TURING 图灵原版数学·统计学系列 36

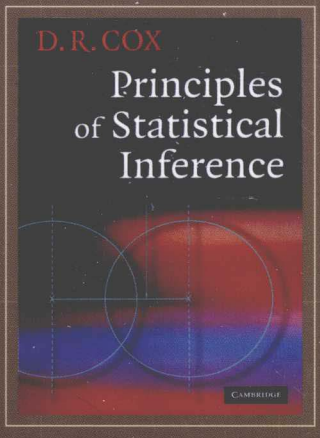
CAMBRIDGE

Principles of Statistical Inference

统计推断原理

（英文版）

[英] D. R. Cox 著



人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原版数学·统计学系列

Principles of Statistical Inference

统计推断原理

（英文版）

[英] D. R. Cox 著

人民邮电出版社
北　京

图书在版编目（CIP）数据

统计推断原理＝Principles of Statistical Inference：英文/（英）考克斯（Cox, D. R.）著. —北京：人民邮电出版社，2009.8

（图灵原版数学·统计学系列）

ISBN 978-7-115-21074-6

I. 统… II. 考… III. 统计推断－英文 IV. O212

中国版本图书馆CIP数据核字（2009）第100530号

内 容 提 要

本书是统计学名家名作，包含9章内容和两个附录，前面几章介绍一些基本概念，如参数、似然、主元等，然后介绍显著性检验、渐进理论以及比较复杂的统计推断问题．还特别介绍了实验设计中基于随机化的统计推断．核心概念的解释非常清晰，即使跳过其中的数学细节，也能使读者理解．

本书可作为工科、管理类学科专业本科生、研究生的教材或参考书，也可供教师、工程技术人员自学之用．

图灵原版数学·统计学系列

统计推断原理（英文版）

◆ 著 [英] D. R. Cox

责任编辑 明永玲

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号

邮编 100061 电子函件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京铭成印刷有限公司印刷

◆ 开本：700×1000 1/16

印张：14.5

字数：232千字 2009年8月第1版

印数：1-2 000册 2009年8月北京第1次印刷

著作权合同登记号 图字：01-2009-2915号

ISBN 978-7-115-21074-6/O1

定价：49.00元

读者服务热线：（010）51095186 印装质量热线：（010）67129223

反盗版热线：（010）67171154

版 权 声 明

Preface

Most statistical work is concerned directly with the provision and implementation of methods for study design and for the analysis and interpretation of data. The theory of statistics deals in principle with the general concepts underlying all aspects of such work and from this perspective the formal theory of statistical inference is but a part of that full theory. Indeed, from the viewpoint of individual applications, it may seem rather a small part. Concern is likely to be more concentrated on whether models have been reasonably formulated to address the most fruitful questions, on whether the data are subject to unappreciated errors or contamination and, especially, on the subject-matter interpretation of the analysis and its relation with other knowledge of the field.

Yet the formal theory is important for a number of reasons. Without some systematic structure statistical methods for the analysis of data become a collection of tricks that are hard to assimilate and interrelate to one another, or for that matter to teach. The development of new methods appropriate for new problems would become entirely a matter of ad hoc ingenuity. Of course such ingenuity is not to be undervalued and indeed one role of theory is to assimilate, generalize and perhaps modify and improve the fruits of such ingenuity.

Much of the theory is concerned with indicating the uncertainty involved in the conclusions of statistical analyses, and with assessing the relative merits of different methods of analysis, and it is important even at a very applied level to have some understanding of the strengths and limitations of such discussions. This is connected with somewhat more philosophical issues connected with the nature of probability. A final reason, and a very good one, for study of the theory is that it is interesting.

The object of the present book is to set out as compactly as possible the key ideas of the subject, in particular aiming to describe and compare the main ideas and controversies over more foundational issues that have rumbled on at varying levels of intensity for more than 200 years. I have tried to describe the

various approaches in a dispassionate way but have added an appendix with a more personal assessment of the merits of different ideas.

Some previous knowledge of statistics is assumed and preferably some understanding of the role of statistical methods in applications; the latter understanding is important because many of the considerations involved are essentially conceptual rather than mathematical and relevant experience is necessary to appreciate what is involved.

The mathematical level has been kept as elementary as is feasible and is mostly that, for example, of a university undergraduate education in mathematics or, for example, physics or engineering or one of the more quantitative biological sciences. Further, as I think is appropriate for an introductory discussion of an essentially applied field, the mathematical style used here eschews specification of regularity conditions and theorem–proof style developments. Readers primarily interested in the qualitative concepts rather than their development should not spend too long on the more mathematical parts of the book.

The discussion is implicitly strongly motivated by the demands of applications, and indeed it can be claimed that virtually everything in the book has fruitful application somewhere across the many fields of study to which statistical ideas are applied. Nevertheless I have not included specific illustrations. This is partly to keep the book reasonably short, but, more importantly, to focus the discussion on general concepts without the distracting detail of specific applications, details which, however, are likely to be crucial for any kind of realism.

