1

Learning from Data, and Tools for the Task

Chapter Summary

We begin by illustrating the interplay between questions driven by scientific curiosity and the use of data in seeking the answers to such questions. Graphs provide a useful window through which meaning can be extracted from data. Numeric summary statistics and probability distributions provide a form of quantitative scaffolding for models of random as well as nonrandom variation. Simple regression models foreshadow the issues that arise in the more complex models considered later in the book. Frequentist and Bayesian approaches to statistical inference are touched upon, the latter primarily using the Bayes Factor as a summary statistic which moves beyond the limited perspective that p-values offer. Resampling methods, where the one available dataset is used to provide an empirical substitute for a theoretical distribution, are also introduced. Remaining topics are of a more general nature. Section 1.9 will discuss the use of RStudio and other such tools for organizing and managing work. Section 1.10 will include a discussion on the important perspective that replication studies provide, for experimental studies, on the interplay between statistical analysis and scientific practice. The checks provided by independent replication at another time and place are an indispensable complement to statistical analysis. Chapter 2 will extend the discussion of this chapter to consider a wider class of models, methods, and model diagnostics.

A Note on Terminology – Variables, Factors and More!

Much of data analysis is concerned with the statistical modeling of relationships or associations that can be gleaned from data, with a mathematical formula used to specify the model. There is an example at the beginning of Section 1.1.6.

The word *variable* will be used when data values are numeric. These include counts, as for example in `count` from the `DAAG::ACF1` data frame which has numbers of aberrant lesions in the lining of a rat's colon. The term *factor* will be used when values are on a categorical scale. Thus, in the data frame `DAAG::kiwishade`, `yield` is a variable with values such as 101.11, and `block` is a factor with levels `east`, `north`, and `west`. A factor may also represent values on an ordinal scale. Thus the factor `tint` in the data frame `DAAG::tinting` has ordered levels `no`, `lo`, and `hi`.

Continuous measurements can be further classified as having either an *interval* scale or a *ratio* scale. Variables defined on an interval scale can take positive or negative values, and differences in the data values are meaningful. Variables defined on a ratio scale are usually positive only, so quotients are more meaningful.

1.1 Questions, and Data That May Point to Answers

Accounts of observed phenomena become part of established science once we know the circumstances under which they will recur. This process is relatively straightforward when applied to the study of regular events, such as a solar eclipse or the ocean tide levels in the Bay of Fundy in eastern Canada. Mathematical models that are based on sound physical principles can provide very accurate predictions for such events. Not everything is so readily predictable. How effective is a particular vaccine in preventing COVID-19-associated hospitalizations? How fast will a wildfire spread through a region with known topography and vegetation under given wind, temperature, and moisture conditions? Data from a suitable experiment or series of experiments may be able to go at least part of the way towards providing an answer. Thus, results from prescribed burns in designated forest stands where all relevant variables have been measured can provide a starting point for assessing the rate of spread of surface fires.

Or it may be necessary to rely on whatever data are already available. How effective are airbags in reducing the risk of death in car accidents? Data on car accidents in the United States over the period 1997–2002 are available. While careful and critical analyses of these data can help answer the question, caveats apply when the interest is in effectiveness at a later time and in another country. There have been important advances in the subsequent two decades in airbag design, manufacture, and systems that control deployment.

In Canada, there is a tendency for car passengers to use seatbelts at a higher rate than in the United States, so that efficacy assessments based on the American data have to be tempered when applied to the Canadian experience. The decision on which of the available datasets is best designed to provide an answer, and the choice of model, have called for careful and critical assessment. The help pages `?DAAG::nassCDS` and `?gamclass::FARS` provide further commentary. There is a strong interplay between the questions that can reasonably be asked, and the data that are available or can be collected. Keep in mind, also, that different questions, asked of the same data, may demand different analyses.

