Springer Series in Statistics

David Siegmund

Sequential Analysis

Tests and Confidence Intervals

Springer-Verlag
New York Berlin Heidelberg Tokyo

David Siegmund

Sequential Analysis

Tests and Confidence Intervals

With 13 Illustrations

Springer-Verlag
New York Berlin Heidelberg Tokyo

David Siegmund
Department of Statistics
Stanford University
Stanford, CA 94305
U.S.A.

AMS Classification: 62L10

Library of Congress Cataloging in Publication Data
Siegmund, David
Sequential analysis.
(Springer series in statistics)
Bibliography: p.
Includes index.
1. Sequential analysis. 2. Statistical hypothesis testing.
3. Confidence intervals. I. Title. II. Series.
QA279.7.S54 1985 519.2 85-7942

Typeset by Asco Trade Typesetting Ltd., Hong Kong.
Printed and bound by R. R. Donnelley and Sons, Harrisonburg, Virginia.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-96134-8 Springer-Verlag New York Berlin Heidelberg Tokyo
ISBN 3-540-96134-8 Springer-Verlag Berlin Heidelberg New York Tokyo

Springer Series in Statistics

Springer Series in Statistics

D. F. Andrews and A. M. Herzberg, Data: A Collection of Problems from Many Fields for the Student and Research Worker. xx, 442 pages, 1985.

F. J. Anscombe, Computing in Statistical Science through APL. xvi, 426 pages, 1981.

J. O. Berger, Statistical Decision Theory: Foundations, Concepts, and Methods, 2nd edition. xiv, 425 pages, 1985.

P. Brémaud, Point Processes and Queues: Martingale Dynamics. xviii, 354 pages, 1981.

K. Dzhaparidze, Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series. xii, 300 pages, 1985.

R. H. Farrell, Multivariate Calculation. xvi, 367 pages, 1985.

L. A. Goodman and W. H. Kruskal, Measures of Association for Cross Classifications. x, 146 pages, 1979.

J. A. Hartigan, Bayes Theory. xii, 145 pages, 1983.

H. Heyer, Theory of Statistical Experiments. x, 289 pages, 1982.

H. Kres, Statistical Tables for Multivariate Analysis. xxii, 504 pages, 1983.

M. R. Leadbetter, G. Lindgren and H. Rootzén, Extremes and Related Properties of Random Sequences and Processes. xii, 336 pages, 1983.

R. G. Miller, Jr., Simultaneous Statistical Inference, 2nd edition. xvi, 299 pages, 1981.

F. Mosteller, D. S. Wallace, Applied Bayesian and Classical Inference: The Case of *The Federalist* Papers. xxxv, 301 pages, 1984.

D. Pollard, Convergence of Stochastic Processes. xiv, 215 pages, 1984.

J. W. Pratt and J. D. Gibbons, Concepts of Nonparametric Theory. xvi, 462 pages, 1981.

L. Sachs, Applied Statistics: A Handbook of Techniques. xxviii, 706 pages, 1982.

E. Seneta, Non-Negative Matrices and Markov Chains. xv, 279 pages, 1981.

D. Siegmund, Sequential Analysis: Tests and Confidence Intervals. xii, 272 pages, 1985.

V. Vapnik, Estimation of Dependences based on Empirical Data. xvi, 399 pages, 1982.

K. M. Wolter, Introduction to Variance Estimation. xii, 428 pages, 1985.

Preface

The modern theory of Sequential Analysis came into existence simultaneously in the United States and Great Britain in response to demands for more efficient sampling inspection procedures during World War II. The developments were admirably summarized by their principal architect, A. Wald, in his book *Sequential Analysis* (1947).

In spite of the extraordinary accomplishments of this period, there remained some dissatisfaction with the sequential probability ratio test and Wald's analysis of it. (i) The open-ended continuation region with the concomitant possibility of taking an arbitrarily large number of observations seems intolerable in practice. (ii) Wald's elegant approximations based on "neglecting the excess" of the log likelihood ratio over the stopping boundaries are not especially accurate and do not allow one to study the effect of taking observations in groups rather than one at a time. (iii) The beautiful optimality property of the sequential probability ratio test applies only to the artificial problem of testing a simple hypothesis against a simple alternative.

