CHAPTER 1

Introduction

Things vary. If they were all the same, we would not need to collect data and analyze them. Some of that variability is not desirable, but we have tools to recognize that and constructively deal with it. A typical example is an imaging system, starting with your everyday camera or a printer. Manufacturers put a lot of effort into minimizing noise and maximizing consistency of those devices. How is that done? The best way is to start with understanding the system and then measuring its variability. Once you have your measurements, or data, you will need statistical methods to understand and analyze them, so that proper conclusions can be drawn. This is where this book becomes handy. We will show you how to deal with data, how to distinguish between different types of variability, and how to separate the real information from noise.

Statistics is the science of the collection, modeling, and interpretation of data. In this book, we are going to demonstrate how to use statistics in the fields of imaging, optics, and photonics. These are very broad fields—not easy to define. They deal with various aspects of the generation, transmission, processing, detection, and interpretation of electromagnetic radiation. Common applications include the visible, infrared, and ultraviolet ranges of the electromagnetic spectrum, although other wavelengths are also used. This plethora of different measurements makes it difficult to extract useful information from data. The strength of statistics is in describing large amounts of data in a concise way and then drawing general conclusions, while minimizing the impact of data noise on our decisions.

Here are some examples of real, practical problems we are going to deal with in this book.

Example 1.1 (Eye Tracker Data). Eye tracking devices are used to examine people's eye movements as they perform certain tasks (see Pelz et al. (2000)). This information is used in research on the human visual system, in psychology, in product design, and in many other applications. In eye tracking experiments, a lot of data are collected. In a study of 30 shoppers, lasting 20 min per shopper, over one million video frames are generated. In order to reduce the amount of data, fixation periods are identified when a shopper fixes her gaze at one spot. This reduces the number of

Statistics for Imaging, Optics, and Photonics, Peter Bajorski.

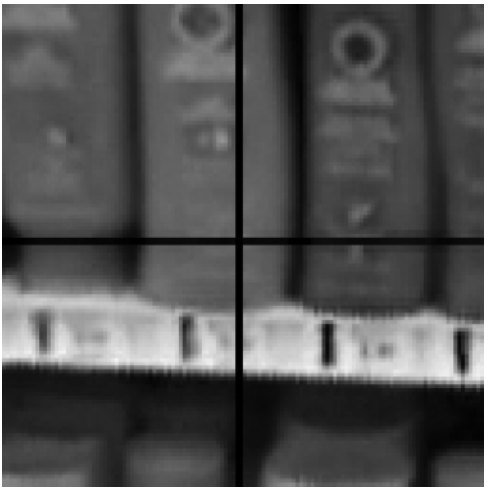

Figure 1.1 Shampoo bottles on a store shelf. The cross shows the spot the shopper is looking at.

frames to under 100,000, but those images still need to be labeled in order to describe what the shoppers are looking at. Many of those images are fixations on the same product, but possibly from a different angle. The image frame might also be slightly shifted. Our goal is to find the groups of images of the same product. One approach could be to compare the images pixel by pixel, but that would not work well when the image is shifted. One could also try to segment the image into identifiable objects and then compare the objects from different images, but that would require a lot of computations. Another approach is to ignore the spatial structure of the image and describe the image by how the three primary colors mix in the image.

Figure 1.1 shows a sample fixation image used in a paper by Kinsman et al. (2010). The cross in the image shows the spot the shopper is looking at. This 128 by 128 pixel image was recorded with a camcorder in the RGB (red, green, and blue) channels. This means that each pixel is represented by a mixture of the three colors. Mathematically, we can describe the pixel with three numbers, each representing the intensity of one of the colors. For educational purposes, we select here a small subset of all pixels and use only the red and green values. Figure 1.2a shows a scatter plot of this small subset. We can see some clusters, or concentrations, of points. Each cluster corresponds to a group of pixels with a given mix of color. The group in the top right corner of the graph is a mix of a large amount of red with a large amount of green.

Our goal is to find those clusters automatically and describe them in a concise way. This is called unsupervised learning because we learn about the clusters without prior information (supervision) about the groups. One possible solution is shown in Figure 1.2b, where five clusters are identified and described by the elliptical shapes. This provides a general structure for the data. In a real implementation, this needs to be done on all 16,384 pixels in a three-dimensional space of the red, green, and blue intensity values. Methods for efficient execution of such tasks will be shown in this book.