1.1.1 A Sample Is a Window into the Wider Population

The population comprises all the data that might have been. The sample is the data that we have. Subjects for a sample to be surveyed should be selected randomly. In a clinical trial, it is important to allocate subjects randomly to different treatment groups.

Suppose, for example, that names on an electoral roll are numbered from 1 to 9384. The following uses the function `sample()` to obtain a random sample of 12 individuals:

```
## For the sequence below, precede with set.seed(3676)
sample(1:9384, 12, replace=FALSE)  # NB: `replace=FALSE` is the default
```

```
 [1] 2263 9264 4490 8441 1868 3073 5430   19 1305 2908 5947  915
```

The numbers are the numerical labels for the 12 individuals who are included in the sample. The task is then to find them! The option `replace=FALSE` gives a *without replacement* sample, that is, it ensures that no one is included more than once.

A more realistic example might be the selection of 1200 individuals, perhaps for purposes of conducting an opinion poll, from names numbered 1 to 19,384, on an electoral roll. Suitable code is:

```
chosen1200 <- sample(1:19384, 1200, replace=FALSE)
```

The following randomly assigns 10 plants (labeled from 1 to 10, inclusive) to one of two equal-sized groups, control and treatment:

```
## For the sequence below, precede with set.seed(366)
split(sample(seq(1:10)), rep(c("Control","Treatment"), 5))
```

```
$Control
[1]  5  7  1 10  4

$Treatment
[1] 8 6 3 2 9
```

```
# sample(1:10) gives a random re-arrangement (permutation) of 1, 2, ..., 10
```

This assigns plants 3, 5, 10, 2, and 7 to the control group. This mechanism avoids any unwitting preference for placing healthier-looking plants in the treatment group.

The simple independent random sampling scheme can be modified or extended in ways that take account of structure in the data, with random sampling remaining a part of the data-selection process.

Cluster Sampling

Cluster sampling is one of many probability-based variants on simple random sampling. See Barnett (2002). The function `sample()` can be used as before, but now the numbers from which a selection is made correspond to clusters. For example, households or localities may be selected, with multiple individuals from each. Standard inferential methods then require adaptation to account for the fact that it is the clusters that are independent, not the individuals within the clusters. Donner and Klar (2000) describe methods that are designed for use in health research.

**A Note on With-Replacement Samples*

For data that can be treated as a random sample from the population, one way to get an idea of the extent to which it may be affected by random variation is to take with-replacement random samples from the one available sample, and to do this

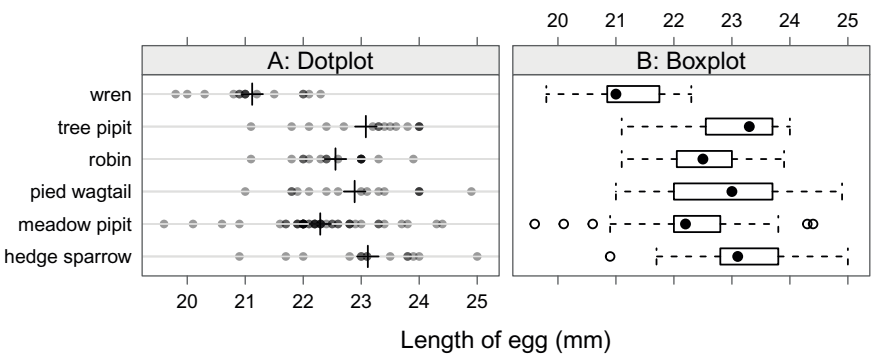

Figure 1.1 Dotplot (Panel A) and boxplot (Panel B) displays of cuckoo egg lengths. In Panel A, points that overlap have a more intense color. Means are shown as +. The boxes in Panel B take in the central 50 percent of the data, from 25 percent of the way through the data to 75 percent of the way through. The dot marks the median. Data are from Latter (1902).

repeatedly. The distribution that results can be an empirical substitute for the use of a theoretical distribution as a basis for inference.

