

# Preface

This book is about the “information-theoretic” approaches to rigorous inference based on Kullback–Leibler information. My objective in writing this book is to provide an introduction to making rigorous statistical inferences from data and models about hypotheses in the life sciences. The goal of this primer is to explain the information-theoretic approaches with a focus on application. I stress science philosophy as much as statistical theory and I wade into some ideas from information theory because it is so interesting. The book is about hypothesizing science alternatives and providing quantitative evidence for these.

In 1973 Hirotugu Akaike made a world class discovery when he found a linkage between K–L information and statistical theory through the maximized log-likelihood function. Statistical theory developed since the mid-1970s allows science hypotheses, represented by mathematical models, to be ranked from best to worst. In addition, the discrete probability of each hypothesis  $i$ , given the data, can be easily computed. These can be viewed as Bayesian posterior model probabilities and are quite important in making inferences about the science hypotheses of interest. The likelihood of each hypothesis, given the data, and evidence ratios between hypotheses  $i$  and  $j$  are also available, and easy to interpret. All of these new features are simple to compute and understand and go far beyond traditional methods. While many of the examples are biological, I hope students and scientists in other fields (e.g., social sciences, medicine, economics, and many other disciplines) can learn from this primer. Several examples are ecological as that has been my interest; however, the specific examples used are far less important than the science context and trying to understand new approaches; I could not include an example from all of the many subdisciplines.

Tosio Kitagawa (1986) noted that the information-theoretic methods are

*“... a challenge to conventional statistics as well as a proposal for a new approach to statistical analysis. The reader may find some aspects of the approach controversial insofar as they imply a criticism of conventional mathematical statistics, such as the use of statistical tests, individual sampling distribution theory, and statistical tables.”*

I find that some people are still struggling with these new approaches 20 years later. Perhaps this reticence is healthy for science as new ideas must be carefully evaluated and scrutinized; however, we must not let “progress ride on a hearse” either.

I have tried to write this as a science textbook; in a sense it is a companion to the books I have written on this subject with Ken Burnham in 1998 and 2002. Those books contain statistical theory, derivations, proofs, some comparisons with other approaches, and were written at a more advanced level. The present primer tries to be well above a “cookbook” but well below a highly technical treatment; this is a book largely for people new to these effective approaches to empirical science. The book provides a consistent strategy (the concepts of evidence and evolving hypothesis sets) for rapid learning and a way of thinking about science and discovery; a road map of sorts. I provide several examples and many insights on modeling; however, I must say clearly that this is not a primer on modeling.

In the first 4 chapters I cover some material to motivate thinking about plausible science hypotheses (the most important issue), data, information, K–L information, and various measures of evidence and support for members of a set of science hypotheses and their corresponding models. Several examples continue through the chapters as new developments are introduced. At this point, the basics of model based inference under the “information-theoretic” approach will have been laid out. But then, like many good novels – there is a twist. Instead of trying to identify the best science hypothesis (and its model) from the set of hypotheses, I refocus on making formal inference based on all the models – “multimodel inference.” In many cases it is desirable to make predictions from all the hypotheses in an *a priori* set – one facet of multimodel inference. These procedures allow model averaging and unconditional measures of precision. Those people thinking this jump will surely be difficult will be pleasantly surprised. The main approaches to multimodel inference under this approach can be understood in 1–2 h of lecture and discussion – they are relatively simple but effective. I hope readers will conceptualize their empirical work in science as multimodel inference. This mental image will help focus on the importance of deriving a set of good, plausible science hypotheses (the hard thinking), gathering quality data, and using modern methods to provide quantitative evidence for each of the science hypotheses of interest.

I want to be quick to say that there are other valid approaches to making inferences from empirical data and I make no effort to deny these. There are general theories related to cross validation, nonparametric statistics, bootstrapping, and

Bayesian approaches to mention only a few. In addition, there are a number of new theories for model selection for linear models; I have omitted reference to these special cases but admit that, with further development, they may someday have wider application. Of the four general theories I noted, only the Bayesian approaches have the breadth and depth of those based on information theory. All have their strengths and I encourage some understanding of these approaches. I will make passing reference to some of these alternatives. I am pro-Bayesian and am interested in areas of commonality between the information-theoretic methods and Bayesian methods. Frequentists and Bayesians have waged a long and protracted philosophical war; I do not want to see the information-theoretic approaches join the conflict.

I consider the various null hypothesis testing approaches to be only of historical interest at this stage (2007), except perhaps in the analysis of data from strict experiments where the design stipulates a single model (i.e., design based inference). In general I think scientists serious about their work must move beyond testing sterile null hypotheses to modern methods and the substantial advantages they provide. I offer several comparisons.

This primer is written to be useful for seniors in excellent undergraduate science programs at top universities. Perhaps more realistically, the book is aimed at graduate students, post-doctoral fellows, as well as established scientists in academia, government agencies, and various science institutes. A basic statistical background is essential to easily understand the material in this book: sampling theory, simple experimental designs, measures of variability and covariability (e.g., sampling variances and covariances, standard errors, coefficients of variation, various approaches to confidence intervals, and sampling correlations), “regression” (e.g.,  $\beta_i$  as partial regression coefficients,  $R^2$ , residual variance  $\sigma^2$ , residual sum of squares RSS), and goodness-of-fit concepts.

