Essential Statistical Concepts for the Quality Professional.


The main points of Essential Statistical Concepts for the Quality Professional are (1) Preface and Introduction, (2) Chapters 1–10 (pp. 1–177), (3) Appendices A-G (pp. 181–304), (4) Glossary (pp. 305–468). As you can see, the chapters make up well under half of the book, with much of the remaining material comprising the appendices and the glossary.

The Preface and Introduction provide an introduction to problem solving using statistical methods. This includes a good description of Deming’s Plan—Do—Check—Act cycle, which the author modifies to Plan—Do—Check/Study—Act—Ingrain.

Chapters 1 through 10 cover a myriad of topics, usually in little depth. There are many cases where the topics are simply mentioned, with no guidance on how to use the method or even a reference on where to study more. For example, in just under four pages (pp. 131–135) the author has sections that are entitled: Discriminant Analysis, Log-Linear Models, Logistic Regression, Factor Analysis, and Cluster Analysis. In these four pages, there is not one reference to point the reader to a more in-depth presentation. There are, however, references in a selected bibliography at the end of each chapter.

The most striking aspect of Chapters 1–10 is that there is not a single example given to help the reader see how the method can be applied. The “List of Illustrations” on p. xiii lists only seven figures for the entire book, and none of these involve any data. Most readers who find the need to look up statistical methods will want to see examples of how the methods (and the formulas) are applied. This book provides many formulas, with minimal explanation and no worked out examples. Although it would seem that control charts are an essential concept for quality professionals, there is no chapter on control charts (although Appendix C does cover this topic).

As a rule, the book defines population statistics such as the mean and variance as averages across a finite population, rather than as expected values. For example, on p. 244, we find

\[ \sigma^2 = \frac{\sum (x - \mu)^2}{N} \]

(yes, the equal sign is missing) rather than the more familiar and generalizable

\[ \sigma^2 = E[(x - \mu)^2] \]

“Expected value” is never defined, although the term is used once in Appendix C on p. 237 in the context of the sampling distribution of W. The reader is left to guess what W means.

Appendix A describes the use of Minitab. This chapter is well-written and could be informative as long as the reader has the same version of Minitab. (The author does not say what version of Minitab he used.) Even small changes in software can affect presentations such as Appendix A. Nothing is more frustrating, for example, than to be told to click on a menu item that is not there, or has been renamed. Thus, while this appendix will be helpful for a while, it will quickly become dated. Appendix C is a collection of dozens of formulas, often, as is the case with the W mentioned above, the context is absent. There are a number of errors in Appendix C and elsewhere. For example, just on pp. 232–233, there are three mistakes (formulas 17, 19, and 25), not counting the claims that “e = 2.71828” and “r = 3.14159.” Formula 34 gives “Summary measures of the sampling distribution of P” with no indication of what P means. Closer inspection suggests that the author is talking about the sampling distribution of \( \hat{p} = X/n \) (the number of successes divided by the number of trials), that is, the usual estimator of the success probability for the binomial distribution. In addition, Appendix C covers formulas for control charts, including familiar ones like this for the \( \bar{X} \) chart: \( UCL = \bar{X} + A_2 \bar{R} \). Readers familiar with statistical process control will recognize the numerous control chart constants such as \( A_2 \). Although many of these control chart constants are mentioned, there is no table for the reader to look up values for them. In fact, there are no tables for any distribution in the book. Thus, a reader who needs to study how to apply a particular chart will have to look it up in this book, find another book with an example, and (possibly) another book with the required tables. Appendix F is entitled “Hypothesis Testing—Selected Explanations and Examples.” This chapter does contain some examples, but in some cases this involves just numbers with no context. The reader must also be cautious because there are numerous errors. On p. 259, the author says “Since \( |z_0| < |z_{0.05}| \), reject \( H_0 \) [We deal with a t-test, we use absolute values].” The inequality is obviously wrong, but the second statement is wrong too; when we have a one-sided alternative, for example, we do not use absolute values. Further down the page, in an example of the two-sample t-test, we see that “\( n_1 = n_2 = 4 \) df=8.”

The Glossary is quite thorough, containing over 1500 entries. One might quibble with a few definitions, such as that for “treatment” which confuses factor with treatment. The definition of “confounding” covers only the confounding that occurs in an experimental design when, for example, main effects are confounded with interactions, and not the more general idea that we are unable to separate estimates of effects. Clearly, the glossary is the strongest part of the book.

As a reference book on the use of statistical concepts in quality, this book fails. A typical reader will require more thorough explanations and illustrations on real or realistic data.