We can randomly sample from the set $\{1, 2, \ldots, 10\}$, allowing repeats, thus:

```
sample(1:10, replace=TRUE)
```

```
 [1]  1  3  7  5  5 10  3  3  2  9
```

```
## sample(1:10, replace=FALSE) returns a random permutation of 1,2,...10
```

With-replacement sampling is the basis of *bootstrap* sampling. The effect is that of repeating each value an infinite number of times, and then taking a without-replacement sample. Subsections 1.8.3 and 1.8.4 will demonstrate the methodology.

1.1.2 Formulating the Scientific Question

Questions should be structured with a view both to the intended use of results, and to the limits of what the available data allow. Predictions of numbers in hospital from COVID-19 two weeks into the future do not demand the same level of scientific understanding or detailed data as needed to judge who among those infected are most likely to require hospitalization.

Example: A Question About Cuckoo Eggs

Cuculus canorus is one of several species of cuckoos that lay eggs in the nests of other birds. The eggs are then unwittingly adopted and hatched by the hosts. Latter (1902) collected the data in `DAAG::cuckoos` as shown in Figure 1.1 in order to investigate claims in Newton and Gadow (1896, p. 123) that the cuckoo eggs tend to match the eggs of the host bird in size, shape, and color. Panel A is a dotplot

Table 1.1 *Mean lengths of cuckoo eggs, compared with mean lengths of eggs laid by the host bird species. The table combines information from the two* DAAG *data frames* `cuckoos` *and* `cuckoohosts`.

Host species	Meadow pipit	Hedge sparrow	Robin	Wagtails	Tree pipit	Wren	Yellow hammer
Length (cuckoo)	22.3 (45)	23.1 (14)	22.5 (16)	22.6 (26)	23.1 (15)	21.1 (15)	22.6 (9)
Length (host)	19.7 (74)	20.0 (26)	20.2 (57)	19.9 (16)	20 (27)	17.7 (-)	21.6 (32)

(Numbers in parentheses are numbers of eggs)

display of the raw data. Panel B is the more summary boxplot form of display (to be discussed further in Section 1.1.5) that is designed to give a rough indication of how variation between groups compares with variation within groups.[1]

Table 1.1 adds information that suggests a relationship between the size of the host bird's eggs and the size of the cuckoo eggs that were laid in that nest. Observe that apart from several outlying egg lengths in the meadow pipit nests, the length variability within each host species' nest is fairly uniform.

In the paper (Latter, 1902) that supplied the cuckoo egg data of Figure 1.1 and Table 1.1, the interest was in whether cuckoos do in fact match the eggs that they lay to the host eggs, and if so, in assessing which features match and to what extent.

Uniquely among the birds listed, the architecture of wren nests makes it impossible for the host birds to see the cuckoo's eggs, and the cuckoo's eggs do not match the wren's eggs in color. For the other species the color does mostly match. Latter concluded that the claim in Newton and Gadow (1896) is correct, that the eggs that cuckoos lay tend to match the eggs of the host bird in ways that will make it difficult for hosts to distinguish their own eggs from the cuckoo eggs.

Issues with the data in Table 1.1 and Figure 1.1 are as follows.

- The cuckoo eggs and the host eggs are from different nests, collected over the course of several investigations. Data on the host eggs are from various sources.
- The host egg lengths for the wren are indicative lengths, from Gordon (1894).

There is thus a risk of biases, different for the different sources of data, that limit the inferences that can be drawn. How large, then, relative to statistical variation, is the difference between wrens and other species? Would it require an implausibly large bias to explain the difference? A more formal comparison between lengths for the different species based on an appropriate statistical model will be a useful aid to informed judgment.

