Chapter 1
Introduction and Motivation

Statistics is the science and art of making decisions based on quantitative evidence. This introductory chapter motivates the study of statistics by describing where and how it is used in all endeavors. It gives examples of applications, a little history of the subject, and a brief overview of the structure and content of the remaining chapters.

Almost all fields of study (including but not limited to physical science, social science, business, and economics) collect and interpret numerical data. Statistical techniques are the standard ways of summarizing and presenting the data, of turning data from an accumulation of numbers into usable information. Not all numbers are the same. No group of people are all the same height, no group has an identical income, not all cars get the same gas mileage, not all manufactured parts are absolutely identical. How much do they differ? Variability is the key concept that statistics offers. It is possible to measure how much things are not alike. We use standard deviation, variance, range, interquartile range, and MAD (median absolute deviation from the median) as measures of not-the-sameness. When we compare groups we compare their variability as well as their range.

Statistics uses many mathematical tools. The primary tools—algebra, calculus, matrix algebra, analytic geometry—are reviewed in Appendix I. Statistics is not purely mathematics. Mathematics problems are usually well specified and have a single correct answer on which all can agree. Data interpretation problems calling for statistics are not yet well specified. Part of the data analyst's task is to specify the problem clearly enough that a mathematical tool may be used. Different answers to the same initial decision problem may be valid because a statistical analysis requires assumptions about the data and its manner of collection, and analysts can reasonably disagree about the plausibility of such assumptions.

Statistics uses many computational tools. In this book, we use R (R Core Team, 2015) as our primary tool for statistical analysis. R is an exceptionally well-developed tool for statistical research and analysis, that is for exploring and

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_1

designing new techniques of analysis, as well as for analysis. We discuss installation and use of R in Appendix A.

We make liberal use of graphs in our presentations. Data analysts are responsible for the *display* of data with graphs and tables that summarize and represent the data and the analysis. Graphs are often the output of data analysis that provide the best means of communication between the data analyst and the client. We study a variety of display techniques.

While producing this book, we designed many innovative graphical displays of data and analyses. We introduce our displays in Section 1.3.4. We discuss the displays throughout the book in the context of their associated statistical techniques. These discussions are indexed under the term *graphical design*. In the appendix to Chapter 4, we summarize the large class of newly created graphs that are based on Cartesian products.

The R code for all the graphs and tables in this book is included in the **HH** package for R (Heiberger, 2015). See Appendix B for a summary of the **HH** package. We consider the **HH** package to be an integral part of the book.

Statistics is an art. Skilled use of the mathematical tools is necessary but not sufficient. The data analyst must also know the subject area under study (or must work closely with a specialist in the subject area) to ensure an appropriate choice of statistical techniques for solving a problem. Experience, good judgment, and considerable creativity on the part of the statistical analyst are frequently needed.

Statistics is "the science of doing science" and is perhaps the only discipline that interfaces with all other sciences. Most statisticians have training or considerable knowledge in one or more areas other than statistics. The statistical analyst needs to communicate successfully both orally and in writing with the client for the analysis.

Statistics uses many communications skills, both written and oral. Results must be presented to the client and to the client's management. We discuss some of the mechanics of writing programs and technical reports in Appendices K, L, M, N, and O.

A common statistical problem is to discover the characteristics of an unobservable population by examining the corresponding characteristics of a sample *randomly* selected from the population and then (inductively) inferring the population characteristics (parameters) from the corresponding sample characteristics (statistics). The task of selecting a random sample is not trivial. The discipline of statistics has developed a vast array of techniques for inferring from samples to populations, and for using probabilities to quantify the quality of such inferences.

Most statistical problems involve simultaneous consideration of several related measurements. Part of the statistician's task is to determine the interdependence among such measures, and then to account for it in the analysis.