Steven E. RIGDON
Saint Louis University

Introduction to Statistical Methods for Biosurveillance.


Bioterrorism is increasingly becoming a prominent threat to countries all over the world. From a public health perspective, the early detection of a potential bioterrorism incident is critical, as it allows timely communication with critical organizations, such as health providers, who attempt to limit the damage. Effective biosurveillance is a critical capability for quickly identifying and monitoring health threats that are not visible to health providers. Introduction to Statistical Methods for Biosurveillance introduces the state-of-the-art in terms of design and evaluation of statistical methods for biosurveillance, and is intended for readers with a modest prior background in statistics and probability.

The book is well written and easy to read, with a useful organization. Each chapter opens with an overview and a clear definition of its objectives, and ends with a discussion that summarizes the chapter and leads to the next one. A list of additional reading is also given. When relevant, a table of the mathematical notation used in the chapter is provided right after the objectives. Though exercises at the end of chapters would be beneficial, I find this book suitable as a textbook for both masters- and doctoral-level students in business and applied statistics. I would also recommend this book as a useful reference for researchers with an interest in disease surveillance.

Part I of the book ("Introduction") presents the conceptual framework of biosurveillance, with an emphasis on the topic of syndromic surveillance. The first chapter begins with definition and objectives of biosurveillance, epidemiological surveillance, and syndromic surveillance. In essence, this chapter sets the reader’s expectations and defines what is within the scope of the book and what is beyond the scope.

The second chapter of the introduction is a 27-page description of common types of surveillance data that are used in the book. Unfortunately, this chapter does not mention more advanced data types, such as textual data (e.g., physicians’ notes) and network or geographic data (e.g., geographic distances between hospitals, or networks of individuals), nor are they mentioned later in the book. The author then touches upon data preparation techniques, that is, data cleaning, coding, and imputation. This chapter is a must-have foundation for the statistical methods that follow, which assume availability of data, their completeness, and their quality. I found the second half of this chapter (data preparation) too short for the target audience, which is defined in the preface as students in an “advanced undergraduate or beginning graduate-level [statistics and probability] course.” In particular, data coding requires much more reading than the author suggests, or at least a moderate programming background.

Following the introduction is a thorough survey of the most commonly used statistical methods for disease monitoring. The statistical methods in the book are divided into two parts: Situational Awareness (Part II), where the focus is on understanding current data in the absence of disease activity, and Early Event Detection (Part III) that, as conveyed by its title, discusses statistical (univariate and multivariate) methods for early detection of health events.
fundamentals of each method are carefully reviewed with examples of their use and application to biosurveillance.

The distinction between situational awareness and early event detection is clear and important. By separating the statistical methods into these two parts, the author makes a coherent statement: first understand the baseline and the data at hand. If necessary, process the data and prepare them for further analysis. Then, monitor the data for early detection of anomalies.

Further to my previous comment, I found Chapter 4 in Part II (“Descriptive Statistics for Comprehending the Situation”) to be a great strength of the book. The chapter focuses on statistical tools useful for comprehending the data, including descriptive statistics, exploratory techniques, and data visualization. Applying a combination of these tools provides an understanding of the data dynamics and characteristics, which constitute the core of situational awareness.

Chapter 5 in Part II (“Statistical Methods for Projecting the Situation”) surveys methods for data projection. The author argues that the methods presented are useful both for situational awareness and for data preprocessing as preparation for event detection. As such, I feel that data projection and preprocessing should be combined and constitute a separate section in the book. From the book, an untrained reader might think that the goals of data projection and data preprocessing are essentially the same. In practice, however, although the same techniques are used, they serve different purposes. When projection is used in situational awareness, the objective is to project (forecast) the next state, given the current state, which may or may not contain an outbreak. In contrast, for data preprocessing, the methods intend to project (or isolate) only the baseline, assuming the absence of outbreaks. The distinction between projection and preprocessing may affect how the methods are tuned. For example, with smoothing methods, we might choose a smoothing coefficient with a larger value when the goal is preprocessing, compared to projection purposes (where over-fitting is a concern). Similarly, choosing a smoothing coefficient that minimizes MSE, as suggested on page 123, is a good approach only when the goal is preprocessing.

However, when “projecting the current situation into the future” is considered, where the goal is “to provide the decision maker with an understanding of what the near-term trends are likely to be” (p. 111), a local, rather than global MSE minimization approach, is typically more appropriate.