Stripped down code for Figure 1.1 is:

```
library(latticeExtra)  # Lattice package will be loaded and attached also
cuckoos <- DAAG::cuckoos
## Panel A: Dotplot without species means added
dotplot(species ~ length, data=cuckoos)  ## `species ~ length` is a 'formula'
## Panel B: Box and whisker plot
bwplot(species ~ length, data=cuckoos)
## The following shows Panel A, including species means & other tweaks
av <- with(cuckoos, aggregate(length, list(species=species), FUN=mean))
```

[1] Subsection A.5.1 has the code that combines the two panels, for display as one graph.

```
dotplot(species ~ length, data=cuckoos, alpha=0.4, xlab="Length of egg (mm)") +
  as.layer(dotplot(species ~ x, pch=3, cex=1.4, col="black", data=av))
  # Use `+` to indicate that more (another 'layer') is to be added.
  # With `alpha=0.4`, 40% is the point color with 60% background color
  # `pch=3`: Plot character 3 is '+'; `cex=1.4`: Default char size X 1.4
```

1.1.3 Planning for a Statistical Analysis

First steps in any coordinated scientific endeavor must include clear identification of the question of interest, followed by careful planning. Consultation with subject-matter specialists, as well as with specialists in statistical aspects of study design, will help avoid obvious mistakes in any of the steps: designing the study, collecting and/or collating data, carrying out analyses, and interpreting results.

If new data are to be acquired, one must decide if a designed experiment is feasible. In human or animal experimentation, such as in clinical trials to test a new drug therapy, ethics are an immediate concern. Data from experiments appear throughout this text – examples are the data on the tinting of car windows that is used for Figure 7.8 in Section 7.5, and the kiwifruit shading data that is discussed in Subsection 1.3.2. Such data can, if the experiment has been well designed with a view to answering the questions of interest, give reliable results. Always, the question must be asked: "How widely do the results generalize?". For example, we might be interested in knowing to what extent the results for the kiwifruit shading conditions can be generalized to other locations with different soil types and weather conditions.

Understand the Data

Most standard elementary statistical methods assume that sample values were all chosen independently and with equal probability from the relevant population. If the data were from an observational study, such as in the cuckoo eggs example of Subsection 1.1.2, special care is required to consider what biases may have been induced by the method of data collection, and to ensure that they do not not lead to incorrect conclusions.

Temporal and spatial dependence are common forms of departure from independence, often leading to more complicated analyses. Data points originating from points that are close together in time and/or space are often more similar. Tests and graphical checks for dependence are necessarily designed to detect specific forms of dependence. Their effectiveness relies on recognizing forms of dependence that can be expected in the specific context.

If the data were acquired earlier and for a different purpose, details of the circumstances that surrounded the data collection are especially important. Were they from a designed experiment? If so, how was the randomization carried out? What factors were controlled? Was there a hierarchical structure to the data, such as would occur in a survey of students, randomly selected from classes, which are themselves randomly selected from schools, and so on? If the data were collected as part of an observational study, such as in the cuckoos example of Subsection

1.1.2, special care is required to ensure that hidden biases induced by the method of data collection do not lead to incorrect conclusions. Biases are likely when data are obtained from "convenience" samples that have the appearance of surveys but which are really poorly designed observational studies. Online voluntary surveys are of this type. Similar biases can arise in experimental studies if care is not taken. For example, an agricultural experimenter may pick one plant from each of several parts of a plot. If the choice is not made according to an appropriate randomization mechanism, a preference bias can easily be introduced.

Nonresponse, so that responses are missing for some respondents, is endemic in most types of sample survey data. Or responses may be incomplete, with answers not provided to some questions. Dietary studies based on the self-reports of participants are prone to measurement error biases. With experimental data on crop or fruit yields, results may be missing for some plots because of natural disturbances caused by animals or harsh weather. One ignores the issue at a certain risk, but treating the problem is nontrivial, and the analyst is advised to determine as well as possible the nature of the missingness. It can be tempting simply to replace a missing height value for a male adult in a dataset by the average of the other male heights. Such a *single imputation* strategy will readily create unwanted biases. Males that are of smaller than average weight and chest measurement are likely to be of smaller than average height. *Multiple imputation* is a generic name for methodologies that, by matching incomplete observations as closely as possible to other observations on the variables for which values are available, aim to fill in the gaps.