The word "statistics" derives from the political science collections of numerical data describing demographics, business, politics that are useful for

management of the "state". The development of statistics as a scientific discipline dates from the end of the 19th century with the design and analysis of agricultural experiments aimed at finding the best combination of fertilization, irrigation, and variety to maximize crop yield. Early in the 20th century, these ideas began to take hold in industry, with experiments designed to maximize output or minimize cost. Techniques for statistical analysis are developed in response to the needs of specific subject areas. Most of the techniques developed in one subject field can be applied unchanged to other subjects.

1.1 Statistics in Context

We write as if the statistician and the client are two separate people. In reality they are two separate roles and the same person often plays both roles. The client has a problem associated with the collection and interpretation of numerical data. The statistician is the expert in designing the data collection procedures and in calculating and displaying the results of statistical analyses.

The statistician's contribution to a research project typically includes the following steps:

1. Help the client phrase the question(s) to be answered in a manner that leads to sensible data collection and that is amenable to statistical analysis.
2. Design the experiment, survey, or other plan to approach the problem.
3. Gather the data.
4. Analyze the data.
5. Communicate the results.

In most statistics courses, including the one for which this book is designed, much of the time is spent learning how to perform step 4, the science of statistics. However, step 2, the art of statistics, is very important. If step 2 is poorly executed, the end results in step 5 will be misleading, disappointing, or useless. On the other hand, if step 4 is performed poorly following an excellent plan from step 2 and a correct execution of step 3, a reanalysis of the data (a new step 4) can "save the day".

Today (2015) there are more than 18,000 statisticians practicing in the United States. Most fields in the biological, physical, and social sciences require training in statistics as educational background. Over 100 U.S. universities offer graduate degrees in statistics. Most firms of any size and most government agencies employ statisticians to assist in decision making. The profession of *statistician* is highly placed in the *Jobs Rated Almanac* Krantz (1999). A shortage of qualified statisticians to fill open positions is expected to persist for some time American Statistical Association (2015).

1.2 Examples of Uses of Statistics

Below are a few examples of the countless situations and problems for which statistics plays an important part in the solution.

1.2.1 Investigation of Salary Discrimination

When a group of workers believes that their employer is illegally discriminating against the group, legal remedies are often available. Usually such groups are minorities consisting of a racial, ethnic, gender, or age group. The discrimination may deal with salary, benefits, an aspect of working conditions, mandatory retirement, etc. The statistical evidence is often crucial to the development of the legal case.

To illustrate the statistician's approach, we consider the case of claimed salary discrimination against female employees. The legal team and statistician begin by developing a defensible list of criteria that the defendant may legally use to determine a worker's salary. Suppose such a list includes years of experience (`yrsexp`), years of education (`yrsed`), a measure of current job responsibility or complexity (`respon`), and a measure of the worker's current productivity (`product`). The statistician then obtains from a sample of employees, possibly following a subpoena by the legal team, data on these four criteria and a fifth criterion that distinguishes between male and female employees (`gender`) . Using regression analysis techniques we introduce in Chapter 9, the statistician considers two statistical models, one that explains `salary` as a function of the four stipulated permissible criteria, and another that explains `salary` as a function of these four criteria plus `gender`. If the model containing the predictor `gender` predicts salary appreciably better than does the model excluding `gender` and if, according to the model with `gender` included, females receive significantly less salary than males, then this may be regarded as statistical evidence of discrimination against females. Tables and graphs based on techniques discussed in Chapters 15, 17, and 4 (and other chapters) are often used in legal proceedings.

In the previous section it is pointed out that two statisticians can provide different analyses because of different assumptions made at the outset. In this discrimination context, the two legal teams may disagree over the completeness or relevance of the list of permissible salary determinants. For example, the defense team may claim that females are "less ambitious" than males, or that women who take maternity or child care leaves have less continuous or current experience than men. If the court accepts such arguments, this will undermine the plaintiff statistician's finding of the superiority of the model with the extra predictor.