Ideally, the reader would have had some introduction to Fisher’s likelihood approaches (e.g., maximum likelihood estimates, profile likelihood intervals). It is hard to understand why there is so much emphasis on least squares approaches even in graduate courses for nonstatistics majors as this narrow approach comes at the expense of the much more general and useful likelihood methods. In addition, likelihood is foundational to the Bayesian approaches. Readers with the required background will find the quantitative issues easy; it is the deeper conceptual issues that will challenge nearly everyone (e.g., model selection bias). This is the fun and rewarding part of science – thinking hard. Readers lacking exposure to null hypothesis testing will find the material here easier to understand than their counterparts. Still, readers should expect to have to reread some material and contemplate the examples given to chase a full understanding of the material.

A *Remarks* section is found near the end of most chapters and some people will find these unordered comments interesting; however, I suggest this material might best be saved for a second reading. This material includes historical comments, technical notes, and other tangential issues that I thought might

interest many readers. In a sense, the *Remarks* are a grouping of what would otherwise be “footnotes,” which I often find interesting, but sometimes distracting from the main points. Most chapters end with some nontraditional exercises. Comments on answers to some of the exercises can be found at [www.springer.com/978-0-387-74073-7](http://www.springer.com/978-0-387-74073-7) Each chapter includes a photo and short biography of people who have made major contributions to this literature. I think it is important to recognize and learn from people who came before us and made substantial contributions to science.

This is not an introductory text as I assume a basic knowledge of statistics, the ability to conceptualize science hypotheses ( $H_i$ ), represent these by mathematical models ( $g_i$ ), obtain estimates of model parameters ( $\theta$ ), their sampling covariance matrix ( $\Sigma$ ), goodness-of-fit tests, and residual analysis. Given this backdrop, new and deeper questions can be asked and answers quantified effectively and simply. I believe this material is fun and exciting if you are a scientist who is serious about advanced work on problems where there are substantial stochasticities and complexities. The material is very broad, but I say less about models for multivariate responses and random effects (as opposed to so-called fixed effects) models.

Many people in the life sciences leave graduate programs with little or no exposure to quantitative thinking and methods and this is an increasingly serious issue, limiting both their contributions to science and their career growth. Many PhD-level people lack any working knowledge of calculus, statistics, matrix algebra, computer programming, numerical methods, and modeling. Some people think that is why they are in biology – “because then I don’t have to learn that quantitative stuff.” I can certainly understand the sentiment; however, there are ample reasons to reconsider, even later in life. Quantification becomes essential in real world problems as a science matures in a given discipline.

In a sense, undergraduate students are taught a small fraction of material that is *already known* in their field and associated disciplines. People need this background information. Graduate work is quite different (or should be), as students are taught effective philosophies and methods to help them learn how to understand things *new* to their field of science. First one wants to know the current “edge” of knowledge on some issue. Second, one wants to push that edge further as new things are learned from the science process. These are things that cannot be found in a book or on the Internet; the discovery of *new* things – this is what science is all about. Good undergraduate programs try to blur the line between these extremes and this is healthy for science. Graduate programs try to help students shift gears into considering methodologies and philosophies for rapid learning of new things; things that no one has discovered (yet). These are often the harder, more complex issues as our predecessors have solved the easier problems. The information-theoretic approaches represent an effective science strategy and allow one to shift into 6th or even 7th gear and that is what makes learning this material both important and fun.

I wanted to write a short book and try to make some main points that I think are important to people coming to these subjects for the first time. This is a

book about doing empirical science. I do not expect everyone to agree with every philosophical or analytical aspect. Few of the ideas are originally mine as I have taken from thoughts and results from many others as I try to synthesize the more fundamental issues for the reader. This synthesis comes from about 40 years of experience, studying the work of others and trying to form a coherent philosophy about an effective way to do empirical science. I hope people will take what they find useful and be willing to forge ahead in areas where they have found better approaches. I intend to remain interested in this broad subject and will always enjoy hearing comments from colleagues, many of whom I have not yet met.

I want to fully acknowledge my closest friend and colleague over the last 34 years, Ken Burnham. I (reluctantly) wrote this text alone as I am trusting Ken to complete his book on experimental design. Ken has had an powerful influence on my thinking about science philosophy, statistics, information theory, and model based inference. Several other people helped in various ways and I am proud to acknowledge Peter Caley and Jim Hone for their help with my use of their ferret data and Lianne Ball and Paul Doherty for their help with the Palm Springs ground squirrel example. I benefited greatly from extensive review comments offered by Peter Beerli, Barry Grand, Benedikt Schmidt, and Bill Thompson. I also want to thank Bill Gould, Paul Lukacs, Dave Otis, and Eric Stolen for their advice and encouragement. The photo of Thomas Chamberlin was provided by the Edgar Fahs Smith collection at the University of Pennsylvania. John Kimmel at Springer was both patient and encouraging as he helped me through the writing and publishing process.

Fort Collins, CO

David R. Anderson  
June, 2007



<http://www.springer.com/978-0-387-74073-7>

Model Based Inference in the Life Sciences

A Primer on Evidence

Anderson, D.R.

2008, XXIV, 184 p. 8 illus., Softcover

ISBN: 978-0-387-74073-7