The subject has an enormous literature and to avoid overburdening the reader I have given, by notes at the end of each chapter, only a limited number of key references based on an admittedly selective judgement. Some of the references are intended to give an introduction to recent work whereas others point towards the history of a theme; sometimes early papers remain a useful introduction to a topic, especially to those that have become suffocated with detail. A brief historical perspective is given as an appendix.

The book is a much expanded version of lectures given to doctoral students of the Institute of Mathematics, Chalmers/Gothenburg University, and I am very grateful to Peter Jagers and Nanny Wermuth for their invitation and encouragement. It is a pleasure to thank Ruth Keogh, Nancy Reid and Rolf Sundberg for their very thoughtful detailed and constructive comments and advice on a preliminary version. It is a pleasure to thank also Anthony Edwards and Deborah Mayo for advice on more specific points. I am solely responsible for errors of fact and judgement that remain.

The book is in broadly three parts. The first three chapters are largely introductory, setting out the formulation of problems, outlining in a simple case the nature of frequentist and Bayesian analyses, and describing some special models of theoretical and practical importance. The discussion continues with the key ideas of likelihood, sufficiency and exponential families.

Chapter 4 develops some slightly more complicated applications. The long Chapter 5 is more conceptual, dealing, in particular, with the various meanings of probability as it is used in discussions of statistical inference. Most of the key concepts are in these chapters; the remaining chapters, especially Chapters 7 and 8, are more specialized.

Especially in the frequentist approach, many problems of realistic complexity require approximate methods based on asymptotic theory for their resolution and Chapter 6 sets out the main ideas. Chapters 7 and 8 discuss various complications and developments that are needed from time to time in applications. Chapter 9 deals with something almost completely different, the possibility of inference based not on a probability model for the data but rather on randomization used in the design of the experiment or sampling procedure.

I have written and talked about these issues for more years than it is comfortable to recall and am grateful to all with whom I have discussed the topics, especially, perhaps, to those with whom I disagree. I am grateful particularly to David Hinkley with whom I wrote an account of the subject 30 years ago. The emphasis in the present book is less on detail and more on concepts but the eclectic position of the earlier book has been kept.

I appreciate greatly the care devoted to this book by Diana Gillooly, Commissioning Editor, and Emma Pearce, Production Editor, Cambridge University Press.

List of examples

Contents

1
Preliminaries

Summary. Key ideas about probability models and the objectives of statistical analysis are introduced. The differences between frequentist and Bayesian analyses are illustrated in a very special case. Some slightly more complicated models are introduced as reference points for the following discussion.

1.1 Starting point

We typically start with a subject-matter question. Data are or become available to address this question. After preliminary screening, checks of data quality and simple tabulations and graphs, more formal analysis starts with a provisional model. The data are typically split in two parts $(y : z)$, where y is regarded as the observed value of a vector random variable Y and z is treated as fixed. Sometimes the components of y are direct measurements of relevant properties on study individuals and sometimes they are themselves the outcome of some preliminary analysis, such as means, measures of variability, regression coefficients and so on. The set of variables z typically specifies aspects of the system under study that are best treated as purely explanatory and whose observed values are not usefully represented by random variables. That is, we are interested solely in the distribution of outcome or response variables conditionally on the variables z; a particular example is where z represents treatments in a randomized experiment.

We use throughout the notation that observable random variables are represented by capital letters and observations by the corresponding lower case letters.

A model, or strictly a family of models, specifies the density of Y to be

$$f_Y(y : z; \theta), \tag{1.1}$$

where $\theta \subset \Omega_\theta$ is unknown. The distribution may depend also on design features of the study that generated the data. We typically simplify the notation to $f_Y(y;\theta)$, although the explanatory variables z are frequently essential in specific applications.

To choose the model appropriately is crucial to fruitful application.

We follow the very convenient, although deplorable, practice of using the term *density* both for continuous random variables and for the probability function of discrete random variables. The deplorability comes from the functions being dimensionally different, probabilities per unit of measurement in continuous problems and pure numbers in discrete problems. In line with this convention in what follows integrals are to be interpreted as sums where necessary. Thus we write

$$E(Y) = E(Y;\theta) = \int y f_Y(y;\theta) dy \tag{1.2}$$

for the expectation of Y, showing the dependence on θ only when relevant. The integral is interpreted as a sum over the points of support in a purely discrete case. Next, for each aspect of the research question we partition θ as (ψ, λ), where ψ is called the *parameter of interest* and λ is included to complete the specification and commonly called a *nuisance parameter*. Usually, but not necessarily, ψ and λ are *variation independent* in that Ω_θ is the Cartesian product $\Omega_\psi \times \Omega_\lambda$. That is, any value of ψ may occur in connection with any value of λ. The choice of ψ is a subject-matter question. In many applications it is best to arrange that ψ is a scalar parameter, i.e., to break the research question of interest into simple components corresponding to strongly focused and incisive research questions, but this is not necessary for the theoretical discussion.