1.2.2 Measuring Body Fat

In Chapters 8, 9, and 13 we discuss an experiment designed to develop a way to estimate the percentage of fat in a human body based only on body measurements that can be made with a simple tape measure. The motivation for this investigation is that measurement of body fat is difficult and expensive (it requires an underwater weighing technique), but tape measurements are easy and inexpensive to obtain. At the outset of this investigation, the client offered data consisting of 15 inexpensive measurements and the expensive body fat measurement on each of 252 males of various shapes and sizes. Our analysis in Chapter 9 demonstrates that essentially all of the body fat information in the 15 other measurements can be captured by just three of these other measurements. We develop a regression model of body fat as a function of these three measurements, and then we examine how closely these three inexpensive measurements alone can estimate body fat.

1.2.3 Minimizing Film Thickness

In Section 13.3.1 we discuss an experiment that seeks to find combinations of `temperature` and `pressure` that minimize the thickness of a film deposited on a substrate. Each of these factors can affect thickness, and the complication here is the possibility that the optimum amount of one of these factors may well depend on the chosen amount of another factor. Modeling such *interaction* between factors is key to a proper analysis. The statistician is also expected to advise on the extent of sensitivity of thickness to small changes in the optimum mix of factors.

1.2.4 Surveys

Political candidates and news organizations routinely sample potential voters for their opinions on candidates and issues. Results gleaned from samples selected by contemporary methods are often regarded as sufficiently reliable to influence candidate behavior or public policy.

The marketing departments of retail firms often sample potential customers to learn their opinions on product composition or packaging, and to help find the best matches between specialized products and locales for their sale.

Manufacturers sample production to determine if the proportion of output that is defective is excessive. If so, this may lead to the decision the output should be scrapped, or at least that the production process be inspected and corrected for problems.

All three of these examples share statistical features. The data are collected using techniques discussed in Section 3.11. The initial analysis is usually based on techniques of Chapter 5.

1.2.5 Bringing Pharmaceutical Products to Market

The successful launching of a new pharmaceutical drug is a huge undertaking in which statisticians are key members of the investigative team. After candidate drugs are found to be effective for alleviation of a condition, experiments must be run to check them for toxicity, safety, side effects, and interactions with other drugs. Once these tests are passed, statisticians help to determine the optimum quantity and spacing of dosages. Much of the testing is done on lab animals; only at the later stages are human subjects involved. The entire process is performed in a manner mandated by government regulatory agencies (such as the Food and Drug Administration (FDA) in the United States, The European Medicines Agency (EMA) in the European Union, or the Ministry of Health, Labour and Welfare (MHLW) in Japan). Techniques are based on material developed in all chapters of this book.

1.3 The Rest of the Book

1.3.1 Fundamentals

Chapters 2 through 5 discuss data, types of data analysis, and graphical display of data and of analyses.

Chapter 2 describes data acquisition and how to get the data ready for its analysis. We emphasize that an important early step in any data analysis is graphical display of the data.

Chapter 3 provides an overview of basic concepts—probability, distributions, estimation, testing, principles of inference, and sampling—that are background material for the remainder of the book. Several common distributions are discussed and illustrated here. Others appear in Appendix J. Two important fitting criteria—least squares and maximum likelihood—are introduced. Random sampling is a well-defined technique for collecting data on a subset of the population of interest. Random sampling provides a basis for making inferences that a haphazard collection of data cannot provide.

A variety of graphical displays are discussed and illustrated in Chapter 4. The graphs themselves are critically important analysis tools, and we show examples

where different display techniques help in the interpretation of the data. On occasion we display graphs that are intermediate steps leading to other graphs. For example, Figure 14.17 belongs in a final report, but Figure 14.15, which suggests the improved and expanded Figure 14.17, should not be shown to the client.

Chapter 5 introduces some of the elementary inference techniques that are used throughout the rest of the book. We focus on tests on data from one or two normal distributions. We show the algebra and graphics for finding the center and spread of the distributions. These algebraic and graphical techniques are used in all remaining chapters.

