawn as vertical lines at the two edges of the box and the median is the bold vertical line drawn through the middle of the box. The whiskers are the horizontal lines drawn at the two edges of the box. There are three extreme observations plotted as individual points beyond the whisker at the right of the plot. Comparing the lengths of the two whiskers, we see that the data shows an un-even distribution where there is a larger spread of values in the right hand side of the median, or the right tail. Such a distribution of

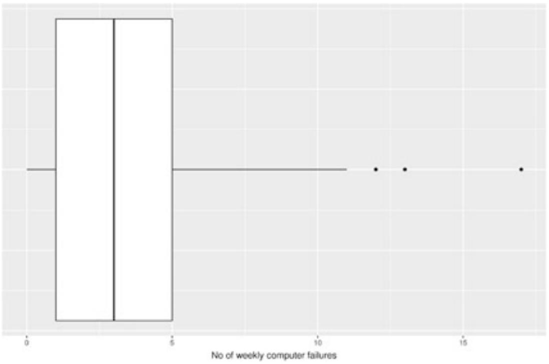

Fig. 1.11 A boxplot of computer failure data

the data is often called to be positively (or right) skewed. We will learn how to draw such a plot using `R` in Chap. 2.

Exploration of statistical data uses many other plots such as the stem and leaf plot, histogram, barplot and pie chart. However, this textbook does not provide discussion of such plots for brevity. Instead, the interested reader is referred to school level elementary statistics textbooks for detailed discussions on such topics.

1.3.4 Summary

This section has introduced three measures of location: mean, median and mode, each of which is optimal under a different consideration. We have also introduced three measures of variability range, sd and the IQR, each of which has the same unit as the original data.

1.4 Exercises

1.1 (Addition with the Summation Symbol $\sum$)

- Assume $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are real numbers not all zero. Also, a, b, k are real numbers and $n > 0$ is an integer.
- We write $\displaystyle\sum_{i=1}^{n} x_i$ to denote the sum $x_1+x_2+\cdots+x_n$. We should always include the limits and the dummy (e.g. i), i.e., $\displaystyle\sum_{i=1}^{n} x_i$, and we do not encourage the notation $\sum x$ since it does not make it clear what numbers are being added up. Also note that $\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$, i.e, the letter i or j we write for the dummy does not matter.

1. Prove that $\sum_{i=1}^{n} k\,x_i = k \sum_{i=1}^{n} x_i$.
2. Prove that $\sum_{i=1}^{n} (k + x_i) = n\,k + \sum_{i=1}^{n} x_i$.
3. Prove that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.
4. Prove that $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.
5. Suppose that the values of $x_1, \ldots, x_n$ are known and we want to minimise the sum $\sum_{i=1}^{n} (x_i - a)^2$ with respect to the variable a. Prove that $\sum_{i=1}^{n} (x_i - a)^2$ is minimised when $a = \bar{x}$ by using the derivative method described below.

   To optimise $f(a)$, we first solve the equation $f'(a) = 0$. We then see if $f''(a)$, evaluated at the solution, is positive or negative. The function $f(a)$ attains a local **minimum** at the solution if the sign is positive. The function $f(a)$ attains a local **maximum** at the solution if the sign is negative. There is neither a minima nor a maxima if the second derivative is zero at the solution. Such a point is called a *point of inflection.*

1.2 (Mean-Variance)

1. Suppose we have the data: $x_1 = 1,\ x_2 = 2, \ldots, x_n = n$. Find the mean and the variance. For variance use the divisor n instead of $n - 1$.
2. Suppose $y_i = ax_i + b$ for $i = 1, \ldots, n$ where a and b are real numbers. Show that:
   (a) $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i = a\bar{x} + b$ and
   (b) $\text{Var}(y) = a^2 \text{Var}(x)$ where $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
   and for variance it is possible to use either the divisor n, i.e. $\text{Var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ or $n - 1$, i.e. $\text{Var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$. The divisor does not matter as the results hold regardless. **Hint:** For the second part, start with the left hand side, $\text{Var}(y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$ and substitute y_i and $\bar{y}$ in terms of x_i and $\bar{x}$.
3. Suppose $a \le x_i \le b$ for $i = 1, \ldots, n$. Show that $a \le \bar{x} \le b$.

1.3 (Variance Inequality)

1. Prove that for any set of numbers $x_1,\ x_2, \ldots, x_n$,

$$\left(x_1^2 + x_2^2 + \cdots + x_n^2\right) \ge \frac{(x_1 + x_2 + \ldots x_n)^2}{n},$$

   i.e. sum of squares of n numbers is greater than equal to the square of the sum divided by n. **Hint:** You may start by assuming $\sum_{i=1}^{n} (x_i - \overline{x})^2 \ge 0$ and then expand the square within the summation symbol.

1.4 (Additional Data)

1. Assume that $x_1, \ldots, x_n, x_{n+1}$ are given real numbers. Prove that:
   (a) $\bar{x}_{n+1} = \frac{x_{n+1}+n\bar{x}_n}{n+1}$
   (b) $ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$.
2. Assume that $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ are real numbers not all zero, where m and n are positive integers. Let $z_1, \ldots, z_{n+m}$ denote the combined $m+n$ observations, i.e. $\mathbf{z} = (x_1, \ldots, x_m, y_1, \ldots, y_n)$ without loss of generality.

   Let $\bar{x}$, s_x^2, $\bar{y}$, s_y^2, $\bar{z}$, s_z^2 denote the sample mean and variance pair of the x's, y's and z's respectively.
   (a) Prove that $\bar{z}$ is given by:

$$\bar{z} = \frac{m\,\bar{x} + n\,\bar{y}}{m+n}.$$

   (b) Prove that the sample variance of the z values is given by:

$$s_z^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-1} + \frac{m(\bar{x}-\bar{z})^2 + n(\bar{y}-\bar{z})^2}{m+n-1}.$$

- These two formulae allow us to calculate the mean and variance of the combined data easily.

1.5 (Two Variables and the Cauchy-Schwarz Inequality) Suppose that $(x_1, y_1), (x_2, y_2). \ldots, (x_n, y_n)$ are given pairs of numbers.

1. Prove that

$$\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})x_i = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}.$$

2. Prove the Cauchy-Schwarz Inequality.

$$\left(x_1^2 + x_2^2 + \cdots + x_n^2\right)\left(y_1^2 + y_2^2 + \cdots + y_n^2\right) \geq (x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2.$$

   Hint: You can either try the induction method or use the fact that for any set of numbers $a_1, \ldots a_n$ and $b_1, \ldots b_n$.:

$$\sum_{i=1}^{n}(a_i - b_i)^2 \geq 0, \quad \text{and then substitute } a_i = \frac{x_i}{\sqrt{\sum_{i=1}^{n} x_i^2}},\ b_i = \frac{y_i}{\sqrt{\sum_{i=1}^{n} y_i^2}}.$$