It is often helpful to distinguish between the primary features of a model and the secondary features. If the former are changed the research questions of interest have either been changed or at least formulated in an importantly different way, whereas if the secondary features are changed the research questions are essentially unaltered. This does not mean that the secondary features are unimportant but rather that their influence is typically on the method of estimation to be used and on the assessment of precision, whereas misformulation of the primary features leads to the wrong question being addressed.

We concentrate on problems where Ω_θ is a subset of R^d, i.e., d-dimensional real space. These are so-called *fully parametric* problems. Other possibilities are to have semiparametric problems or fully nonparametric problems. These typically involve fewer assumptions of structure and distributional form but usually contain strong assumptions about independencies. To an appreciable

extent the formal theory of semiparametric models aims to parallel that of parametric models.

The probability model and the choice of ψ serve to translate a subject-matter question into a mathematical and statistical one and clearly the faithfulness of the translation is crucial. To check on the appropriateness of a new type of model to represent a data-generating process it is sometimes helpful to consider how the model could be used to generate synthetic data. This is especially the case for stochastic process models. Understanding of new or unfamiliar models can be obtained both by mathematical analysis and by simulation, exploiting the power of modern computational techniques to assess the kind of data generated by a specific kind of model.

1.2 Role of formal theory of inference

The formal theory of inference initially takes the family of models as given and the objective as being to answer questions about the model in the light of the data. Choice of the family of models is, as already remarked, obviously crucial but outside the scope of the present discussion. More than one choice may be needed to answer different questions.

A second and complementary phase of the theory concerns what is sometimes called *model criticism*, addressing whether the data suggest minor or major modification of the model or in extreme cases whether the whole focus of the analysis should be changed. While model criticism is often done rather informally in practice, it is important for any formal theory of inference that it embraces the issues involved in such checking.

1.3 Some simple models

General notation is often not best suited to special cases and so we use more conventional notation where appropriate.

Example 1.1. *The normal mean.* Whenever it is required to illustrate some point in simplest form it is almost inevitable to return to the most hackneyed of examples, which is therefore given first. Suppose that $Y_1, \ldots, Y_n$ are independently normally distributed with unknown mean μ and known variance σ_0^2. Here μ plays the role of the unknown parameter θ in the general formulation. In one of many possible generalizations, the variance σ^2 also is unknown. The parameter vector is then (μ, σ^2). The component of interest ψ would often be μ

but could be, for example, σ^2 or μ/σ, depending on the focus of subject-matter interest.

Example 1.2. *Linear regression.* Here the data are n pairs $(y_1, z_1), \ldots, (y_n, z_n)$ and the model is that $Y_1, \ldots, Y_n$ are independently normally distributed with variance σ^2 and with

$$E(Y_k) = \alpha + \beta z_k. \tag{1.3}$$

Here typically, but not necessarily, the parameter of interest is $\psi = \beta$ and the nuisance parameter is $\lambda = (\alpha, \sigma^2)$. Other possible parameters of interest include the intercept at $z = 0$, namely α, and $-\alpha/\beta$, the intercept of the regression line on the z-axis.

Example 1.3. *Linear regression in semiparametric form.* In Example 1.2 replace the assumption of normality by an assumption that the Y_k are uncorrelated with constant variance. This is semiparametric in that the systematic part of the variation, the linear dependence on z_k, is specified parametrically and the random part is specified only via its covariance matrix, leaving the functional form of its distribution open. A complementary form would leave the systematic part of the variation a largely arbitrary function and specify the distribution of error parametrically, possibly of the same normal form as in Example 1.2. This would lead to a discussion of smoothing techniques.

Example 1.4. *Linear model.* We have an $n \times 1$ vector Y and an $n \times q$ matrix z of fixed constants such that

$$E(Y) = z\beta, \quad \text{cov}(Y) = \sigma^2 I, \tag{1.4}$$

where β is a $q \times 1$ vector of unknown parameters, I is the $n \times n$ identity matrix and with, in the analogue of Example 1.2, the components independently normally distributed. Here z is, in initial discussion at least, assumed of full rank $q < n$. A relatively simple but important generalization has $\text{cov}(Y) = \sigma^2 V$, where V is a given positive definite matrix. There is a corresponding semiparametric version generalizing Example 1.3.

Both Examples 1.1 and 1.2 are special cases, in the former the matrix z consisting of a column of 1s.

Example 1.5. *Normal-theory nonlinear regression.* Of the many generalizations of Examples 1.2 and 1.4, one important possibility is that the dependence on the parameters specifying the systematic part of the structure is nonlinear. For example, instead of the linear regression of Example 1.2 we might wish to consider

$$E(Y_k) = \alpha + \beta \exp(\gamma z_k), \tag{1.5}$$