1.3.2 Linear Models

Chapters 6 through 13 build on the techniques developed in Chapter 5. The word "linear" means that the equations are all linear functions of the model parameters and that graphs of the analyses are all straight lines or planes.

In Chapter 6 we extend the t-tests of Chapter 5 to the comparison of the means of several (more than two) populations.

With $k > 2$ populations, there are only $k - 1$ independent comparisons possible, yet we often wish to make $\binom{k}{2}$ comparisons. In Chapter 7 we discuss the concept of *multiple comparisons*, the way to make valid inferences when there are more comparisons of interest than there are degrees of freedom. We introduce the fundamental concept of "contrasts", direct comparisons of linear combinations of the means of these populations, and show several potentially sensible ways to choose $k - 1$ independent contrasts. We introduce the *MMC* plot, the mean–mean plot for displaying arbitrary multiple comparisons.

Chapters 8 through 11 cover regression analysis, the process of modeling a continuous response variable as a linear function of one or more predictor variables.

In Chapter 8 we plot a continuous response variable against a single continuous predictor variable and develop the least-squares procedure for fitting a straight line to the points in the plot. We cast the algebra of least squares in matrix notation (relevant matrix material is in Appendix I) and apply it to more than one predictor variable. We introduce the statistical assumptions of a normally distributed error term and show how that leads to estimation and testing procedures similar to those introduced in Chapter 5.

Chapter 9 builds on Chapter 8 by allowing for more than one predictor for a response variable and introducing additional structure, such as interactions, among the predictor variables. We show techniques for studying the relationships of the predictors to each other as well as to the response.

Chapter 10 shows how dummy variables are used to incorporate categorical predictors into multiple regression models. We begin to use dummy variables to encode the contrasts introduced in Chapter 6, and we continue using dummy variables and contrasts in Chapters 12, 13, and 14. We show how the use of continuous (concomitant) variables (also known as covariates) can enhance the modeling of designed experiments.

Chapter 11 evaluates the models, introduces diagnostic techniques for checking assumptions and detecting outliers, and uses tools such as transformation of the variables to respond to the problems detected.

In Chapter 12 we extend the analysis of one-way classifications of continuous data to several types of two-way classifications. We cast the analysis of variance into the regression framework with dummy variables that code the classification factors with sets of contrasts.

In Chapters 13 and 14 we consider the principles of experimental design and their application to more complex classifications of continuous data. We discuss the analysis of data resulting from designed experiments.

1.3.3 Other Techniques

The analysis of tabular categorical data is considered in Chapter 15. We discuss contingency tables, tables in which frequencies are classified by two or more factors. For 2×2 tables or sets of 2×2 tables we use odds ratios or the Mantel–Haenszel test. For larger tables we use χ^2 analysis. We discuss several situations in which contingency tables arise, including sample surveys and case–control studies.

In Chapter 16 we briefly survey nonparametric testing methods that don't require the assumption of an underlying normal distribution.

Chapter 17 is concerned with logistic regression, the statistical modeling of a response variable which is either dichotomous or which represents a probability. We place logistic regression in the setting of generalized linear models (although we do not fully discuss generalized linear models in this volume). We extend the graphical and algebraic analysis of linear regression to this case.

We conclude in Chapter 18 with an introduction to ARIMA modeling of time series. Time series analysis makes the explicit assumption that the observations are *not* independent and models the structure of the dependence.

1.3.4 New Graphical Display Techniques

This book presents many new graphical display techniques for statistical analysis. Most of our new displays are based on defining the panels of a multipanel graphical display by a Cartesian product of sets of variables, of transformations of a variable, of functions of a fit, of models for a fit, of numbers of parameters, or of levels of a factor. The appendix to Chapter 4 summarizes how we use the Cartesian products to design these new displays and gives a reference to an example in the book for each. The displays, introduced throughout this book's 18 chapters, cover a wide variety of statistical methods. The construction and interpretation of each display are provided in the chapter where it appears.