In response to these issues and to new motivation from the direction of controlled clinical trials numerous modifications of the sequential probability ratio test were proposed and their properties studied—often by simulation or lengthy numerical computation. (A notable exception is Anderson, 1960; see III.7.) In the past decade it has become possible to give a more complete theoretical analysis of many of the proposals and hence to understand them better.

The primary goal of this book is to review these recent developments for the most part in the specific framework of Wald's book, i.e., sequential hypothesis testing in a non-Bayesian, non-decision-theoretic context. In contrast to the sequential probability ratio test, the emphasis is on closed (truncated) sequential tests defined by non-linear stopping boundaries and often applied

to grouped data. In particular the repeated significance tests of Armitage (1975) and the group repeated significance tests of Pocock (1977) are given an extensive theoretical treatment. To the extent that there is a unifying theme to the book, it is an attempt to understand repeated significance tests theoretically, to criticize them, and to suggest some improvements as a response to this criticism.

A secondary goal is to redress to some extent the imbalance in the literature of sequential analysis between considerations of experimental design and considerations of statistical inference. Here the term "experimental design" refers to selection of an experiment, hence to selection of the data to be observed, from a set of possibilities. Since choosing a sequential test includes the problem of choosing a stopping rule, it should properly be considered an aspect of experimental design.

Inferential summaries of the data actually obtained in an experiment include attained significance levels (p-values) and confidence intervals, which play important roles in fixed sample statistics, but until recently have been almost totally ignored in discussions of sequential analysis. Chapters III and IV examine these concepts, which turn out to have certain implications for the selection of a stopping rule, hence for the experimental design itself.

Some additional subjects which can be studied by the methods developed to investigate truncated sequential tests are discussed briefly. Thus cusum tests are introduced in Chapter II and a simple approximation for their average run length is given in Chapter X. However, they are not systematically compared with other competing possibilities for detecting a change of distribution. Fixed precision confidence intervals receive a similar cursory treatment in Chapter VII.

I have attempted to make this book more widely accessible than most of the literature on which it is based by delaying as long as possible the introduction of "heavy" mathematics. Chapters I–VII emphasize the statistical ideas accompanied by enough of the mathematical development that the reader can regard himself as a participant rather than a spectator observing the opening of a black box. The proofs of many results are delayed until after their significance and application have been discussed; in fact most proofs fall short of the currently accepted mathematical standard. Woodroofe (1982) has given a more precise and extensive development of the mathematical theory.

The requirements of exposition have led to numerous compromises. The most significant is the use of Brownian motion in Chapters I–VI to provide basic qualitative insight. The corresponding discrete time results are stated so that they can be used for numerical calculations, but their generally more difficult justification is then deferred. Consequently the mathematical methods have been chosen to a certain extent because they provide a unified treatment for the range of problems considered, linear and non-linear in discrete and continuous time. In many cases this meant not using what appears to be the "best" method for a particular problem. For example, the methods pioneered by Woodroofe, which often seem to deliver the best result in discrete time,

have not been developed completely; and the purely continuous time methods of Jennen and Lerche (1981, 1982) are not discussed at all.

Finally it should be noted that the book is primarily concerned with very simple models—especially those involving the normal distribution. There exists an extensive literature devoted to approximating more complex models by simple ones. A notable early contribution is due to Cox (1963). See also Hall and Loynes (1977) and Sen (1982). Digressions illustrating the nature of such approximations appear in III.9 and again in V.5, which is concerned with the log rank test of survival analysis.