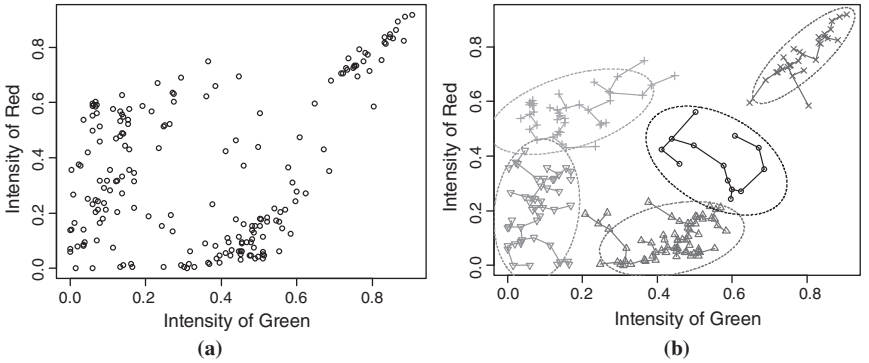

Figure 1.2 Understanding structure of data. Original data are shown in panel (a) and clusters of the same data with elliptical descriptors are shown in panel (b).

Example 1.2 (Printing Data). Printer manufacturers want to ensure high consistency of printing by their devices. There are various types of calibrations and tests that can be done on a printer. One of them is to print a page of random color patches such as those shown in Figure 1.3. The patches are in four basic colors of the CMYK color model used in printing: cyan, magenta, yellow, and black. In a given color, there are several gradations, from the maximum amount of ink to less ink, where the patch has a lighter color if printed on a white background. For a given gradation of color, there are several patches across the page printed in the same color. Our goal is to measure the consistency of the color in all those patches. We also want to monitor printing quality over time, including possible changes in quality after the printer's idle time. An experiment was performed to study these issues, and the resulting data set is used throughout this book. Methods for exploratory analysis of such data and then for statistical inference will be discussed.

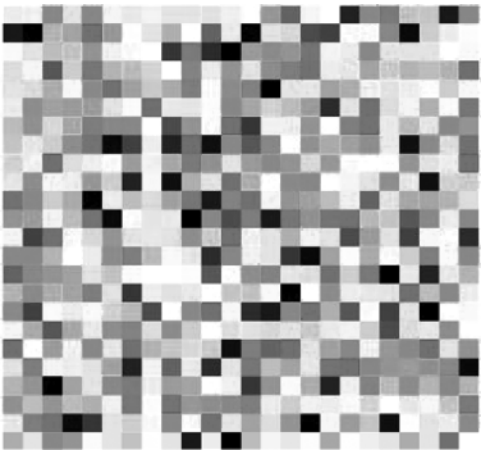

Figure 1.3 Random color patches for printing-quality testing.

Example 1.3 (Remote Sensing Data). Remote sensing is a broad concept of taking measurements, or making observations, from a distance. Here, we concentrate on spectral images of the Earth from high altitudes by way of aircraft or satellite. Digital images consist of pixels, each pixel representing a small area in the image. In a standard color photograph, a pixel can be represented by a mixture of three primary colors—red, green, and blue. Each color represents a certain wavelength range of the visible light. Different materials reflect light in different ways, which is why they have different colors. Colors provide a lot of information about our environment. A color photograph is more informative than a black-and-white one. Even more information can be gathered when the visible spectrum is divided into, let's say, 31 spectral bands and the reflectance is measured separately in each band. Now we can see a difference between two materials that look the same to a human eye. In the same way, we can measure reflectance of electromagnetic waves in other (invisible) wavelengths, including infrared, ultraviolet, and so on. The amount of information increases considerably, but this also creates many challenges when analyzing such data. Each pixel is now represented by a spectral curve describing reflectance as a function of wavelength. The spectral curves are often very spiky with not much smoothness in them. It is then convenient to represent them in their original digitized format, that is, as p-dimensional vectors, where p is the number of spectral wavelengths. The number p is often very large, sometimes several hundred or even over a thousand. This creates major difficulties with visualization and analysis of such data. In Figure 1.2, we saw a scatter plot of two-dimensional data, but what do we do with 200-dimensional data? This book will show you how to work in very high dimensional spaces and still be able to extract the most important information.

Remote sensing images are used in a wide range of applications. In agriculture, one can detect crop diseases from aerial images covering large areas. One example that we are going to use in this book is an image of grass area, where a part of the image was identified as representing diseased grass. Our goal is to learn how to recognize diseased grass based on a 42-dimensional spectral vector representing a pixel in the image. We can then use this information to classify spectra in future images into healthy or diseased grass. This learning process is called supervised learning because we have prior information from the image on how the healthy grass and the diseased grass look in terms of their spectrum. Once we know how to differentiate the two groups based on the spectra, we can apply the method to large areas of grass.

