Preface

The fact that you use biostatistics in your work does not say much about who you are. You may be a physician who has collected some data and is trying to write up a publication, or you may be a theoretical statistician who has been consulted by a physician, who has in turn collected some data and is trying to write up a publication. Whichever you are, or if you are something in between, such as a biostatistician working in a pharmaceutical company, the chances are that your perception of statistics to a large extent is driven by what a particular statistical software package can do. In fact, many books on biostatistics today seem to be more or less extended manuals for some particular statistical software. Often there is only one software package available to you, and the analysis you do on your data is governed by your understanding of that software. This is particularly apparent in the pharmaceutical industry.

However, doing biostatistics is not a technical task in which the ability to run software defines excellence. In fact, using a piece of software without the proper understanding of why you want to employ statistical methods at all, and what these methods actually provide, is bad statistics, however well versed you are in your software manual and code writing. The fundamental ingredient of biostatistics is not a software package, but an understanding of (1) whatever biological/medical aspect the data describe and (2) what it is statistics actually contribute. Statistics as a science is a subdiscipline of mathematics, and a proper description of it requires mathematical formulas. To hide this mathematical content within the inner workings of a particular software package must lead to an insufficient understanding of the true nature of the results, and is not beneficial to anyone.

Despite its title, this book is not an introduction to biostatistics aimed at laymen. This book is about the concepts, including the mathematical ones, of the more elementary aspects of biostatistics, as applied to medical problems. There are many excellent texts on medical statistics but one cannot cover everything and many of them emphasize the technical aspects of producing an analysis at the expense of the mathematical understanding of how the result is obtained. In this book the emphasis is reversed. These other books have a more systematic treatment of different types of problems and how you obtain the statistical results on different types of data. The present volume differs from others in that it is more concerned with ideas, both the particular aspects concerned with the role of statistics in the scientific process of obtaining evidence, and the mathematical ideas that constitute the basis of the subject. It is not a textbook, but should be seen as complementary to more traditional textbooks; it looks at the subject from a different angle, without being in conflict with them. It uses non-conventional and alternative approaches to some statistical concepts, without changing their meaning in any way. One such difference is that key computational aspects are often replaced by graphs, to illustrate *what* you are doing instead of *how*.

The ambition to discuss a wide range of concepts in one book is a challenge. Some concepts are philosophical in nature, others are mathematical, and we try to cover both. Broadly speaking, the book is divided into three major parts. The first part, Chapters 1–5, is concerned with what statistics contribute to medical research, and discusses not only the underlying philosophy but also various issues that are related to the art of drawing conclusions

from statistical output. For this we introduce the concept of the confidence function, which helps us obtain both p-values and confidence intervals from graphics alone. In this part of the book we mostly discuss only the simplest of statistical data, in the form of proportions. We need a background to have the discussion on, and this simple case contains almost all of the conceptual problems in statistics.

The second part consists of Chapters 6–8, and is about generalizing frequency data to more general data. We emphasize the difference between the observed and the infinite truth, how population distributions are estimated by empirical (observed) distributions. We also introduce bivariate distributions, correlation and the important law of nature called 'regression to the mean'. These chapters show how we can extend the way we compare proportions for two groups to more general data, and in the process emphasize that in order to analyze data, you need to understand what kind of group difference you want to describe. Is it a horizontal shift (like the t-test) or a vertical difference (non-parametric tests)? A general theme here, and elsewhere, is that model parameters are mostly estimated from a natural condition, expressed as an estimating equation, and not really from a probability model. There are intimate connections between these, but this view represents a change to how estimation is discussed in most textbooks on statistics.

The third part, the next four chapters, is more mathematical and consists of two subparts: the first discusses how and why we adjust for explanatory variables in regression models and the other is about what it is that is particular about survival data. There are a few common themes in these chapters, some of which are build-ups from the previous chapters. One such theme is heterogeneity and its impact on what we are doing in our statistical analysis. In biology, patients differ. With some of the most important models, based on Gaussian data, this does not matter much, whereas it may be very important for non-linear models (including the much used logistic model), because there may be a difference between what we think we are doing and what we actually are doing; we may think we are estimating individual risks, when in fact we are estimating population risks, which is something different. In the particular case of survival data we show how understanding the relationship between the population risk and the individual risks leads to the famous Cox proportional hazards model.

The final chapter, Chapter 13, is devoted to a general tie-up of a collection of mathematical ideas, spread out in the previous chapter. The theme is estimation, which is discussed from the perspective of estimating equations instead of the more traditional likelihood methods. You can have an estimating equation for a parameter that makes sense, even though it cannot be derived from any appropriate statistical model, and we will discuss how we still can make some meaningful inference.

As the book develops, the type of data discussed grows more and more complicated, and with it the mathematics that is involved. We start with simple data for proportions, progress to general complete univariate data (one data point per individual), move on to consider censored data and end up with repeated measurements. The methods described are developed by analogy and we see, for example, the Wilcoxon test appear in different disguises.