I would like to thank a number of people who have directly or indirectly contributed to this project. Herbert Robbins introduced me to the subject of Sequential Analysis and has been a constant source of intellectual guidance and inspiration. T. L. Lai worked with me on two papers which form the mathematical foundation of the presentation given here. Michael Woodroofe's research, beginning with his brilliant 1976 paper, has been a rich source of ideas and a stimulus to me to improve mine. In addition I want to thank Michael Hogan, Steve Lalley, and Thomas Sellke for many helpful discussions and technical assistance during the past several years. Some of their specific contributions are mentioned in the bibliographical notes. I thank Peter Elderon and Inchi Hu for proofreading and helpful suggestions on exposition. Thanks are also due to Jerri Rudnick and Judi Davis for their superb typing and cheerful retyping. And finally I want to thank the Office of Naval Research and the National Science Foundation for their support of my research.

Stanford, California
May, 1985

David Siegmund

Contents

CHAPTER I

Introduction and Examples

In very general terms there are two reasons for introducing sequential methods into statistical analysis. One is to solve more efficiently a problem which has a fixed sample solution. The other is to deal with problems for which no fixed sample solution exists. It is the first category which is the primary concern of this book, but we begin here with a few comments about the second.

Some problems are intrinsically sequential and cannot be discussed without considering their sequential aspects. An important example is a control system with unknown dynamics, about which something can be learned as the system operates. Dynamic programming is one method for dealing with problems of this sort. A beautiful recent summary is given by Whittle (1982, 1983).

Another intrinsically sequential problem is the fixed precision estimation of a parameter in the presence of an unknown nuisance parameter. It is almost obvious that one cannot give a confidence interval of prescribed length for the mean of a normal distribution based on a sample of some fixed size n if one does not know the variance of the distribution. (See Dantzig, 1940, for a formal proof.) However, by taking data sequentially one can use the data to estimate the variance and the estimated variance to determine a (random) sample size which will permit the mean to be estimated by a fixed length confidence interval. See Stein (1945) and Chapter VII. (In spite of its apparent omnipotence the method of dynamic programming appears not to have been applied to this problem.)

The principal subject of this book is sequential hypothesis testing and related problems of estimation. In contrast to the preceding examples, for most of the problems studied in detail there exist fixed sample solutions, and the reason for introducing sequential methods is to provide greater efficiency in some sense to be defined. Many of the problems might be attacked by dynamic programming. In fact, dynamic programming is a far reaching generalization

of the method originally developed in the pioneering papers of Wald (1947b), Wald and Wolfowitz (1948), and perhaps most importantly Arrow *et al.* (1949) to find Bayes solutions to problems of sequential hypothesis testing. Nevertheless, because we shall be primarily concerned with problems having vaguely specified loss functions, for the most part we shall ignore the possibility of finding optimal solutions and concentrate instead on procedures which can be directly compared with and improve upon those used most often in practice, to wit fixed sample size procedures evaluated in the classical terms of significance level, power, and sample size.

The simplest sequential test is a so-called curtailed test. Suppose that a machine produces items which may be judged good or defective, and we wish to infer on the basis of a random sample whether the proportion of defectives in a large batch of items exceeds some value p_0. Assume that the inference will be based on the number S_m of defectives in a random sample of size m. If m is a small proportion of the batch size, then S_m has approximately a binomial distribution with mean mp, where p is the true proportion of defectives in the batch; and a reasonable rule to test the hypothesis H_0: $p \leq p_0$ against H_1: $p > p_0$ is to reject H_0 if $S_m \geq r$ for some constant r, which at the moment need not be specified more precisely. If the sample is drawn sequentially and for some value k less than m the value of S_k already equals r, one could stop sampling immediately and reject H_0. More formally, let T denote the smallest value of k for which $S_k = r$ and put $T' = \min(T, m)$. Consider the procedure which stops sampling at the random time T' and decides that $p > p_0$ if and only if $T \leq m$. If one considers these two procedures as tests of H_0 against H_1, their rejection regions, to wit $\{T \leq m\}$ and $\{S_m \geq r\}$, are the same events, and hence the two tests have the same power function. Since the test which stops at the random time T' never takes more observations and may take fewer than the fixed sample test, it has a reasonable claim to be regarded as more efficient.