The diseased grass does not look much different from the healthy grass, if you are assessing it visually or looking at a color photograph. However, there is more information in 42 dimensions, but how can we find it and see it? In this book, we will show you methodologies for finding the most relevant information in 42 dimensions. We will also find the most informative low-dimensional views of the data. Figure 1.4 shows an optimal way of using two dimensions for distinguishing between three types of grass pixels—the healthy grass (Group 1), less severely diseased grass (Group 2), and severely diseased grass (Group 3). The straight lines show the optimum separation between the groups for the purpose of classification, and the ellipses show an attempt to describe the variability within the groups. In this book, we will show you how to construct such separations, how to evaluate their efficiency,

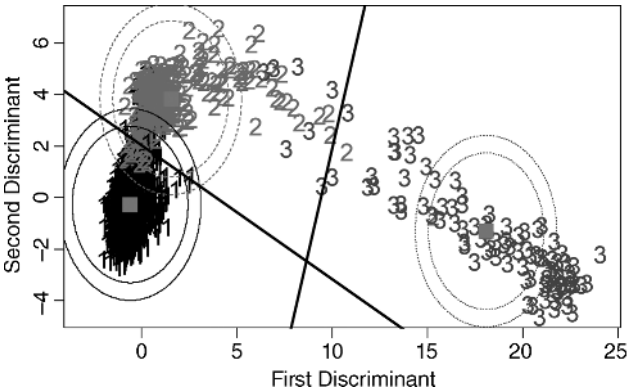

Figure 1.4 A two-dimensional representation of a 42-dimensional set of image pixels representing healthy grass (Group 1), less severely diseased grass (Group 2), and severely diseased grass (Group 3).

how to describe the variability within groups, and then check how reliable such descriptions are.

Example 1.4 (Statistical Thinking). Even before any data are collected, we need to utilize statistical thinking so that our study is scientifically valid and the conclusions are representative of the intended scope of the study. Whenever we use data and try to analyze them, we need to take the following three steps:

1. Formulate the practical problem at hand as a statistical problem.
2. Solve the problem using statistics. This usually involves the collection and analysis of data.
3. Translate the problem solution back to the real-world application.

The purpose of this book is to show you how to solve practical problems by using this statistical approach. Let's say you are a quality engineer at Acme Labs producing plastic injection molding parts. You are part of a team assigned to provide a sensor for automatically detecting whether the produced parts have an acceptable shade of a chosen color. Many steps are needed to accomplish the task, but here we give an example of two steps where statistics would be useful:

1. Define what it means that a color shade is acceptable or not.
2. Find and test an instrument that would measure the color with sufficient precision at a reasonable cost.

The color shade acceptability is somewhat subjective and will depend on the observer and viewing conditions when the material is compared visually. See Berns (2000) for a more detailed discussion of color and color measurement. In this book, we will focus on instrumental color measurement. The produced parts of nominally the same color

will vary slightly in the color shade, possibly due to variation in the production process. Instrumental measurements of the color will also vary. All those sources of variability can be measured and described using statistical methods. It would be best to know the variability of all produced parts, all possible measurements made with a given instrument, and all possible observers. However, it is either impossible or impractical to gather all that knowledge. Consequently, in statistics we deal with samples, and we determine to what extent a sample represents the whole population that it is attempting to describe.

Throughout this book, we are going to use the examples described above, as well as many others, to illustrate real-world applications of the discussed statistical methods.

1.1 WHO SHOULD READ THIS BOOK

This book is primarily intended for students and professionals working in the fields of imaging, optics, and photonics. Hence, all examples are from these fields. Those are vast areas of research and practical applications, which is why the examples are written in a simplified format, so that nonexperts can relate to the problem at hand. Nevertheless, this book is about statistics, and the presented tools can be potentially useful in any type of data analysis. So, practitioners in other fields will also find this book useful.

The reader is expected to have some prior experience with quantitative analysis of data. We provide a gentle and brief introduction to data analysis and concentrate on explaining the associated concepts. If a reader needs more practice with those tools, it is recommended that other books, with a more thorough coverage of fundamentals, are studied first.

Some experience with vector and matrix algebra is also expected. Familiarity with linear algebra and some intuition about multidimensional spaces are very helpful. Some of that intuition can be developed by working slowly through Chapter 5.

This book is not written for statisticians, although they may find it interesting to see how statistical methods are applied in this book.

1.2 HOW THIS BOOK IS ORGANIZED

This chapter is followed by two chapters that review the fundamentals needed in subsequent chapters. Chapter 2 covers the tools needed for exploratory data analysis as well as the probability theory needed for statistical inference. In Chapter 3, we briefly introduce the fundamental concepts of statistical inference. The regression models covered in Chapter 4 are very useful in statistical analysis, but that material is not necessary for understanding the remaining chapters. On the other hand, two supplements to that chapter provide the fundamental information about vector and matrix algebra as well as random vectors, all needed in the following chapters.