The mathematical complexity increases, more or less monotonically, with chapter number, but also within chapters. On most occasions, if the math becomes too complicated for you to understand the idea, you should move to the next chapter, which in most cases start out simpler. The mathematical theory is not described in a coherent and logical way, but as it applies locally to what is primarily a statistical discussion, and it is described in a variety of different ways: to some extent in running text, with more complex matters isolated in stand-alone text boxes,

while even more complex aspects are summarized in appendices. These appendices are more like isolated overviews of some piece of mathematics relevant to the chapter in question. All mathematical notation is explained, albeit sometimes rather intuitively, and for some readers it may be wise to 'hum' their way through some more complicated formulas. In that way it should be possible to read at least half the book with only minor mathematical skills, as long as one is not put off by the existence of such equations as one comes across. (If you are put off by formulas, you need to get another book.) As already mentioned, at least some of the repetitive (and boring) calculations in statistics have been replaced by an extensive use of graphs. In this way the book attempts to do something that is probably considered almost impossible by most: to simultaneously speak to peasants in peasant language and the learned in Latin (this is a free translation of an old Swedish saying). But there is a price to pay for targeting a wide audience: we cannot give each individual reader the explanation that he or she would find the most helpful. No one will find every single page useful. Some parts will be only too trivial to some, whereas some parts will be incomprehensible to others. There are therefore different levels at which this book can be read.

If you are medically trained and have worked with statistics, in particular p-values, to some extent, your main hurdle will probably be the mathematics. Your priority should be to understand what things intuitively mean, not only the statistical philosophy but also different statistical tests. There is no specific predefined level of mathematics, above basic high-school math, that you need to master for most parts of the book. You only need some basic understanding of what it is a formula tries to say, in order to grasp the story, and you do not need to understand the details of different formulas. To understand a mathematical formula can mean different things, and all formulas definitely do not need to be understood by everyone. The non-trivial mathematics is essentially only that of differentiation and integration, in particular the latter, which most people in the target readership are expected to have encountered at least to some degree. An integral is essentially only a sum, albeit made up of a vast number of very, very small pieces. If you see an integral, it may well suffice to look upon it as a simple sum and, instead of getting agitated, leave such higher-calculus formulas to be read by those with more mathematical interest and skill.

On a second level, you may be a reader who has had basic training in statistics and is working with biostatistics. Being a professional statistician nowadays does not necessarily mean that you have much mathematical training. Hopefully you can make sense of most of the equations, but you may need to consult a standard textbook or other references for further details.

The third level is when you are well versed in reading mathematical textbooks, deriving formulas and proving theorems. For you, the main reason for reading this book may be to get an introduction to biostatistics in order to see whether you want to learn more about the subject. For you, the lack of mathematical details should not be a problem; most left-out steps are probably easily filled in. At this point I beg the indulgence of any mathematician who has ventured into this book and who sees that mathematical derivations are not completely rigorous but sacrificed for the sake of more intuitive 'explanation'. It must also be noted that this book is not an introduction to what to consider when you work as a biostatistician. It may be helpful in some respects, but there is most often an initial hurdle to such work, not addressed in this book, which is about being able to translate biological or medical insight and assumptions to the proper statistical question.

These three levels represent a continuum of mathematical skills. But remember that this book is not a textbook. We use mathematics as a tool for description, an essential tool, but we do not transform biostatistics into a mathematical subdiscipline. One aspect of mathematics is notation. Proper and consistent use of mathematical notation is fundamental to mathematics. In this book we do not have such aspirations, and are therefore occasionally slack in our use of notation. Our notation is not consistent between chapters, and sometimes not even within chapters. Notation is always local, optimized for the present discussion, sacrificing consistency throughout. On most occasions we use capital letters to denote stochastic variables and lower case letters to denote observations, but occasionally we let lower case letters denote stochastic variables. Sometimes we are not even explicit about the change from the observation to the corresponding stochastic variable. Another example is that is not always well defined whether a vector is a column vector or row vector, it may change state almost within a sentence. If you know the importance of this distinction, you can probably identify which it is from the context. This sacrifice is made because I believe it increases readability.

All chapters end with some suggestions on further reading. These are unpretentious and incomplete listings, and are there to acknowledge some material from which I have derived some inspiration when writing this book.

I am deeply grateful to Professor Stephen Senn for the strong support he has given to the project of finalizing this book and for the invaluable advice he has given in the course of so doing. It has been a long-standing wish of mine to write this book, but without his support it is very doubtful that it would ever have happened. I also want to give credit to all those (to me unknown) providers of information on the internet from which I have borrowed, or stolen, a phrase now and then, because it sounded much better than the Swenglish way I would have written it myself. In addition, I want to thank a number of present or past colleagues at the AstraZeneca site where I have worked for the past 25 years, but which the company has decided to close down at more or less the same time as this book is published, in particular Tore Persson and Tobias Rydén, who, despite conflicting priorities, provided helpful comments. Finally, I also want to thank my father and Olivier Guilbaud for input at earlier stages of this project.

This book was written in LaTeX, and the software used for computations and graphics was the high level matrix programming language GAUSS, distributed by Aptech Systems of Maple Valley, Washington. Graphs were produced using the free software *Asymptote*.

The Cochrane Collaboration logo in Chapter 3 is reproduced by permission of Cochrane Library.

Anders Källén
Lund, October 2010