The preceding discussion has the appearance of delivering a positive benefit at no cost. However, the situation is not so clear if a second consideration is also to estimate p, say by means of a confidence interval. To continue the discussion with a slightly different example, suppose that $X(t)$, $t > 0$, is a Poisson process with mean value λt, and we would like to test H_0: $\lambda \leq \lambda_0$ against H_1: $\lambda > \lambda_0$. This problem might be regarded as an approximation to the preceding one, for if p is small the process of failures is approximately a Poisson process. However, the Poisson formulation might also apply to a reliability analysis of items having exponentially distributed lifetimes, which (in the simplest experimental design) are put on test serially with each failed item being immediately replaced with a good one. Then λ is the reciprocal of the mean time to failure of the items. It is clear from the discussion of the preceding paragraph that instead of a fixed time test which observes $X(t)$ until $t = m$ and rejects H_0 whenever $X(m) \geq r$, one can curtail the test at the stopping time $T' = \min(T, m)$, where T denotes the first time t such that $X(t) = r$, and reject H_0 whenever $T \leq m$.

Now consider the problem of giving an upper confidence bound for λ (hence

a lower confidence bound for the mean lifetime of an item). The standard fixed sample $(1-\alpha)\times 100\%$ confidence bound is $\lambda_2^*[X(m)]$, where $\lambda_2^*(n)$ is defined as the unique solution of

$$P_\lambda\{X(m)\le n\}=\alpha. \tag{1.1}$$

Since

$$P_\lambda\{X(t)\le n\}=P_\lambda\{w_{n+1}>t\} \tag{1.2}$$

where w_n is the waiting time for the nth event of the Poisson process, and since λw_n has a gamma distribution with parameter n (chi-square distribution with parameter $2n$), the value of $\lambda_2^*(n)$ is easily determined. For the curtailed test having exactly the same power function as a given fixed sample test, the corresponding confidence bound is slightly different. In analogy with (1.1) (see also Problem 1.1) define $\lambda_1^*(t)$ to be the solution of

$$P_\lambda\{T>t\}=\alpha. \tag{1.3}$$

Then a $(1-\alpha)\times 100\%$ upper confidence bound for λ based on the data $(T', X(T'))$ is

$$\lambda^*[T', X(T')]=\begin{cases}\lambda_1^*(T') & \text{if } T\le m\\ \lambda_2^*[X(m)] & \text{if } T>m\end{cases} \tag{1.4}$$

(see Problem 1.2 for a proof). The relation (1.2) between $X(t)$ and w_n makes it easy to determine $\lambda_1^*(t)$.

Lower confidence bounds, $\lambda_{*2}[X(m)]$ and $\lambda_*[T', X(T')]$ may be similarly defined. Confidence intervals may be obtained by combining upper and lower confidence bounds in the usual way. It turns out that $\lambda_*[T', X(T')]\le \lambda_{*2}[X(m)]$ with equality if and only if $X(m)\le r$, so one price of curtailment is a smaller lower confidence bound for λ.

The relation between $\lambda_2^*[X(m)]$ and $\lambda^*[T', X(T')]$ is not so simple.[1] Since the Poisson distributions have monotone likelihood ratio, the confidence bound $\lambda_2^*[X(m)]$ for the fixed sample size m is optimal in the strong sense of being uniformly most accurate (see Lehmann, 1959, p. 78ff. or Cox and Hinkley, 1974, p. 213). Since the statistician who observes $X(m)$ could by sufficiency define a randomized upper confidence bound with exactly the same coverage probability as (1.4), it follows that the fixed sample upper confidence bound is uniformly more accurate than that defined by (1.4). Hence less accuracy at the upper confidence bound is also a price of curtailment. (It is easy to see that the distributions of $(T', X(T'))$ have monotone likelihood ratio and hence that the upper confidence bound (1.4) is itself uniformly most accurate in the class of procedures which depend on the sample path $X(t)$ only until time T' (cf. Problem 1.7). We shall see that the method used to define (1.4) can be

[1] The material in this paragraph plays no role in what follows. It can be omitted by anyone not already familiar with the relevant concepts.

adapted to a variety of sequential tests, but it is very rare that the resulting confidence bounds have an easily described optimal property.)