Starting with Chapter 5, this book is about multivariate statistics dealing with various structures of data on multiple variables. We lay the foundation for the

multidimensional considerations in Chapter 5. This is where a reader comfortable with univariate statistics could start reading the book. Chapter 6 covers basic multivariate statistical inference that is needed in specific scenarios but is not necessary for understanding the remaining parts of the book. Principal component analysis (PCA) discussed in Chapter 7 is a very popular tool in the fields of imaging, optics, and photonics. Most professionals in those fields are familiar with PCA. Nevertheless, we recommend reading that chapter, even for those who believe they are familiar with this methodology. We are aware of many popular misconceptions, and we clarified them in that chapter. Each of the remaining chapters moves somewhat separately in three different directions, and they can be read independently. Chapter 8 covering canonical correlation analysis is difficult technically. In Chapter 9, we describe classification, also called supervised learning, which is used to classify objects into populations. Clustering, or unsupervised learning, is discussed in Chapter 10, which can be read independently of the majority of the book material.

1.3 HOW TO READ THIS BOOK AND LEARN FROM IT

Statistics is a branch of mathematics, and it requires some of the same approaches to learning it as does mathematics. First, it is important to know definitions of the terms used and to follow the proper terminology. Knowing the proper terminology will not only make it easier to use other books on statistics, but also enable easier communication with statisticians when their help is needed. Second, one should learn statistics in a sequential fashion. For instance, the reader should have a good grasp of the material in Chapters 2 and 3 before reading most of the other parts of this book. Finally, when reading mathematical formulas, it is important to understand all notation. You should be able to identify which objects are numbers, or vectors, or matrices—which are known, which are unknown, which are random or fixed (nonrandom), and so on. The meaning of the notation used is usually described, but many details can also be guessed from the context, similar to everyday language. When writing your own formulas, you need to make sure that a reader will be able to identify all of the features in your formulas.

As with all areas of mathematics and related fields, it is critical to understand the basics before the more advanced material can be fully mastered. The particular difficulty for many nonstatisticians is the full appreciation of the interplay between the population, the model, and the sample. Once this is fully understood, everything else starts to fall into place.

Each chapter has a brief list of problems to practice the material. The more difficult problems are marked by a star. We recommend that the readers' main exercise be the recreation of the results shown in the book examples. Once the readers can match their results with ours, they would most likely master the mechanics of the covered methodologies, which is a prerequisite for their deeper understanding. Most concepts introduced in this book have very specific geometric interpretations that help in their understanding. We use many figures to illustrate the

concepts and elicit the geometric interpretation. However, readers are encouraged to sketch their own graphs when reading this book, especially some representations of vectors and other geometric figures.

Real-world applications are usually complex and require a considerable amount of time to even understand the problem. For educational purposes, we show simplified versions of those real problems with smaller data sets and straightforward descriptions, so that nonexperts can easily relate to them.

We often provide references where the proofs of theorems can be found. This is not meant as a recommendation to read those proofs, but simply as potential further reading for more theoretically inclined readers. We provide derivations and brief proofs only in cases when they are simple and provide helpful insight and illustration of the introduced concepts.

In this book, we try to keep the mathematical rigor at an intermediate level. For example, the main statistical theme of distinguishing between the population and the sample quantities is emphasized, but only in places where it is necessary. In other places, readers will need to keep track of those subtleties on their own, using their statistical thinking skills, hopefully developed by that time.

We use mathematical notation and formulas generously, so readers are encouraged to overcome their fear of formulas. We treat mathematical language as an indispensable tool to describe things precisely. As with any other language learning, it becomes easier with practice. And once you know it, you find it useful, and you cannot resist using it.

We abstain from a mathematical tradition of reminding the reader that the introduced objects must exist before one can use them. We usually skip the assumptions that sets are nonempty and the numbers we use are finite. For example, if we write a definite integral, we implicitly assume that it exists and is a finite number.

1.4 NOTE FOR INSTRUCTORS

The author has used the multivariate material of this book in a 10-week graduate course on multivariate statistics for imaging science students. With the additional material developed for this book and the review of the univariate statistics, the book is also suitable for a similar 15-week course. The author's experience is that some review of the material in Chapters 2, 3 and 4 is very helpful for students for a better understanding of the multivariate material. The computational results and graphs in this book were created with the powerful statistical programming language R (see R Development Core Team (2010)). However, students would usually use their preferred software, such as ENVI/IDL or MATLAB. It is our belief that students benefit from implementing statistical techniques in their own computational environment rather than using a statistical package that is chosen for the purpose of the course and possibly never again used by the students. This is especially true for students dealing with complex data such as those used in imaging, optics, and photonics.

1.5 BOOK WEB SITE

The web site for this book is located at

http://people.rit.edu/∼pxbeqa/ImagingStat

It contains data sets used in this book, color versions of some of the book figures (if the color is relevant), and many other resources.