Causal Inference

With data from carefully designed experiments, it is often possible to infer causal relationships. Perhaps the most serious danger is that the results will be generalized beyond the limits imposed by the experimental conditions.

Observational data, or data from experiments where there have been failures in design or execution, is another matter. Correlations do not directly indicate causation. A and B may be correlated because A drives B, or because B drives A, or because A and B change together, in concert with a third variable. For inferring causation, other sources of evidence and understanding must come into play.

What Was Measured? Is It the Relevant Measure?

The `DAAG::science` and `DAAG::socsupport` data frames are both from surveys. The former concerns student attitudes towards science in Australian private and public school systems. The latter concerns social and emotional support resources as they might relate to psychological depression in a sample of individuals.

In either case it is necessary to ask: "What was measured?" This question is itself amenable to experimental investigation. For the dataset `science`, what did students understand by "science"? Was science, for them, a way to gain and test knowledge of the world? Or was it a body of knowledge? Or, more likely, was it a label for their experience of science laboratory classes (interesting sights, smells

and sounds perhaps) and field trips? Answers to other questions included in the survey shed some limited light.

In the `socsupport` dataset, an important variable is the Beck Depression Inventory or `BDI`, which is based on a 21-question multiple-choice self-report. It is the outcome of a rigorous process of development and testing. Since its first publication in 1961, it has been extensively used, critiqued, and modified. Its results have been well validated, at least for populations on which it has been tested. It has become a standard psychological measure of depression (see, e.g., Streiner et al., 2014).

For therapies that are designed to prolong life, what is the relevant measure? Is it survival time from diagnosis? Or is a measure that takes account of quality of life over that time more appropriate? Two such measures are "Disability Adjusted Life Years" (DALYs) and "Quality Adjusted Life Years" (QALYs). Quality of life may differ greatly between the therapies that are compared.

Use Relevant Prior Information in the Planning Stages

Information from the analysis of earlier data may be invaluable both for the design of data collection for the new study and for planning data analysis. When prior data are not available, a pilot study involving several experimental runs can sometimes provide such information.

Graphical and other checks are needed to identify obvious mistakes and/or quirks in the data. Graphs that draw attention to inadequacies may be suggestive of remedies. For example, they may indicate a need to numerically transform the data, such as by taking a logarithm or square root, in order to more accurately meet the assumptions underlying a more formal analysis. At the same time, one should keep in mind the risk that use of the data to influence the analysis may bias results.

Subject Area Knowledge and Judgments

Data analysis results must be interpreted against a background of subject area knowledge and judgment. Some use of qualitative judgment is inevitable, relating to such matters as the weight that can be placed on claimed subject area knowledge, the measurements that are taken, the details of study design, the analysis choices, and the interpretation of analysis results. These, while they should be as informed as possible, involve elements of qualitative judgment. A well-designed study will often lead to results that challenge the insights and understandings that underpinned the planning.

The Importance of Clear Communication

When there are effective lines of communication, the complementary skills of a data analyst and a subject matter expert can result in effective and insightful analyses. When unclear about the question of interest, or about some feature of the data, analysts should be careful not to appear to know more than is really the case. The subject-matter specialist may be so immersed in the details of their problem that, without clear signals to the contrary, they may assume similar knowledge on the part of the analyst.

Data-Based Selection of Comparisons

In carefully designed studies where subjects have been assigned to different groups, with each group receiving a different treatment, comparisons of outcomes between the various groups, and of subgroups within those groups (e.g., female/male, old/young) will be of interest. Among what may be many possible comparisons, the comparisons that will be considered should be specified in advance. Prior data, if available, can provide guidance. Any investigation of other comparisons may be undertaken as an exploratory investigation, a preliminary to the next study.