The preceding discussion illustrates qualitatively both the advantages (smaller sample size) and the disadvantages (less accurate estimation) associated with a sequential test. In Chapters III and IV these tradeoffs are studied quantitatively.

Remark 1.5. The reader interested in the foundations of statistics may find it interesting to think about various violations of the likelihood principle (Cox and Hinkley, 1974, p. 39) which occur in the sequel. One example is in the definition of confidence bounds. For a Bayesian with a prior distribution for λ which is uniform on $(0, \infty)$, an easy calculation shows that for any stopping rule τ, $\lambda_2^*[X(\tau)]$ defined above is a $1 - \alpha$ posterior probability upper bound for λ, i.e. $P\{\lambda \leq \lambda_2^*[X(\tau)] | \tau, X(\tau)\} = 1 - \alpha$. In particular, for the fixed sample experiment the confidence and posterior probability bounds agree. But for the sequential experiment, the particular stopping rule plays an important role in the determination of a confidence bound with the effect that the "confidence" of the posterior probability upper bound is strictly less than $1 - \alpha$ (see also Problem 1.5).

Although the methods described in the following chapters can be adapted to the investigation of a wide variety of sequential procedures, the primary concrete example studied in detail is the repeated significance test and some of its modifications. Let $x_1, x_2, \ldots$ be independent, normally distributed random variables with unknown mean μ and known variance σ^2, which without loss of generality can be taken equal to 1. Let $S_n = x_1 + \cdots + x_n$. The standard fixed sample .05 level significance test of H_0: $\mu = 0$ against H_1: $\mu \neq 0$ is to reject H_0 if and only if $|S_n| \geq 1.96n^{1/2}$. Here n is the arbitrary, but fixed sample size of the experiment. Suppose now that if H_1 is actually true it is desirable to discover this fact after a minimum amount of experimentation, but no similar constraint exists under H_0. Such might be the case in a clinical trial where x_i represents the difference in responses to two medical treatments in the ith pair of a paired comparison experiment. If H_0 is true, the two treatments are equally good, and from the patients' point of view the experiment could continue indefinitely. However, if H_1 is true, one or the other treatment is superior, and the trial should terminate as soon as possible so that all future patients can receive the better treatment.

An ad hoc solution to the problem of the preceding paragraph is the following. Let $b > 0$ and let m be a maximum sample size. Sample sequentially, stopping with rejection of H_0 at the first $n \leq m$, if one exists, such that $|S_n| > bn^{1/2}$. Otherwise stop sampling at m and accept (do not reject) H_0. The significance level of this procedure is

$$\alpha = \alpha(b, m) = P_0\{|S_n| > bn^{1/2} \text{ for some } n \leq m\}, \tag{1.6}$$

which means that b must be somewhat larger than 1.96 (depending on m) in order that $\alpha(b, m) = .05$.

Tests of this sort were criticized by Feller (1940), who alleged that they were used in extrasensory perception experiments without making the necessary adjustment in the value of b to account for the sequential nature of the experiment. (For these experiments, S_n might count the excess of correct over incorrect guesses by a subject who supposedly can predict the outcome of a coin toss before being informed of the result.) Feller also complained that there was no definite value of m, so that one should consider the significance level to be

$$\lim_{m\to\infty} \alpha(b, m),$$

which is known to equal 1 (for example, as a consequence of the law of the iterated logarithm). Robbins (1952) gave an upper bound for $\alpha(b, m)$ and posed the problem of giving a good approximation to α.

Such repeated significance tests were studied by Armitage *et al.* (1969) and by MacPherson and Armitage (1971), who evaluated their significance level, power, and expected sample size by lengthy numerical computations. The theoretical research from which this book has developed began with Woodroofe's (1976) and Lai and Siegmund's (1977) approximation for α (cf. (4.40)), which was followed by a series of papers approximating the power and expected sample size of repeated significance tests, extending the results to more general models, and suggesting certain modifications of the test itself (see Chapters IV and V).