Part III: Early Event Detection (EED) is very comprehensive. It presents a wide range of important statistical techniques for anomaly detection and their adjustments for biosurveillance. I especially liked the distribution-free definition, where the data are not assumed to follow a normal distribution (as they usually are nonnormal). The organization and flow of this section is especially appealing. Definitions, objectives, and types of EED methods are presented in the first chapter. These are followed by statistical methods for EED, naturally divided into univariate methods (Chapter 7) and multivariate methods (Chapter 8).

For each of the methods, the author presents the fundamentals and how to appropriately apply the method to the biosurveillance problem. Implementation issues are also discussed and illustrated.

Of all sections, Part IV: Putting It All Together, is the most important in the book. This fairly short part conveys the practical aspects of the statistical methods, their interpretations, and performance evaluation in the context of biosurveillance. Though examples are presented throughout Parts I through III, in Part IV, the author takes another step and presents the use and performance of the methods on real biosurveillance datasets. This applied part contrasts with the theoretical and mathematical nature of the book up to this point.

The book ends with two useful appendices: “A Brief Review of Probability, Random Variables, and Some Important Distributions,” and “Simulating Biosurveillance Data.” The appendices are aimed at the “tail” users: users with little to no statistical training (the first appendix), and readers within the professional healthcare community that use this book as a reference volume (the second appendix).

In the preface to Introduction to Statistical Methods for Biosurveillance, the author states that the book’s focus is on basic methods for two reasons. One, these are commonly used in practice, and two, they usually perform as well as (if not better than) their more complex alternatives. I think this is the essence of efficient biosurveillance. Often I read papers that introduce mathematically complex surveillance methods that theoretically outperform currently used methods, such as those presented in the book. However, complexity has its price: for the mathematics to work, some assumptions on the underlying data have to be made. These assumptions are commonly violated when real biosurveillance data are involved. Moreover, with health practitioners being the end users of such methods, the implementation and interpretation of complicated methods makes them largely unaccepted, and hence they are rarely used in practice.

I. YAHAN
The Bar Ilan Graduate School of Business Administration

Introduction to Statistics Through Resampling Methods and R (2nd ed.).

The second edition of Introduction to Statistics Through Resampling Methods and R is again intended to guide students through introductory topics, including some resampling methods, while making use of the R programming language. Much of the second edition remains the same as the first edition, with a few notable changes. In particular, several of the chapters have been restructured. The third and fourth chapters from the first edition “Distributions” and “Testing Hypotheses” have been split into three chapters in the new edition: “Chapter 3: Two Naturally Occurring Probability Distributions,” “Chapter 4: Estimation and the Normal Distribution,” and “Chapter 5: Testing Hypotheses.” There is a short, new chapter (Chapter 7) entitled “Guide to Entering, Editing, Saving, and Retrieving Large Quantities of Data Using R” that provides instructions for dealing with datasets that are generally too large to type in R directly (which is what is commonly done in the earlier chapters). Additionally, a small appendix with selected answers to 10 of the book’s exercises is provided. These 10 solutions, along with R code/datasets, are also available for download at booksupport.wiley.com. Unfortunately, while there are a sizable number of exercises in the book (the back cover claims over 250 exercises), solutions are not available for most of them, and instructors will be forced to write solutions themselves. Finally, the small appendix devoted to S-Plus has been removed, and the text now appears to be solely devoted to R.

While some of the mistakes and poorly explained passages described in the review of the first edition (Hardin, 2006) have been fixed, in my opinion, a few serious issues still remain in the second edition. Although resampling methods seem to be given a little more emphasis in this edition, the book still seems light on these methods, given the use of “resampling” in its title. Readers looking for an in-depth resampling-based introductory statistics book will probably want to look elsewhere, as most of the book still has very little to do with resampling methods.

Additionally, not much detail is given in the provided R code (many of the snippets of code have few to no comments provided). In my opinion, this is still one of the weakest parts of the book—very little time is spent showing readers how to use R and few details are given throughout. While Section 1.3.1 “Learning to Use R” starts to provide basic information about R, it is only two pages long and doesn’t go in to much depth. Accessing the help menu is only briefly mentioned in Section 1.4.1 (page 8), with a figure that shows little about how the help menu operates.

Ultimately, I agree with the review of the first edition that a firm understanding of programming is needed to do more than just copy code from the book, and I would be hesitant to recommend it for an introductory statistics course.

Ivan P. RAMLER
St. Lawrence University

REFERENCE

Predicting Presidential Elections and Other Things (2nd ed.).

At first glance, Predicting Presidential Elections and Other Things seems like it is another popular-statistics book following in the footsteps of Freakonomics.