Preface

Statistical design is one of the fundamentals of our subject, being at the core of the growth of statistics during the previous century. Design played a key role in agricultural statistics and set down principles of good practice, principles that still apply today. Statistical design is about understanding where the variance comes from, and making sure that is where the replication is. Indeed, it is probably correct to say that these principles are even more important today. Fisher (1947) compared a dataset to a sample of gold ore. The finest analysis could only extract the proportion of gold contained in the ore. But a good design could produce a sample with more gold.

There are plenty of "Design of Experiments" books available, many of which do a fine job of describing not only how to design experiments, but also how to analyze them. So why bother with another book? There were two main reasons.

The first reason is the observation that many of our "standard" analyses have become driven by the default setting of one's favorite computer package and, unfortunately, many times these default settings provide an incorrect analysis. More frightening is the fact that sometimes such default analyses have found their way into textbooks.

The second reason is that, although design books have gotten broader in coverage of designs and often have much to say about analysis, the basic theoretical underpinnings are not always covered, and if they are, they are not covered in sufficient detail to understand how to construct the correct analysis and to understand why the computer package default analysis may be incorrect. Without fully understanding what the correct analysis *should be*, it is impossible to design a good experiment.

So ... the goal is to describe the principles that drive good design, which are also the principles that drive good statistics. Moreover, this will be done with detail and attention paid to the theoretical background – only by having more than a passing familiarity with the fundamental theory can one truly understand statistical design.

This book is not an encyclopedia of designs. There is no attempt to cover all designs, and no attempt to teach data analysis at all. Although we will analyze many datasets, we will usually use common anova techniques and not preach about transformations, heterogeneity, missing data, and all the other good stuff that a good data analyst needs to know. We assume that the design has been run with enough success so that rescue techniques are not needed for the analysis. We also assume that the student has been exposed to such data analysis strategies.

We will cover the most popular designs in depth, both with theory and examples, and datasets from consulting sessions and research publications. We emphasize basic principles and careful modeling and, armed with such tools, the student should be able to apply these principles in any situation

This book grew out of a course on Statistical Design taught at Cornell during the 1990s, and at Florida in the 2000s. Most of the examples and datasets are from consulting sessions with graduate students writing theses or professors writing papers, and the subjects span everything from planting alfalfa in the field to harvesting brain stem cells. (Most of my career has been spent at Colleges of Agriculture – Rutgers, Cornell, and split between Arts and Sciences and Agriculture at Florida. This will show in the examples.) I have found that although the data and the lab techniques have changed, the statistical principles remain quite constant.

Now, for the more important details:

○ The level of the text is for first or second year graduate students. The students should be familiar with standard statistical methodology (anova, blocking, multiple regression) that one would get from a typical one-year methods sequence (from books such as Ott and Longnecker 2000, Rawlings *et al.* 1998).

○ The material in the text is about right for a one-semester course. There is probably a bit more than can be comfortably covered, so some picking and choosing will be necessary. However, marching through the text is a reasonable strategy.

○ The chapters cover, for the most part, the standard material of a design book, with mostly real examples, and applications of design in real situations. Although we cover many microarray designs, we do not have a special section, instead treating them as the topics apply. The only unusual chapter is Chapter 4, where the concept of blocking is explored further, and the effect of a random factor is examined. To me, blocks are not about being samples from a larger population (which can be difficult to justify) but rather about the correlation structure that they induce.

○ Most chapters have a section *Technical Notes*, which contains the underlying theory in detail. The level in these notes is "anything goes"; we use a lot of matrix algebra, some calculus, and also some statistical concepts such as likelihood and sufficiency. These are in-depth looks that will enhance the understanding of the advanced student, but skipping these

sections will not hamper the beginner. To fully appreciate these sections it would be good if the student has had a course from a book like *Statistical Inference* by Casella and Berger (2001).

○ The exercises are divided into "Essential" and "Accompaniment". Everyone should do the essentials. The accompaniments tend to be more of a theoretical nature, going into the details of the procedures. I strongly suggest that, if the students are able, they should do these exercises too. I have always found that, for me, true understanding only comes from slogging through the details.

○ The datasets are on the book web page found at

www.stat.ufl.edu/∼casella

and, for most datasets in the examples there will be an accompanying R program. These will not be sophisticated analyses, but rather will serve as a starting point. I make no guarantees about the R programs. Smaller datasets may only appear in text.

My best advice is that if you really want to understand statistical design, read (or even better, reread) Fisher. His ideas, especially about blocking, have greatly influenced my thinking. In fact, what I perceived as mishandling of the randomized complete block design was one of the driving forces behind the text. Fisher, of course, got it right.

There are many gray areas in design – when to pool, how to replicate, etc. – some of which cannot be fully answered with statistical fact. This is where we enter the realm of opinions, where judgments are made more on anecdotal evidence and experience rather than formal calculations. After doing this stuff for over 20 years opinions form about how to do things – I have taken the liberty of sharing those thoughts.

Finally, thanks to all of my mentors. First there was Leon Gleser, my PhD advisor, who taught me to work hard and try to learn as much as possible. My design mentors were many, starting with Virgil Anderson at Purdue, and Walt Federer at Cornell, who not only made me really understand split-plot designs, but also taught me to say "a model" and *never* "the model". And Carl Lowe, the Plant Breeding Professor at Cornell who taught the course "Field Plot Techniques", and knew more about field plot layouts than anyone on the planet. When he retired he gave me all of his notes and examples, many of which appear as examples and exercises (and in the Appendix). I also thank the students and colleagues who suffered through my learning process, who listened as I thought through all of this, read through notes, and solved problems: Mihai Giurcanu, Jamie Jarabek, David Lansky, Michael Meredith, Deborah Reichert, and Andy Scherrer.

My most special mentor was Myra Samuels. A true sadness in my life is that she died in 1992. She taught me to always ask questions until you really understand. Thanks, Myra.

And to all of my other mentors – you have all helped me more than you know – Bob Bechhoffer, Jim Berger, Larry Brown, Shanti Gupta, Jean-Pierre Habicht, George McCabe, Doug Robson, Bill Strawderman. And all of the others that I have learned from. Thanks.

George Casella
Gainesville, Florida

June 6, 2008