We produced these displays with the functions that are included in the **HH** package available at CRAN (Heiberger, 2015) and CSAN (Heiberger, 2009). We use R because it is especially strong for designing and programming statistical graphics. We encourage readers and software developers to write and publish functions and macros for these displays in other software systems that have a similarly rich graphics environment.

1.3.5 Appendices on Software

Appendix A discusses the installation and use of R. Some of its material was in the First Edition Appendix B.

Appendix B discusses the **HH** package. The scripts for all examples in both the First and Second Editions of the book are included in the **HH** package. The Appendix shows how to use the scripts to duplicate the figures and tables in the book. Some of its materials were in the First Edition Appendix B.

Appendix C "**Rcmdr**" is new. It discusses and illustrates menu-driven access to the functions and graphics in the book. It is based on my R package **RmcdrPlugin.HH**, an add-in for the R package **Rcmdr** that provides the menu system.

Appendix D "RExcel" is new. It discusses the RExcel interface described in my book with Erich Neuwirth (Heiberger and Neuwirth, 2009) describing his RExcel software (Neuwirth, 2014). RExcel provides a seamless integration of R and Excel. RExcel both places R inside the Excel automatic recalculation model and makes the Rcmdr menu system available on the Excel menu bar.

Appendix E "**Shiny**" is new. It discusses and illustrates web-based access to R functions using the **shiny** package written by R-Studio and distributed on CRAN. **shiny** provides an R language interface for writing interactive web pages.

Appendix F "R Packages" gives a very brief discussion of software design. It includes references to the R documentation.

Appendix G "Computational Precision and Floating Point Arithmetic" is new. Computers use *floating point* arithmetic. The floating point system is not identical to the real-number system that we (teachers and students) know well, having studied it from kindergarten onward. In this appendix we show several examples to illustrate and emphasize the distinction.

Appendix H "Other Statistical Software" is new. It tells how to use the datasets for this book with software other than R.

1.3.6 Appendices on Mathematics and Probability

Appendix I "Mathematics Preliminaries" has been expanded from First Edition Appendix F with many more graphs and tables.

Appendix J "Probability Distributions" has been expanded from First Edition Appendix D to include additional probability distributions. It now covers all probability distributions in the R **stats** package, and it now includes a density graph for each distribution.

1.3.7 Appendices on Statistical Analysis and Writing

Appendix K "Working Style" has been split off and expanded from First Edition Appendix E. It includes a discussion of the importance of a good R-aware text editor and defines what that means. It includes a discussion of our process in writing this book and my process in writing and maintaining the **HH** package.

Appendix L "Writing Style" has been split off and expanded from First Edition Appendix E. It discusses some of the basics of clear writing—including typography, presentation of graphs, and alignment in tables, and programming style.

Appendix M "Accessing R through a Powerful Editor—with Emacs and **ESS** as the Example" has been split off and expanded from First Edition Appendix E. A good editor is one of the most important programs on your computer. It is the direct contact with all the documents, including R scripts and R functions, that you write. A good editor will understand the syntax of your programming language (R specifically) and will simplify the running and testing of code. We write in the terminology of Emacs because it is our working environment. Most of what we illustrate applies to other high-quality editors.

Appendix N "LaTeX" has been split off and expanded from First Edition Appendix E. It provides basic information about LaTeX, the document preparation system in which we wrote this book.

Appendix O "Word Processors and Spreadsheets" has been split off and expanded from First Edition Appendix E. Unless there are specific add-ins that understand R, we do not recommend word processing software for working with R. We can recommend spreadsheet software for use as a small-scale database management system and as a way of organizing calculations. Unless you are working with RExcel (discussed in Appendix D) we do not recommend the use of spreadsheets for the actual statistical calculations.