As a preliminary to our study of repeated significance tests, we discuss the sequential probability ratio test in Chapter II. Although it seems unlikely that this test should be used in practice, the basic tools for studying it, to wit Wald's likelihood ratio identity (Proposition 2.24) and Wald's partial sum identity (Proposition 2.18), are fundamental for analyzing more useful procedures. So called cusum procedures for use in quality control are discussed briefly in II.6.

Chapters III–V form the core of the book. The main conceptual ideas are introduced in Chapter III in a context which minimizes the computational problems. Truncated sequential probability ratio tests and Anderson's modification of the sequential probability ratio test are also discussed. Repeated significance tests are studied in detail in Chapter IV. A number of more difficult examples are presented in Chapter V to illustrate the way one can build upon the basic theory to obtain reasonable procedures in a variety of more complicated contexts.

Chapters VI and VII deal with special topics. Chapter VI is concerned with the allocation of treatments in clinical trials, and Chapter VII briefly introduces the theory of fixed precision confidence intervals.

In order to maximize attention to statistical issues and minimize difficult probability calculations, the mathematical derivations of Chapters III and IV are essentially limited to the artificial, but simple case of a Brownian motion process. Corresponding results for processes in discrete time are given without proof and used in numerical examples. Chapters VIII–X provide the mathematical foundation for these results. Chapter XI is concerned with some miscellaneous probability calculations which are conceptually similar but

technically more difficult than those which appear earlier in the book. Four appendices present some background probabilistic material.

The most obvious omission from this book is a discussion of Bayesian sequential tests. Even for the non-Bayesian, the use of prior probability distributions is a useful technical device in problems which can reasonably be treated decision-theoretically (i.e. have action spaces and loss functions). The two principal fields of application of sequential hypothesis testing are sampling inspection and clinical trials. Of these, the former seems often to admit a decision-theoretic formulation, but the latter not. (For a contrary view, see Anscombe, 1963, and for further discussion see IV.6.) Hald (1981) gives a systematic treatment of sampling inspection with ample discussion of Bayesian methods. Other general introductions to sequential Bayesian hypothesis testing without particular applications in mind are given by Ferguson (1967), Berger (1980), and especially Chernoff (1972). To avoid a substantial increase in the length of this book, the subject has been omitted here.

The formal mathematical prerequisites for reading this book have been held to a minimum—at least in Chapters II–VII. It would be helpful to have some knowledge of elementary random walk and Brownian motion theory at the level of Feller (1968), Cox and Miller (1965), or Karlin and Taylor (1975). Appendix 1 attempts to give the reader lacking this background some feeling for the essentials of Brownian motion, devoid of all details. Martingale theory makes a brief appearance in V.5. Appendix 3 presents the necessary background—again informally.

One bit of nonstandard notation that is used systematically throughout the book is $E(X; B)$ to denote $E(XI_B)$. (Here I_B denotes the indicator variable of the event B, i.e. the random variable which equals 1 if B occurs and 0 otherwise. E denotes expectation.) Some of the notation is not consistent throughout the book, but is introduced in the form most convenient for the subject under discussion. The most important example is the notation for exponential families of probability distributions, which are introduced in II.3, but parameterized slightly differently in II.6 (the origin is shifted). They reappear in the original parameterization in Chapter VIII, and in Chapter X they change again to the parameterization of II.6.

Problem sets are included at the end of each chapter. A few problems which are particularly important have been designated with $*$. Those which are somewhat more difficult or require specialized knowledge are marked †.

Problems

1.1. Suppose that the Poisson process $X(t)$ is observed until the time w_r of the rth failure. Show that $\lambda_1^*(w_r)$ is a $(1-\alpha)\,100\%$ upper confidence bound for λ.

1.2. Prove that for λ^* defined by (1.4)

$$P_\lambda\{\lambda^*[T', X(T')] \geq \lambda\} \geq 1 - \alpha \qquad \text{for all } \lambda.$$

Hint: Note that $\lambda_1^*(m) = \lambda_2^*(r-1)$. Consider separately the two cases $\lambda_1^*(m) \geq \lambda$ and $\lambda_1^*(m) < \lambda$.