Data-based selection of one or two comparisons from a much larger number is not appropriate, since large biases may be introduced. Alternatively, there must be allowance for such selection in the assessment of model accuracy. The issues here are nontrivial, and we defer further discussion until later.

Models Must Be Fit for Their Intended Use

Statistical models must, along with the data upon which they rely, be applied according to their intended use. Architects and engineers have in the past relied heavily on scale models for giving a sense of important features of a planned building. For checking routes through the building, for the plumbing as well as for humans, such models can be very useful. They will not give much insight on how buildings in earthquake-prone regions are likely to respond to a major earthquake – a lively concern in Wellington, New Zealand, where the first author now lives. For that purpose, engineers use mathematical equations that are designed to reflect the relevant physical processes. The credibility of predictions will strongly depend on the accuracy with which the models can be shown to represent those processes.

1.1.4 Results That Withstand Thorough and Informed Challenge

Statistical models aim to give real-world descriptions that are adequate for the purposes for which the model will be used. What checks will give confidence that a model will do the task asked of it? As argued in Tukey (1997), there must be exposure to diverse challenges that can build (or destroy!) confidence in model-based inferences. We should trust those results that have withstood thorough and informed challenge.

A large part of our task in this text is to suggest effective forms of challenge. Specific types of challenge may include the following.

- For experiments, carefully check and critique the design.
- Look into what is known of the processes that generated the data, and consider critically how this may affect its use and the reliance placed on it. Are there possible or likely biases?
- Look for inadequacies in laboratory procedure.
- Use all relevant graphical or other summary checks to critique the model that underpins the analysis.
- Where possible, check the performance of the model on test data that reflects

the manner of use of results. (If, for example, predictions are made that will be applied a year into the future, check how predictions made a year ahead panned out for historical data.)
- For experimental data, have the work replicated independently by another research group, from generation of data through to analysis.

In areas where the nature of the work requires cooperation between scientists with a wide range of skills, and where data are shared, researchers provide checks on each other. For important aspects of the work, the most effective critiques are likely to come from fellow researchers rather than from referees who are inevitably more remote from the details of what has been done. Failures of scientific processes are a greater risk where scientists work as individuals or in small groups with limited outside checks.

There are commonalities with the issues of legal and medical decision making that receive extensive attention in Kahneman et al. (2021, p. 372), on the benefits of "averaging," that is, using the perspectives of multiple judges as a basis for decision making when sentencing; the authors comment:

> The advantage of averaging is further enhanced when judges have diverse skills and complementary judgment patterns.

Also needed is a high level of shared understanding.

For observational data, the challenges that are appropriate will depend strongly on the nature of the claims made as a result of any analysis. Dangers of over-interpretation and/or misinterpretation of results gleaned from observational data will be exemplified later in the text.

1.1.5 Using Graphs to Make Sense of Data

Ideas of *Exploratory Data Analysis* (EDA), as formalized by John W. Tukey, have been a strong influence in the development of many of the forms of graphical display that are now in wide use. See Hoaglin (2003). A key concern is that the data should as far as possible speak for itself, prior to or as part of a formal analysis.

A use of graphics that is broadly in an EDA tradition continues to develop and evolve. The best modern statistical software makes a strong connection between data analysis and graphics, combining the computer's ability to crunch numbers and present graphs with that of a trained human eye to detect pattern. Statistical theory has an important role in suggesting forms of display that may be helpful and interpretable.

Graphical Comparisons

Figure 1.1 was a graphical comparison between the lengths of cuckoo eggs that had been laid in the nests of different host species. The boxes that give boxplots their name focus attention on quartiles of the data, that is, the three points on the axis that split the data into four equal parts. The lower end of the box marks the first quartile, the dot marks the median, and the upper end of the box marks the third