Preface

Statistical analysis is at the core of most modern biology, and many biological hypotheses, even deceptively simple ones, are matched by complex statistical models. Prior to the development of modern desktop computers, determining whether the data fit these complex models was the province of professional statisticians. Many biologists instead opted for simpler models whose structure had been simplified quite arbitrarily. Now, with immensely powerful statistical software available to most of us, these complex models can be fitted, creating a new set of demands and problems for biologists.

We need to:

- know the pitfalls and assumptions of particular statistical models,
- be able to identify the type of model appropriate for the sampling design and kind of data that we plan to collect,
- be able to interpret the output of analyses using these models, and
- be able to design experiments and sampling programs optimally, i.e. with the best possible use of our limited time and resources.

The analysis may be done by professional statisticians, rather than statistically trained biologists, especially in large research groups or multidisciplinary teams. In these situations, we need to be able to speak a common language:

- frame our questions in such a way as to get a sensible answer,
- be aware of biological considerations that may cause statistical problems; we can not expect a statistician to be aware of the biological idiosyncrasies of our particular study, but if he or she lacks that information, we may get misleading or incorrect advice, and
- understand the advice or analyses that we receive, and be able to translate that back into biology.

This book aims to place biologists in a better position to do these things. It arose from our involvement in designing and analyzing our own data, but also providing advice to students and colleagues, and teaching classes in design and analysis. As part of these activities, we became aware, first of our limitations, prompting us to read more widely in the primary statistical literature, and second, and more importantly, of the complexity of the statistical models underlying much biological research. In particular, we continually encountered experimental designs that were not described comprehensively in many of our favorite texts. This book describes many of the common designs used in biological research, and we present the statistical models underlying those designs, with enough information to highlight their benefits and pitfalls.

Our emphasis here is on dealing with biological data – how to design sampling programs that represent the best use of our resources, how to avoid mistakes that make analyzing our data difficult, and how to analyze the data when they are collected. We emphasize the problems associated with real world biological situations.

In this book

Our approach is to encourage readers to understand the models underlying the most common experimental designs. We describe the models that are appropriate for various kinds of biological data – continuous and categorical response variables, continuous and categorical predictor or independent variables. Our emphasis is on general linear models, and we begin with the simplest situations – single, continuous variables – describing those models in detail. We use these models as building blocks to understanding a wide range of other kinds of data – all of the common statistical analyses, rather than being distinctly different kinds of analyses, are variations on a common theme of statistical modeling – constructing a model for the data and then determining whether observed data fit this particular model. Our aim is to show how a broad understanding of the models allows us to

deal with a wide range of more complex situations.

We have illustrated this approach of fitting models primarily with parametric statistics. Most biological data are still analyzed with linear models that assume underlying normal distributions. However, we introduce readers to a range of more general approaches, and stress that, once you understand the general modeling approach for normally distributed data, you can use that information to begin modeling data with nonlinear relationships, variables that follow other statistical distributions, etc.

Learning by example

One of our strongest beliefs is that we understand statistical principles much better when we see how they are applied to situations in our own discipline. Examples let us make the link between statistical models and formal statistical terms (blocks, plots, etc.) or papers written in other disciplines, and the biological situations that we are dealing with. For example, how is our analysis and interpretation of an experiment repeated several times helped by reading a literature about blocks of agricultural land? How does literature developed for psychological research let us deal with measuring changes in physiological responses of plants?

Throughout this book, we illustrate all of the statistical techniques with examples from the current biological literature. We describe why (we think) the authors chose to do an experiment in a particular way, and how to analyze the data, including assessing assumptions and interpreting statistical output. These examples appear as boxes through each chapter, and we are delighted that authors of most of these studies have made their raw data available to us. We provide those raw data files on a website http://www.zoology.unimelb.edu.au/qkstats allowing readers to run these analyses using their particular software package.

The other value of published examples is that we can see how particular analyses can be described and reported. When fitting complex statistical models, it is easy to allow the biology to be submerged by a mass of statistical output. We hope that the examples, together with our own thoughts on this subject, presented in the final chapter, will help prevent this happening.

This book is a bridge

It is not possible to produce a book that introduces a reader to biological statistics and takes them far enough to understand complex models, at least while having a book that is small enough to transport. We therefore assume that readers are familiar with basic statistical concepts, such as would result from a one or two semester introductory course, or have read one of the excellent basic texts (e.g. Sokal & Rohlf 1995). We take the reader from these texts into more complex areas, explaining the principles, assumptions, and pitfalls, and encourage a reader to read the excellent detailed treatments (e.g. for analysis of variance, Winer *et al.* 1991 or Underwood 1997).

Biological data are often messy, and many readers will find that their research questions require more complex models than we describe here. Ways of dealing with messy data or solutions to complex problems are often provided in the primary statistical literature. We try to point the way to key pieces of that statistical literature, providing the reader with the basic tools to be able to deal with that literature, or to be able to seek professional (statistical) help when things become too complex.

We must always remember that, for biologists, *statistics is a tool* that we use to illuminate and clarify biological problems. Our aim is to be able to use these tools efficiently, without losing sight of the biology that is the motivation for most of us entering this field.

Some acknowledgments

Our biggest debt is to the range of colleagues who have read, commented upon, and corrected various versions of these chapters. Many of these colleagues have their own research groups, who they enlisted in this exercise. These altruistic and diligent souls include (alphabetically) Jacqui

Brooks, Andrew Constable, Barb Downes, Peter Fairweather, Ivor Growns, Murray Logan, Ralph Mac Nally, Richard Marchant, Pete Raimondi, Wayne Robinson, Suvaluck Satumanatpan and Sabine Schreiber. Perhaps the most innocent victims were the graduate students who have been part of our research groups over the period we produced this book. We greatly appreciate their willingness to trade the chance of some illumination for reading and highlighting our obfuscations.

We also wish to thank the various researchers whose data we used as examples throughout. Most of them willingly gave of their raw data, trusting that we would neither criticize nor find flaws in their published work (we didn't!), or were public-spirited enough to have published their raw data.